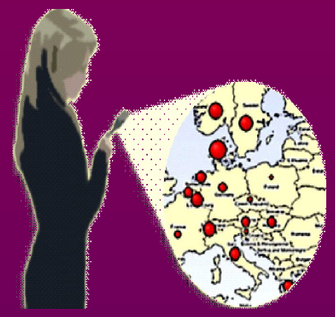


Módszertani Füzetek
2009/1

Szerkesztette:
Füstös László
Szalma Ivett

MTA Szociológiai Kutatóintézete
Társadalomtudományi Elemzések Akadémiai Műhelye (TEAM)

Füstös László

A sokváltozós adatelemzés módszerei*

* Készült az OTKA NN **76722** Fieldwork and data analysis of the European Social Survey Round 4 (ESS R4) keretében.

ELŐSZÓ

A sokváltozós statisztikai módszereket sokan említik többváltozós statisztikai módszerekként. A különbség a két megnevezés között csekélynek tűnik, azonban akik többváltozós statisztikaként említik, nem látják a módszerek lényegét; ti. nem egynél több, hanem nagyon sok változó térében jellemezzük a megfigyelési egységeket, és vagy a változók, vagy a megfigyelési egységek, vagy mindkettő struktúráját vizsgáljuk, képezzük le őket egy redukált, lehetőleg kevés dimenziószámú látens térbe.

A könyv tartalmazza a látens változós modelleket, bemutatja azok elméletét, matematikai statisztikai becslési módszereit, és alkalmazási példáit.

Milyen lépések szükségesek a sokváltozós matematikai statisztikai modellek elsajátításában?

A tanulás folyamatát célszerű két részre osztani: egyik az elméleti, másik a gyakorlati tudás megszerzése. Ha el akarom sajátítanom az adatelemzést, először tisztába kell jönnöm a matematikai statisztika elméletével. Ha az elméleti tudást megszereztem, még korántsem vagyok beavatva az adatelemzés művészetébe. Ennek csak akkor leszek a mestere, művésze, ha sok gyakorlattal a hátam mögött az elméleti tudásom és a gyakorlati tapasztalataim végül egybeolvadnak az intuíciómban – ami minden mesterség, művészet birtoklásának a lényege. De az elmélet és a gyakorlat elsajátításán kívül egy harmadik tényező is szükséges: az adatelemzés mesterségének, művészetének birtokába jutni csak szenvedélyes akarással lehet.

Kívánom, hogy mindenki, aki szenvedélyes akarással próbálja elsajátítani a sokváltozós statisztika módszereit, váljon a sokváltozós modellezés mesterévé, művészévé.

Budapest, 20099. december

Füstös László

TARTALOM

I. rész

Általános bevezető

1. fejezet

Adat, adatelemzés

Adatok halmazán a változók (a megfigyelési egységek, vizsgált személyek jellemzőinek, tulajdonságainak) megfigyelt, mért (manifeszt) értékei összességét értjük.

A teljes adathalmazt az $\mathbf{X} = [x_{ij}]$ mátrix tartalmazza, amelynek annyi sora van, ahány megfigyelést végeztünk. A változók mért értékeit egy-egy oszlopvektorba rendezve x_{ij} jelöli az i-edik megfigyelés j-edik változóra vonatkozó konkrét számértékét.

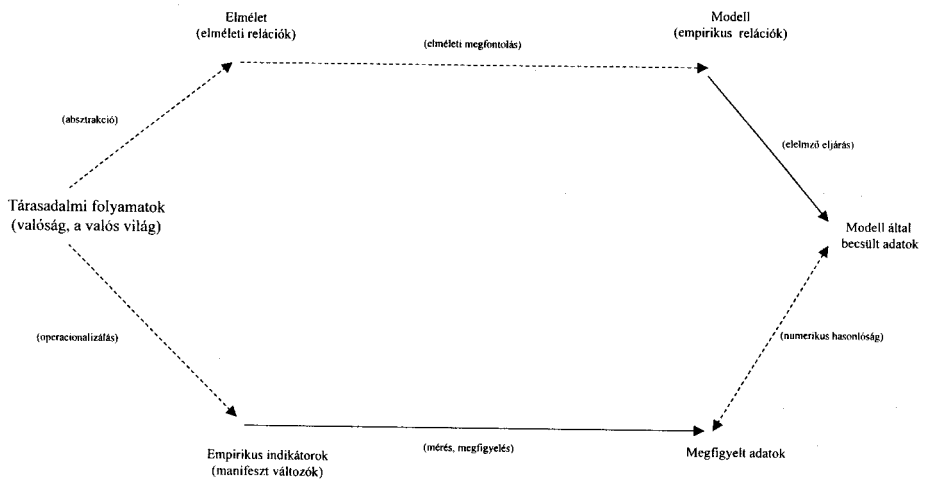
Általában azt mondhatjuk, hogy a sokváltozós adatelemzés statisztikai módszerei ilyen típusú adattáblákból indulnak ki. Az esetek jelentős részében mind a mutatók, mind a megfigyelések összessége minta jellegű. Valamely konkrét rendszert ugyanis általában tetszőleges számú jellemzővel ruházhatunk fel, s a megfigyelések szóba jöhető köre is esetenként igen széles lehet. Az *alkalmazó szaktudomány* igen lényeges alapvető feladata, hogy a rendszert általában vagy a kutatott irányból valóban jellemző változókkal írja le. Gyakorlatilag – ha egy területen még igen kevés ismerettel rendelkezünk – a rendszer struktúráját jól megvilágító, reális jellemzők körét esetleg egyáltalán nem tudjuk eleve rögzíteni. Elemző munkánk ilyenkor sokkal nehezebb, több körültekintést igényel. A sokváltozós módszerek mindenesetre ilyenkor is számos lehetőséget adnak a praktikusan figyelembe veendő változók körének megállapítására.

Komoly figyelmet kell fordítanunk a megfigyelések összességére is. Ha az n számú megfigyelés (pl. vállalat) teljes körű felmérést jelentett (pl. egy ágazat valamennyi vállalata szerepel a vizsgálatban), akkor e tekintetben az összes aktuális információt figyelembe vettük. Egyéb esetekben a megfigyelések reprezentatív jellegére feltétlenül ügyelni kell.

A mondottak értelmében tehát a vizsgált rendszert a mutatók és a megfigyelések alkalmas körével valóban statisztikailag írtuk le. A sokváltozós adatelemzés módszereinek feladata, hogy az induló tábla komplex vizsgálatával lehetővé tegyék a statisztikailag jellemzett háttérrendszer tulajdonságainak, struktúrájának minél alaposabb megismerését.

A társadalomtudományi kutatások egyik legnehezebb kérdése a megfigyelés és az elmélet közötti kapcsolat megteremtése. Megfigyelni, mérni csak ritkán tudunk közvetlenül. A közvetett mérés és az elméleti jellemző átfedése nem tökéletes, mert

– az általánosabb elméleti fogalomnak az indikátorok, a megfigyelt változók csak egy részét fogják át,
– az indikátor valami mást is kifejez, ti. a megfigyelési egység specifikus tulajdonságát.



1.1. ábra. Az elmélet és a mérés viszonya

Amikor tudásunk az adott problémáról nem elégséges ahhoz, hogy elméleti modelleket, hipotéziseket fel tudjunk állítani, a statisztikai módszereket használjuk arra, hogy az adatok struktúráját felderítsük. Az ilyen feltáró eljáráskor a váratlan, nem interpretálható eredményeket vagy eldobjuk, vagy ötletszerűen értelmezzük. A *feltáró elemzés* során korlátozott értelmű magyarázatokhoz, hipotézisekhez, modellekhez juthatunk, amelyek érvényességét bizonyító eljárással igozolni kell.

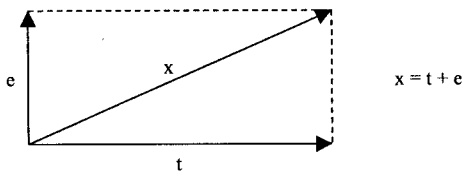
Igazoló elemzést akkor végzünk, amikor elméleti modell felállítása után vizsgáljuk adathalmazunkat. Az elmélet származhat a „közmegegyezésből", más adatok feltáró elemzéséből stb. Az elmélet tehát különböző szintű lehet, mindenesetre ki kell elégítenie azt a minimális követelményt, hogy formalizálható legyen. A hipotézisekből, illetve a hipotézisek alapján építjük fel a modellt, amely az elméletet reprezentálja. Ezután a modellt teszteljük, és ha jól illeszkedik az adatokhoz, akkor azt mondhatjuk, hogy az elmélet egy igazolását kaptuk.

A sokváltozós módszereknél az elemzés általában valamilyen formában a változók magyarázatára irányul. Amikor egy változó magyarázatáról beszélünk, akkor a változó varianciájának, megfigyelt értékei eltéréseinek magyarázatára gondolunk. Egy változót, amely ugyanazt a konstans értéket veszi fel a megfigyelések során, nem szükséges magyarázni. Azt kívánjuk tudni, hogy miért változott a megfigyelt értéke, mi a forrása a változásnak.

Alapvetően két dolgot kell tisztáznunk:

a) a mérés megbízhatóságát,

b) a mérés érvényességét.



1.2. ábra. A változó értéke a szisztematikus tag és a véletlen komponens eredője

A klasszikus méréselmélet szerint feltételezhetjük, hogy a megfigyelt x változó két közvetlenül nem mérhető komponens eredője. Az egyik a szisztematikus komponens (jele t), és ehhez adódik hozzá a másik komponens, a véletlen tag (jele e). (1.2. ábra)

Feltételezzük, hogy

– $E(e) = 0$ a hiba várható értéke nulla,

– $\rho(e_1, t_2) = 0$ a hiba és a szisztematikus komponens korrelálatlan,

– $\rho(e_1, e_2) = 0$ a különböző mérések hibatagjai korrelálatlanok.

Ezekből a feltételezésekből következik, hogy

$$E(x) = E(t).$$

A mérés megbízhatóságát a szórásnégyzet (variancia, σ^2) segítségével fejezhetjük ki. A $\sigma_x^2 = \sigma_{(t+e)}^2$-ből a fenti feltételek felhasználásával következik, hogy

$$\sigma_x^2 = \sigma_t^2 + \sigma_e^2.$$

A megbízhatóságot (ρ_{xt}^2) az igazi és a megfigyelt szórásnégyzet arányaként definiáljuk:

$$\rho_{xt}^2 = \sigma_t^2/\sigma_x^2,$$

ahol $0 \leq \rho_{xt}^2 \leq 1$.

Ha a mérés nem tartalmaz hibát, akkor $\rho_{xt}^2 = 1$, ha csak hibát tartalmaz, $\rho_{xt}^2 = 0$.

A megbízhatósági együtthatót csak akkor tudjuk becsülni, ha legalább két mérésünk van ugyanarról a dologról:

$$x = t + e,$$

$$x' = t + e'.$$

Ha a hiba szórásnégyzete egyenlő a két mérésnél ($\sigma_e^2 = \sigma_{e'}^2$), akkor x és x' *párhuzamos mérések.*

Ha a két mérés hibájának szórásnégyzete különbözik, akkor x és x' *tau-ekvivalens* mérések.

A párhuzamos méréseknél belátható, hogy

$$\rho_{xx'}^2 = \sigma_t^2/\sigma_{x'}^2,$$

vagyis a párhuzamos mérések közötti korreláció négyzete egyenlő a megbízhatósággal.

A klasszikus méréselmélet a mérés érvényességét két változó esetén a két változó közötti korrelációval méri:

$$\rho_{xy} = \sigma_{xy}/\sigma_x\sigma_y = \sigma_t^2/\sigma_x^2.$$

Párhuzamos mérések esetén a mérés megbízhatósága egyenlő a mérés érvényességével. Általában ez a szabály nem érvényes, viszont belátható, hogy

$$\rho_{xy} \leq \sqrt{\rho_{xt}^2}.$$

A klasszikus méréselméletben a mérési hiba véletlenszerű. Ez sokszor nem teljesül, így a következő mérési modellt állítjuk fel:

$$x = t + s + e,$$

ahol s: szisztematikus hiba,
e: véletlen hiba.

Mivel a véletlen hibára vonatkozó korábbi feltételeink itt is érvényesek, belátható, hogy

$$E(x) = E(t) + E(s).$$

A megfigyelt szórásnégyzet

$$\sigma_x^2 = \sigma_t^2 + \sigma_s^2 + \sigma_e^2 + 2\sigma_{ts},$$

ahol σ_{ts} a két tag közti kovariancia.

A megbízhatóság a nem véletlen komponensek szórásnégyzeteinek aránya:

$$\rho_{xt}^2 = \frac{\sigma_t^2 + \sigma_s^2 + 2\sigma_{ts}}{\sigma_x^2} = \frac{\sigma_x^2 - \sigma_e^2}{\sigma_x^2}.$$

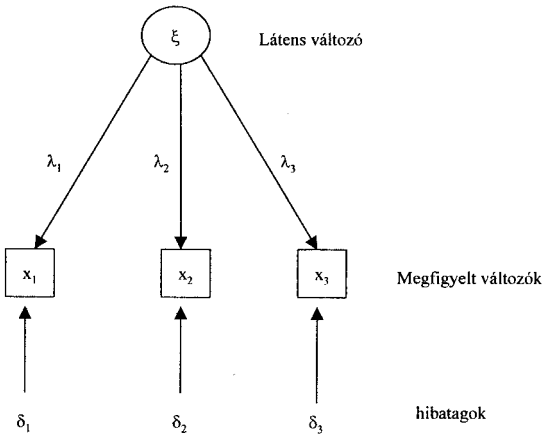
A mérés érvényessége $= \dfrac{\sigma_t^2}{\sigma_x^2}$.

Eddig azt feltételeztük, hogy az elméleti változót mérési hibával, de közvetlenül tudjuk mérni. Most azt az esetet tárgyaljuk, amikor az elméleti változó csak közvetetten mérhető. Ilyenkor az elméleti változóval kapcsolatban levő megfigyelhető változókat mérjük.

A latens változót (ξ) mint a megfigyelt változók legfontosabb közös részét fejezzük ki. (1.3. ábra)

Egyenletekkel felírva a modellt:

$$x_1 = \lambda_1 \xi + \delta_1,$$
$$x_2 = \lambda_2 \xi + \delta_2,$$
$$x_3 = \lambda_3 \xi + \delta_3.$$



1.3. ábra. Mérési modell

A változók kapcsolatrendszerének leírására a gyakorlatban legtöbbször a lineáris modelleket használjuk. A változók közötti valódi kapcsolat azonban lehet, hogy nemlineáris. Ilyenkor a lineáris modell kisebb vagy nagyobb mértékben eltér a valóságtól.

A problémát megoldhatjuk a változók megfelelő transzformációival akkor, ha ismert a kapcsolat típusa. Ha a kapcsolat nem ismert, vagy nem adható meg linearizáló transzformáció, akkor közelítő eljárást alkalmazhatunk.

Az elméleti változók lineáris kapcsolatának egyik alapvető feltételezése a mérési skála folytonossága. Kategorikus változók esetén ez a feltételezés definíció szerint nem teljesül.

Ha feltételezhetjük, hogy a megfigyelt változó mögött meghúzódó latens elméleti változó természete folytonos, akkor a kategorikus megfigyelt változót is folytonosnak tekintjük. De ha az elméleti változó nem folytonos, a megfigyelt változót sem tekinthetjük annak, így a linearitási feltevés sem tartható. Kategorikus elméleti változók esetén a helyes eljárás az, hogy az adatokat a kategóriák szerint csoportokra bontjuk, és a csoportokra külön-külön végezzük el a modell illesztését, majd a becsléseket összevetjük.

2. fejezet

A módszerek klasszifikációja

A sokváltozós technikák az adatmátrix oszlopai (oszlopvektorai), illetve sorai (sorvektorai) között végeznek különböző vizsgálatokat, így a változók, illetve a megfigyelési egységek közötti összefüggések kimutatására alkalmazhatók.

Általánosságban az adattáblázat a megfigyelési egységek, objektumok (egyedek) és a változók (az objektumokon mért jellemzők) dimenziója mellett tartalmazhatja az idő dimenzióját is, vagyis az egyedek az adott változók különböző időpontokra vonatkozó értékeivel is jellemezhetők.

A háromdimenziós adatmátrix alapján Catell különböző elemzési technikákat különböztetett meg (2.1. ábra).

Az *R* technika adott időpontban bizonyos változók értékeit elemzi a megfigyelt objektumokon, és a változók közötti összefüggéseket keresi. Például az urbanizáció szintjét egy meghatározott időpontban mérő mutatók, az infrastrukturális fejlettség és a társadalom különböző szférái (gazdasági szerkezet, kulturális, egészségügyi terület) jellemzőinek a kapcsolatát vizsgálhatjuk pl. faktorelemzéssel a magyar városok mintáján.

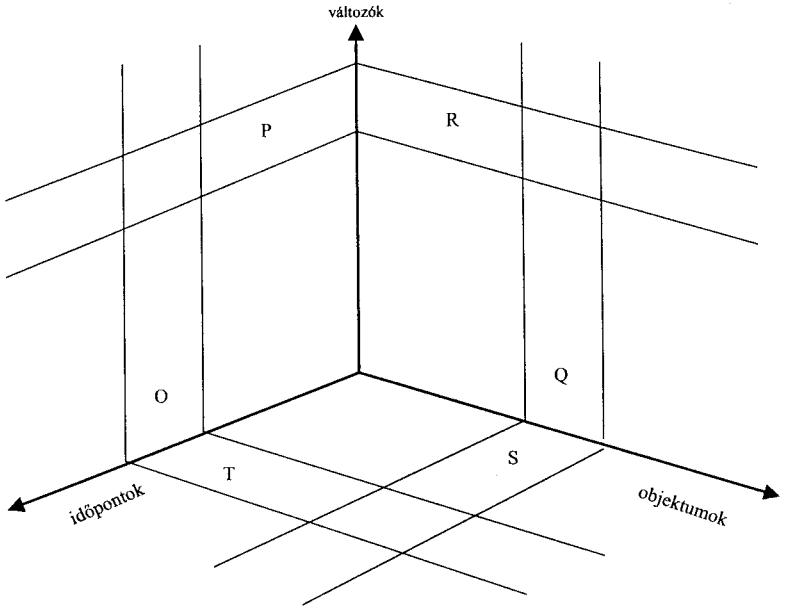
A *Q* technika adott időpontban bizonyos objektumokon figyeli meg a változók értékeit, és az objektumok kapcsolatát írja le. Pl. az urbanizáció mutatói alapján megkísérelhetjük a városokat típusokba, homogén csoportokba sorolni pl. klaszterelemzéssel.

A *P* technika bizonyos változók különböző időpontokban megfigyelt értékeit elemzi adott megfigyelési egységre vonatkozóan. A megfigyelt adatsorok egy-egy változó idősorát jelentik. *P* technika pl. a trendelemzés vagy az idősorok faktorelemzése.

Az *O* technika időpontok közötti összefüggéseket elemez a változók adott időpontokra vonatkozó értékei alapján egy megfigyelési egységre vonatkozóan. Vizsgálhatjuk pl. bizonyos társadalmi jelenségek alapján egy adott történeti időszak „szezonalitásait".

A *T* technika ugyancsak időpontok közötti összefüggéseket elemez, de egy adott változó különböző megfigyelési egységekre vonatkozó értékei alapján. Pl. egy adott populációban milyen eltérést találunk egy kérdés különböző időpontokban való megítélésében.

Az *S* típusú elemzés objektumok azonos változóra vonatkozó idősora alapján az objektumok kapcsolatát vizsgálja. Pl. országok fejlődési típusát elemezzük az országok GDP/fő idősorai alapján.

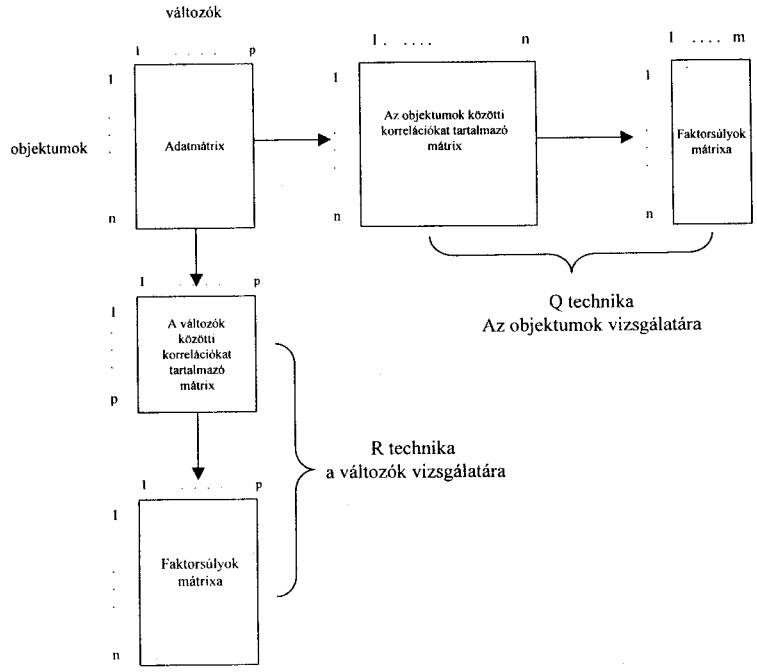


2.1. ábra. A háromdimenziós adattömb szemléltetése

A leggyakrabban alkalmazott *R* és *Q* technikákat egy adott időpontra vonatkozó adatmátrixon illusztráljuk. Vegyük a sokváltozós módszerek közül a faktorelemzés általánosan használt módszerét! A kétfajta megközelítés feldolgozási különbségeit a 2.2. ábra mutatja.

A következőkben a módszereket aszerint csoportosítjuk, hogy az adatmátrixon milyen alapvető műveleteket hajtunk végre. (Itt *R* típusú elemzésre gondolunk, vagyis a mátrix oszlopai között végzünk műveleteket, ahol a mátrix oszlopai a változókat, sorai pedig az objektumokat jelölik. A három séma a módszerek három típusát jelöli (2.3. ábra).

Az a) séma a regressziós modellt reprezentálja: a függő változót (y) a bal oldali x változókkal próbáljuk magyarázni. Az útelemzés a regressziós modell egymás utáni ismétlését jelenti.



2.2. ábra. Az *R* és a *Q* technika megkülönböztetése

A b) séma a kanonikus korreláció modelljét mutatja, amelyben az x változóhalmaz nem megfigyelt változón keresztül határozza meg az y változóhalmaz közös latens változóját.

A c) séma a faktorelemzés modelljét mutatja, amelyben a megfigyelt változókat nem megfigyelt változók határozzák meg.

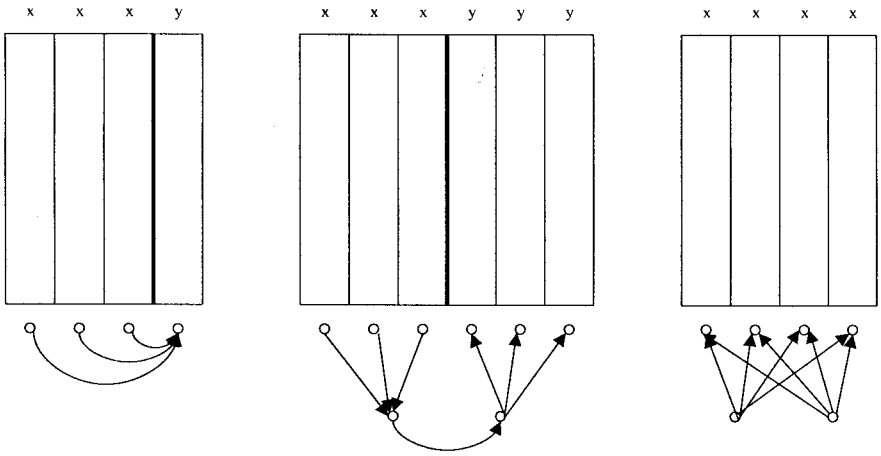
Ebben a három sémában elhelyezhető a többi sokváltozós eljárás is. Nézzük pl. a diszkriminanciaelemzés módszerét! A módszert akkor alkalmazzuk, amikor a mintát csoportokba soroljuk, és a csoportosított objektumokon mérjük meg a változók értékeit. Vegyük például azt az esetet, amikor két csoportot különítünk el! Legyen az y dichotom változó, és mutassa a csoporthoz való tartozást! Vagyis $y = 1$, ha a megfigyelés az I. csoportba kerül, és $y = 0$, ha a II. csoportba soroljuk a kérdéses megfigyelést. Ebben az esetben a diszkriminanciaelemzés az a) séma modelljével azonosítható. A diszkriminanciaelemzés így nem más, mint egy dichotom függő változóval működő regressziós modell. Amennyiben több mint két csoport van a mintában, újabb dichotom változókat vezetünk be. Például három csoport esetén két dichotom változóval tudjuk jelölni a csoporthoz való tartozást. Az $y_1 = 1$, ha a megfigyelés az I. csoportba tartozik, 0 különben, az $y_2 = 1$, ha egy megfigyelés a II. csoportba tartozik, 0 különben. Az (y_1, y_2) változópárral tehát a megfigyeléseket három csoportba egyértelműen be tudjuk sorolni. Az (1,0) az I. csoportot, a (0, 1) a II. csoportot kódolja, a (0,0) értékpár pedig a III. csoport kódja. Ilyen meggondolással a három- vagy több csoportos diszkriminanciaelemzés megfelel a b) séma logikájának. Ennek alapján látható a diszkriminanciaelemzés és a kanonikus korrelációelemzés modelljeinek formális megegyezése.

Nézzük a varianciaelemzés módszerét! Az a) séma x változóit a csoportképző ismérv jelölésére használjuk, az y változó legyen a csoportokban mért függő változó.

a) Többváltozós regresszió
- útelemzés
- diszkriminancia elemzés (kétcsoportos)
- varianciaelemzés

b) Kanonikus korreláció
- diszkriminancia elemzés (többcsoportos)
- többváltozós varianciaelemzés

c) Faktorelemzés
- kanonikus diszkriminancia elemzés
- sokdimenziós skálázás
- latens osztályelemzés
- latens tulajdonságelemzés



2.3. ábra. A sokváltozós módszerek osztályozása az adatmátrix felhasználásával

Ekkor a varianciaelemzés modellje formálisan egybeesik a regresszióelemzés modelljével. Ha nem egy, hanem több (y) változónk van, akkor a b) séma alapján a többváltozós varianciaelemzés modelljének és a kanonikus korreláció modelljének a formális egybeesését láthatjuk.

A c) sémánál is bevezethetünk dichotom változókat, amelyek a csoporthoz tartozást fejezik ki. Ha a megfigyelt változókat olyan faktorokkal akarjuk magyarázni, amelyek a csoportokat a lehető legjobban szétválasztják, a kanonikus diszkriminanciaelemzés modelljéhez jutunk.

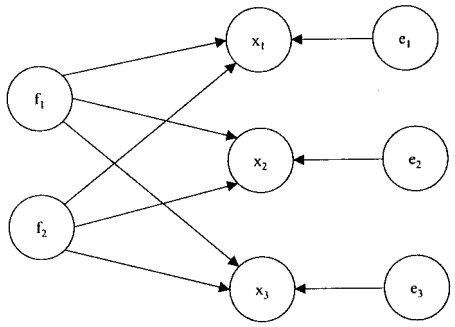
A sokváltozós matematikai statisztikai módszerek klasszifikációját két további szempont egyidejű figyelembevételével is megadhatjuk. Az egyik szempont arra vonatkozik, hogy a változók halmazában megkülönböztetünk-e függőségi viszonyokat. Ha igen, akkor a változók között két vagy több változóhalmazt különítünk el. Ha nem, akkor a változókat együtt elemezzük.

A másik szempont a megfigyelési egységek belső tagozódásához kapcsolódik. Tekinthetjük egy mintának az összes megfigyelést, de feltételezhetünk almintákat is a megfigyelési egységeken belül. A 2.1. táblázat a könyvben szereplő módszerek kétszempontú osztályozását mutatja.

A Q_1 cellában felsorolt sokváltozós módszerek közül az egyik legismertebb eljárás a főkomponens- vagy a faktorelemzés. A faktorelemzés alapgondolata az, hogy a megfigyelt változók kifejezhetők nem megfigyelt változók, faktorok (f) lineáris függvényeiként. (2.4. ábra)

	A változók kapcsolata	
	kölcsönös	oksági
Egy minta	– főkomponens-elemzés – faktorelemzés – kereszttábla-elemzés – sokdimenziós skálázás – korreszpondenciaelemzés – latens osztály- és tulajdonság-elemzés Q_1	– loglineáris modell – útelemzés – kanonikus korrelációelemzés – LISREL-modell – LVPLS-modell – többszempontú módszerek Q_2
Alminták	Q_3 – szóráselemzés – diszkriminanciaelemzés – faktoriális diszkriminancia-elemzés – lineáris szeparáció – klaszterelemzés	Q_4 – többszörös kovarianciaelemzés

2.1. táblázat. A módszerek osztályozása a minta és a változók típusa alapján



2.4. ábra. A faktorelemzés grafikus-modellje

A megfigyelt változók számával megegyező számú faktorral a változókat pontosan le tudjuk írni. Kevesebb faktorral a megfigyelt változók szisztematikus komponenseire kapunk egy becslést. Ebben az esetben a változók mintabeli eltéréseit nem tudjuk teljesen magyarázni, ezért kell szerepeltetnünk a véletlen komponenst.

A *főkomponens-elemzés* a faktorokat a megfigyelt változók csökkenő varianciájú lineáris kombinációjaként (függvényeként) fejezi ki úgy, hogy a faktorok egymással korrelálatlanok legyenek.

Az *arbitráris faktorelemzés* hipotézisvektoroknak megfelelően állít elő maximális varianciájú páronként korrelálatlan faktorokat. A hipotézisvektorokban írjuk elő, hogy az egyes faktorok előállításában mely változók vegyenek nagy súllyal részt.

Az *image-elemzés* a megfigyelt változók szisztematikus komponenseit a többi változó lineáris regeressziós függvényével becsüli. Az image-elemzés minden változót két tagra bont: egyik rész az image, a többi változó lineáris függvényével meghatározható

rész, a másik tag a változónak az a része, amit nem tudunk meghatározni regresszióval, ezt anti-image-nak nevezzük.

Az *alfa-faktorelemzés* alapgondolata, hogy a minta és a teljes sokaság (amiből mind a megfigyelési egységek, mind a változók reprezentatív mintáját vettük) faktorai eltérnek, ezért a minta olyan faktormegoldását keresi, amely maximálisan korrelál a teljes sokaságból számítható faktorokkal.

A *kanonikus faktorelemzés* a változók olyan faktorait keresi meg, amelyek maximálisan korrelálnak a megfigyelt változóhalmazzal.

A kereszttábla két vagy több változó (szempont) szerint rendezett adatokat tartalmaz. Elemzése révén a változók függetlenségét tesztelhetjük, vagy a köztük lévő asszociációs kapcsolat szorosságát mérhetjük.

A *sokdimenziós skálázás* a változóknak és a megfigyeléseknek egy redukált dimenziószámú térben azt a konfigurációját (pontábráját) keresi, amelyben a pontok közötti távolságok jól illeszkednek az eredeti mintatérben számított különbözőségekhez vagy hasonlóságokhoz. Egyes eljárásai keresik bizonyos egyedek preferenciáinak (a változókra vonatkozó preferenciáknak) legjobban megfelelő „ideális pontjait" vagy tengelyeit a redukált térben. Más eljárásai egyedek (egyének) különbözőségeit skálázzák egy közös tér dimenzióihoz viszonyítva.

A korreszpondenciaelemzés során a változók kölcsönös kapcsolatát tételezzük fel, és dimenziócsökkentést hajtunk végre. Ebben a vonatkozásban a faktorelelmzéshez hasonlít leginkább a módszer. Különlegességét az adja, hogy nominális mérési szintű adatokra is alkalmazható, és az eljárás során kapott redukált dimenziószámú térben együtt jelennek meg az adatmátrix sorai és oszlopai.

A *latens tulajdonság-modellben* (latent trait model) a kvalitatív (diszkrét értékekkel rendelkező) megfigyelt változók asszociációit kvantitatív, folytonos latens változókkal magyarázzuk. A latens változókat latens tulajdonságnak, latens jellemzőnek, latens traitnek nevezzük. A trait a pszichológiában használatos kifejezés, olyan általános tulajdonságot, jellemző vonást jelent, amelynek segítségével a személyiségeket meg tudjuk különböztetni. Statisztikai értelemben a latens trait megfelel a közös faktor terminológiának.

A *latens struktúraelemzés* hasonlít a faktorelemzéshez, azonban míg a faktorelemzés folytonos megfigyelt változókat magyaráz folytonos latens változókkal, a latens struktúraelemzésnél mind a manifeszt, mind a latens változók között lehetnek kvalitatív változók is. Ezenkívül a latens struktúraelemzésnél nem kell feltételeznünk sem a gyakrolatban legtöbbször nem teljesülő normalitási feltételt, sem a mérési skála folytonosságát.

A latens struktúraelemzés kitüntetett esete, amikor a megfigyelt változók diszkrét, kategorikus változók (sokszor dichotómok), és egy vagy több kategorikus latens változónk van. Ezt az esetet hívjuk *latens osztályelemzésnek*. A modell alapfeltételezése, hogy a latens változók bármelyik kategóriájában a megfigyelt változók függetlenek egymástól. A manifeszt változók megfigyelt kapcsolatait az adatoknak a latens változó kategóriája szerinti klasszifikációja eredményezi.

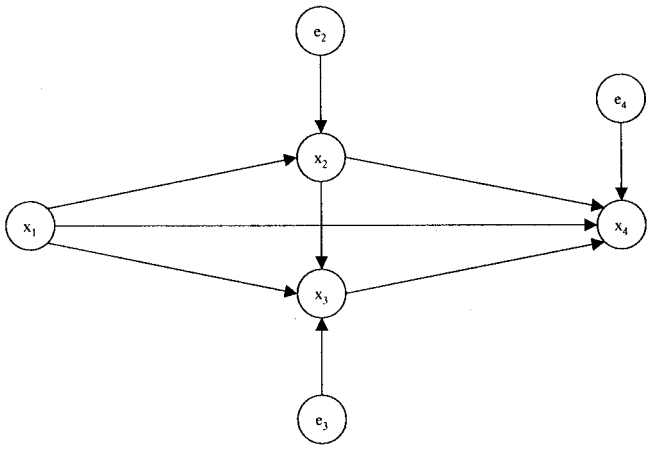
A társadalomtudományban gyakran vizsgáljuk az egyének és a társadalom kapcsolatát. Az egyének és társadalom egymás közötti kapcsolata, interakciója egyrészt jelenti azt a hatást, amely során a társadalom vagy annak egy csoportja befolyásolja az egyén (individuum) viselkedését, vélekedését, másrészt annak a csoportnak a tulajdonságait az adott társadalmi csoportot alkotó egyének befolyásolják. Amennyiben az egyéneket és a társadalmi csoportokat hierarchikus rendszerként fogjuk fel, az egyének és a társadalmi csoportok a hierarchikus rendszer különböző szintjeiként értelmezhetők. Ezeket a szinteket megfigyelhetjük, jellemezhetjük manifeszt változókkal. *A többszintű elemzés*

(multilevel analysis) az egyének és társadalmi csoportok megfigyelt, manifeszt változói közötti kapcsolatokat, interakciókat vizsgálja.

A *klasszifikációs eljárás* egy csoportosított minta egyedeire kiszámítja a megfigyelt értékek alapján a különböző csoportokhoz való tartozás valószínűségeit, így az adott mintatérben megítélhetjük a csoportbeosztás „jóságát".

A 2.1. táblázat 2. blokkjában szereplő *loglineáris modell* a két- vagy többdimenziós kereszttábla gyakoriságainak logaritmusait felhasználva elemzi a változók kapcsolatrendszerét. A loglineáris modellben feltételezhető az is, hogy a változók között kölcsönös kapcsolat van (Q_1), és az is, hogy az általánosított lineáris modellel (GLIM) írható le a változók közötti oksági kapcsolat (Q_2).

Az *útelemzés* módszere (2.5. ábra) oldja fel a regressziós modellnek a magyarázó változók közötti kapcsolat hiányára vonatkozó és a gyakorlatban elég ritkán teljesülő feltételeit. Az útelemzés a változóknak valamilyen szempontú rendezettségét tételezi fel. Ilyen rendezőelv lehet az ok-okozati kapcsolat, de a gyakorlatban leginkább az idő. Ha x_1 időben megelőzi x_2-t, akkor ebből következik, hogy x_1 egyik meghatározója x_2-nek, és nem fordítva.



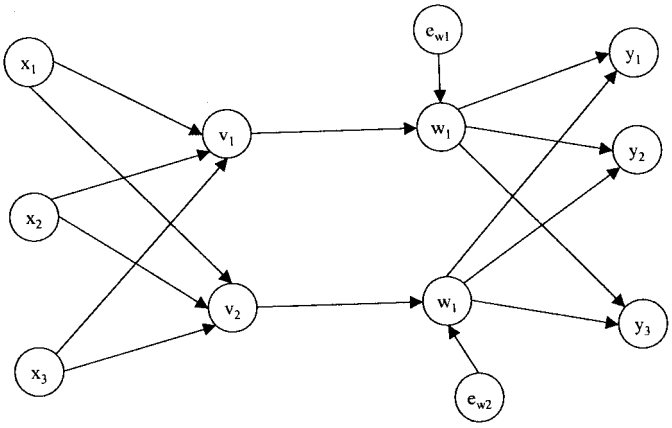
2.5. ábra. Az útelemzés alapproblémájának grafikus szemléltetése

A *kanonikus korreláció* módszere két változóhalmazt különít el, és a függő változók halmazát magyarázza a másik változóhalmazzal. A módszer ezt a kapcsolatot nem megfigyelt változókon keresztül határozza meg. A magyarázó változók halmazának azt a lineáris kombinációját keresi, amely maximálisan megmagyarázza a függő változókat azok lineáris kombinációján keresztül. A kanonikus korreláció két változóhalmaz közötti kapcsolat szorosságát méri a két változóhalmaz maximális korrelációkat adó lineáris függvényein (a kanonikus faktorokon) keresztül (2.6. ábra).

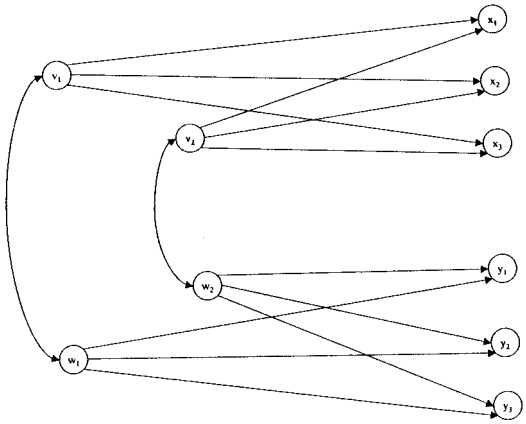
A kanonikus korreláció modellje másféle elrendezésben is ábrázolható, ha kiemeljük azt a sajátosságát, hogy kétszintű faktorelemzésnek tekinthető (2.7. ábra).

A Q_2 cellában találhatók a latens változós modellek is.

A *LISREL-modell* két latens változóhalmaz oksági kapcsolatrendszerét írja le, és a paramétereket maximum likelihood becsléssel határozza meg.



2.6. ábra. A kanonikus korreláció modelljének grafikus szemléltetése



2.7. ábra. A kanonikus korreláció mint kétszintű faktorelemzés

Az LVPLS-modell szintén feltételezi a latens változók oksági kapcsolatát, de a becsléskor a parciális legkisebb négyzetek módszerét alkalmazza.

A *többszempontú módszerek*

A Q_3 cellában található az egy- és többváltozós szóráselemzés.

A *többváltozós varianciaelemzés* valamilyen szempont szerint csoportosított mintában a csoportosító ismérv és egy adott változóhalmaz kapcsolatát vizsgálja. Arra a kérdésre ad választ, hogy a megfigyelt változók terében a csoportok átlagai és szórásai szignifikáns különbséget mutatnak-e. Megmutatja ezen kívül azt, hogy a csoporthoz való tartozást milyen arányban tudjuk magyarázni a változóhalmazzal.

A *diszkriminanciaelemzés* a többváltozós varianciaelemzés eredményeit felhasználva a változók szerepét elemzi a csoportok különválásában. A megfigyeléseket a mintatérből egy olyan diszkrimináló térbe viszi át, ahol a csoportok optimálisan elkülönülnek. A változóknak a diszkrimináns faktorokban szereplő súlyaik szerint kiválaszthatjuk a változók közül azokat, amelyek a csoportok különbözőségét határozottan magyarázzák.

A *faktoriális diszkriminanciaelemzés* a többváltozós varianciaelemzés kiterjesztése több csoportképző ismérv esetére.

A *lineáris szeparáció* az előzetesen megadott csoportosításnak legjobban megfelelő lineáris döntési függvényt állítja elő.

A *klaszterelemzés* hierarchikus változata a teljes mintából kiindulva kívánja a minta egyedeit relatíve homogén csoportokba besorolni úgy, hogy kezdetben még nem rendelkezünk csoportokkal, nem ismerjük a belső tagozódást. A nemhierarchikus klaszterezés alkalmazása során egymást át nem fedő csoportokat képezünk, úgy hogy a csoportok száma az algoritmus során alakul ki, vagy paraméterként előre megadható.

A negyedik cellában csak egy módszert említünk.

A *többszörös kovarianciaelemzés* valamilyen szempont szerint csoportosított mintában vizsgálja a csoportosítási ismérv és a függő változók egy adott halmaza kapcsolatát, miután a függő változók halmazából egy kontroll változóhalmaz hatását kiszűrtük.

3. fejezet

A változók mérési skáláiról

A következőkben a változók mérési skáláiról és a kevert változóhalmaz feldolgozását biztosító skálatranszformációkról lesz szó.[1]

3.1. A változók típusai, mérési skálák

A változók különböző csoportjait különíthetjük el *egyrészt* aszerint, hogy a méréskor a számhozzárendelés milyen tulajdonságait tekintjük érvényesnek, *másrészt* a változók értékkészlete szerint. A két szempont szerinti csoportosításnál a változóknak a gyakorlat szempontjából célszerű kétdimenziós osztályozását kapjuk.

Campbell (1938) szerint a mérés „a számjegyek hozzárendelése nem számokból álló anyagi rendszerek tulajdonságainak reprezentálására a tulajdonságokat meghatározó törvények alapján". Stevens (1951) meghatározásában a „mérés számok hozzárendelése tárgyakhoz vagy eseményekhez szabályoknak valamilyen halmaza szerint". Míg Campbell az esemény, dolog tulajdonságát, addig Stevens magát a dolgot méri. Sorolhatnánk még a részben eltérő definíciókat, azonban nem tekintjük feladatunknak a méréselmélet prob-

[1] Az olvasó ebben a fejezetben számos olyan fogalommal, eljárással találkozhat, amelyeket a könyv későbbi részeiben ismertetünk részletesen.

lémáinak vizsgálatát. Így elfogadhatjuk a bennük lévő közöset, és *mérés* alatt a számok eseményekhez való hozzárendelését értjük valamilyen szabály alapján.

A számok – amelyeket tehát az események reprezentálására kijelölünk – rendelkeznek bizonyos tulajdonságokkal:

a) a számok egymást kizáróak,

b) a számok rendezettek,

c) a számok közötti különbségek rendezettek,

d) a számsoroknak egységes kezdőpontjuk van, és ezt a nulla jelöli.

Az eseményeknek számokkal történő megfeleltetésekor biztosítanunk kell, hogy a számok közötti relációk, tulajdonságok az események, dolgok közötti relációt tükrözzék. A megfeleltetéskor tehát a számoknak csak azon tulajdonságait tekintjük érvényesnek, amelyek érvényesek a dolgokra.

Attól függően, hogy a számok tulajdonságai közül melyek érvényesek a méréskor, *skálatípusokat* különböztetünk meg.

A skálatípusok illusztrálására tekintsünk két objektumot, és jelöljük őket A-val és B-vel. Legyen x az a változó, amelyik A esetén x_A és B esetén x_B értéket vesz fel.

Nominális (névleges) skála

A nominális skála csupán megkülönbözteti ezeket az objektumokat, köztük csak az azonosság vagy a különbözőség viszonyát tételezi fel. Vagyis A-ról és B-ről csak annyit tudunk mondani, hogy $x_A = x_B$ vagy $x_A \neq x_B$.

Ordinális (rendező) skála

Az ordinális skála definiálja az objektumok viszonylagos helyét is, megvalósítja az objektumok rendezését. Vagyis azon túl, hogy különbséget teszünk $x_A = x_B$ és $x_A \neq x_B$ között, azt is mondhatjuk, hogy $x_A > x_B$ vagy $x_A < x_B$.

A nominális, illetve ordinális skálákon mért változókat szokás kategorikus változóknak is nevezni.

Intervallumskála

Az intervallumskálánál a különbségek mértékét is értelmezhetjük. Vagyis az $x_A > x_B$ vagy $x_A < x_B$ rendezésen túl azt is mondhatjuk, hogy A a B-től $x_A - x_B$ egységgel különbözik.

Arányskála

Az arányskála az intervallumskála tulajdonságain túl még értelmezhető kezdőponttal, 0 ponttal is rendelkezik, vagyis ha $x_A > x_B$, akkor nemcsak azt mondhatjuk, hogy A $x_A - x_B$ egységgel nagyobb, hanem azt is, hogy $\frac{x_A}{x_B}$ -szer nagyobb, mint B.

Ez az osztályozás hierarchikus felépítésű, mivel minden skála rendelkezik minden őt megelőző skála tulajdonságával, így pl. az intervallumskála rendelkezik az ordinális és nominális skála tulajdonságaival is. A nominális és ordinális skálán mért változókat gyakran nevezik kategorikus változóknak vagy minőségi változóknak, az intervallum- és arányskálán értelmezett változókat pedig mennyiségi változókként említik.

A változó értékkészlete szerinti osztályozás

Különbséget tehetünk változók között aszerint, hogy hány különböző értéket vehetnek fel, vagyis hogy hány elem tartozik az értékkészletükbe. Ahhoz, hogy a változókat az értékkészlet szerint osztályokba soroljuk, szükségesek a következő fogalmak:

Két halmaz *ekvivalens* (vagy másképpen: egyenlő számosságú), ha az egyik halmaz elemei kölcsönösen egyértelmű módon megfeleltethetők a másik halmaz elemeinek.

a) Egy halmazt *végesnek* nevezünk, ha kölcsönösen egyértelmű módon megfeleltethető n természetes számmal, vagyis ha találunk olyan n természetes számot, amellyel ekvivalens. Ekkor a halmaz elemeit megszámolhatjuk, és a halmaz elemeinek száma véges egész szám.

b) *Megszámlálhatóan végtelennek* nevezünk egy halmazt, ha ekvivalens az összes természetes számok halmazával. Ez a halmaz megszámlálható, de nem adható meg olyan véges természetes szám, amely a halmaz elemeinek a számát jelölné.

c) *Nem megszámlálhatóan végtelennek* azt a halmazt nevezzük, amelynek elemei nem állíthatók kölcsönösen egyértelmű megfeleltetésbe az egész számokkal. Ebben az esetben nem tudunk következő számról beszélni, hiszen minden szám és a „következő" között végtelen sok szám van, vagyis a halmaz elemei nem megszámlálhatóak. Ilyen pl.: a számegyenes reprezentálta valós számok halmaza.

Felhasználva ezeket a fogalmakat, a változók értékkészlete szerint a következő osztályokat képezhetjük:

– *folytonos változó*, amelynek értékkészlete nem megszámlálhatóan végtelen halmaz;

– *diszkrét változó*, értékkészlete véges vagy megszámlálhatóan végtelen;

– *bináris* vagy *dichotom* az a diszkrét változó, amely csak két értéket vehet fel.

A változók kétdimenziós osztályozását mutatja a 3.1. táblázat.

A táblázat együttesen mutatja a változók két osztályozási szempontját, és példákat mutat a közös cellákban. Anélkül, hogy minden cella elemzésébe részletesen belemennénk, nézzük a folytonos intervallumváltozó esetét. A Celsius hőmérsékleti skálán a 0 pont a víz fagyáspontja, a 100 pont a víz forráspontja, a köztük lévő hőterjedelem 100 egyenlő intervallumra van felosztva. A szintén hőmérsékleti Fahrenheit-skála kiindulópontja a szalmiák és jég egyenlő súlyú keverékének hőmérséklete. A Celsius-skála 0 pontjának a Fahrenheit-skála 32 hőmérsékleti értéke felel meg (0 °C = 32 °F). Ezért a 20 °C-os hőmérsékletről nem mondhatjuk, hogy kétszer olyan magas, mint a 10 °C-os, mivel a Fahrenheit-skálán ugyanennek a két hőmérsékletnek az aránya kb. 7 : 5. Azt viszont mondhatjuk, hogy 20 °C *távolsága* a 0 °C-tól kétszer akkora, mint 10 °C-é a 0 °C-tól. Ez igaz a Fahrenheit-skálán is.

A Kelvin-féle hőmérsékleti skála 0 pontja az abszolút zérus, így mondhatjuk, hogy 50 °K kétszer akkora, mint 25 °K.

A bináris változók speciális szerepe

A bináris változók külön szerepeltetése erőltetettnek tűnik, mivel a diszkrét változók csoportjába tartoznak azzal a specialitással, hogy csak két értéket vesznek fel. Ez a specialitás viszont olyan rugalmasságot kölcsönöz, ami miatt célszerű őket külön tárgyalni.

	Az értékkészlet		
Mérési skála típusa	folytonos: nem megszámlálhatóan végtelen értéket vehet fel	diszkrét: véges vagy megszámlálhatóan végtelen számú értéket vehet fel	bináris: csak két értéket vehet fel
Nominális $x_A = x_B$ vagy $x_A \neq x_B$	nem lehet nem megszámlálhatóan végtelen sok különböző osztályt megadni	születési hely, foglalkozás	nő – férfi igen – nem igaz – hamis
Ordinális $x_A > x_B$ vagy $x_A = x_B$ vagy $x_A < x_B$	szubjektív ítéletek különböző dolgok intenzitásáról (fényességről, hangintenzitásról)	nagyon jó – jó – közepes – rossz – nagyon rossz kategóriák	jó – rossz nagy – kicsi magas – alacsony széles – keskeny
Intervallum ha $x_A > x_B$, akkor $x_A - x_B$ egységgel nagyobb A, mint B	hőmérséklet °C Celsius-skála °F Fahrenheit-skála	sorszámok	két különböző érmével működő automata
Arány ha $x_A > x_B$, akkor $\frac{x_A}{x_B}$-szer nagyobb A, mint B	hőmérséklet °K Kelvin-skála	kórházak száma, gyerekek száma, egyéb számlálások	a magántulajdonban lévő üdülőtelkek megengedett száma Magyarországon családonként a szocializmusban

3.1. táblázat.

Nézzük meg, miben is áll ez a rugalmasság. Jelölje egy bináris változó két különböző lehetséges értékét u és v.

Az $f(u) = a+bu$ lineáris transzformációval bármely más x és y számpár előállítható a következő módon:

$$x = a + bu \quad \text{és} \quad y = a + bv.$$

Az a és b együtthatók:

$$b = \frac{x - y}{u - v} \quad \text{és} \quad a = x - bu = y - bv.$$

Tehát minden olyan leképezés, amely az u és v értéket az x és y értékekre képezi le, ekvivalens egy lineáris transzformációval. Azokban az esetekben, amikor az alkalmazott elemzési technikák invariánsak a lineáris transzformációkkal szemben, a bináris változó két kategóriájához rendelt pontszám tetszés szerinti lehet, így a bináris változók esetében csak a nominális skálák létezését tételezzük fel.

Az ismert elemzési módszerek közül a lineáris transzformációval szemben invariánsak a következők:

– korrelációelemzés,

– regresszióelemzés (lineáris),

– diszkriminanciaelemzés,
– faktorelemzés,
– kanonikus korrelációelemzés,
– a bináris változókon alapuló módszerek.

Ezeket a módszereket intervallumskálákon mért változókkal használhatjuk. A bináris változók fenti tulajdonsága miatt a bináris változókként definiált kvalitatív változók közvetlenül felhasználhatók a fenti, intervallummérési szintet feltételező elemzési módszerekben.

3.2. A mérési skálák transzformálása

A többváltozós módszereknél feltételezzük, hogy a változók azonos típusúak, általában intervallumskálán mért változók. A gyakorlati adathalmazoknál ez a feltétel legtöbbször nem teljesül, vagyis eltérő mérési szintű változókkal kell dolgozni. Ilyenkor különböző meggondolásokkal hidalhatjuk át a problémát.

Például, ha a változókat skálatípusok szerint csoportosítjuk, az így kapott csoportokkal külön-külön, a mérési skálának megfelelő elemzési technikát használhatjuk. Most azonban az érdekel bennünket, hogyan lehet különböző skálán mért változókat azonos mérési szintre hozni, és így együttesen szerepeltetni a különböző technikákban.

A skálák átalakításait, a skálatranszformációkat célszerű két csoportra osztani aszerint, hogy hierarchikus rendszerünkben milyen irányú a transzformáció. Így ha nominálisról ordinálisra, ordinálisról intervallumra transzformáljuk a skálákat, akkor úgymond *felértékeljük* őket. Mivel az egymás feletti mérési skálák meghatározása egyre több információt igényel, a felértékelésnél szükségünk van pótlólagos információra vagy újabb feltevések elfogadására.

Ha fordítva járunk el, vagyis *leértékeljük* a skálát, elhagyunk olyan információt, amit az adott skála meghatározásakor figyelembe vettünk. Hogy konkrét esetben melyik mérési skálához igazítjuk a többi változó skáláját, azt mindig több tényező együttes mérlegelése döntheti csak el.

A következőkben szereplő skálatranszformációs technikák csupán egy-egy lehetőséget biztosítanak nekünk, a közülük történő választás problémája azonban mindig megmarad. Ahogyan Anderberg (1973) kifejti: „a jól tájékozott megítélés alternatívát jelent valamennyi ilyen technikával szemben, sőt, néha az egyetlen eszközt adja arra, hogy jelentős, de nem számszerűsíthető szempontokat figyelembe vehessünk”.

A továbbiakban az értékkészlet szerinti megkülönböztetésre nem térünk ki, valamint az arányskálával sem foglalkozunk, mivel a tárgyalt elemzési technikák nem igénylik az arányskálát. A gyakorlatban az arányskála 0 pontjára vonatkozó információtól eltekintünk, így azt intervallumskálaként kezeljük.

A skálaértékelési módokat a következő három pontban tárgyaljuk.

3.2.1 Áttérés intervallumskálárób rdinális skálára

A változó értékeit olyan intervallumokba kell besorolni, olyan kategóriákat kell definiálni, ahol adott kategórián belül nem teszünk különbséget az egyedek között, és a kategóriák között csak a rendezést írjuk elő. Vagyis eltekintünk *egyrészt*:

– az azonos intervallumba eső objektumok közötti különbségektől, *másrészt*

– a különböző intervallumokba eső objektumok közötti különbségek nagyságrendjétől.

A legtöbb technika ezt a kétféle információveszteséget minimalizálja.

Nézőpont változtatása

Néhány esetben nincs is szükség technikai transzformációra, csupán a nézőpontot kell megváltoztatni. Például az életkor esetében, amikor az életkort években, diszkrét értékekkel mérjük. Itt problémát jelenthet az életkor-intervallumok száma és az egyenlő hosszúságú intervallumok választása. Az életkor intervallumainak számát és az intervallumok hosszát a minta és az elemzés együttes szempontjai dönthetik el.

Helyettesítés

Más esetekben célszerű lehet adott változó helyett egy olyan változó keresése és az adott változóval való helyettesítése, amely az eredetit jól jellemzi, vele szoros kapcsolatban áll, ugyanakkor a kívánt skálán mértük. Például ha a legmagasabb iskolai végzettség évei száma helyett a következő kategóriákat tartalmazó változót mérjük:

– általános iskolát végzett,

– ipari iskolát végzett,

– középiskolát végzett,

– főiskolát végzett,

– egyetemet végzett.

Egyenlő hosszúságú kategóriák

A feladat ilyenkor az, hogy a változó értékkészletét egyenlő hosszúságú intervallumokra osszuk fel. Problémát jelenthet az intervallumok számának a meghatározása. Ha előre nem adott az intervallumok száma, akkor úgy járunk el, hogy különböző intervallumszámra megszerkesztjük a hisztogramokat, és az aszerinti csoportosítást fogadjuk el, amelyiknél a hisztogram a legkielégítőbbnek bizonyult. A hisztogram szerkesztése úgy történik, hogy a változó értékkészletét felosztjuk egyenlő hosszúságú intervallumokra. ($\triangle x$ jelentse az intervallum hosszát). A változó minden megfigyelési értékét besoroljuk a megfelelő intervallumba. Jelölje az i-edik intervallumba eső értékek számát f_i. Ezután minden $\triangle x$ szakasz fölé olyan téglalapot rajzolunk, amelynek magassága $\frac{f_i}{n \cdot \triangle x}$ ($i = 1, \ldots, g$) és természetesen az alap $\triangle x$. A kapott téglalapok területei megegyeznek az osztályközök relatív gyakoriságaival. (A hisztogramok készülhetnek különböző hosszúságú $\triangle x_i$ intervallumokkal is.)

Azonos elemszámú kategóriák

Ha az előző módszernél az egyes kategóriákhoz rendelt esetek száma (az egyes intervallumokba eső elemek száma) nagyon eltérő, akkor az intervallumok hosszának változtatásával próbálkozhatunk. Ha feltehetjük, hogy mintánk egyenletes eloszlású, vagyis hogy az egyes intervallumokba esésnek azonos a valószínűsége, akkor az adatokat egyenlő tagszámú intervallumokba osztjuk szét.

Egydimenziós hierarchikus összekapcsolási módszerek

Az intervallumskála átalakítása ordinális mérési skálára végeredményben azt jelenti, hogy a megfigyeléseket intervallumokba, csoportokba soroljuk. A klaszterelemzés módszere éppen a megfigyelések csoportokba sorolását végzi el. Ennek a skálatranszformációnak a problémája valójában tehát megfelel az egydimenziós klaszterelemzés problémájának, így a klaszterelemzés módszerei felhasználhatók az intervallumskála leértékelésére.

A csoportokba rendezéshez definiálni kell a megfigyelések közötti távolság fogalmát:

$$d = |x_j - x_k|,$$

ami két pont (szám) különbségének abszolút értéke.

A következőkben három eljárást említünk meg.

A) Egyszerű összekapcsolás módszere (legközelebbi szomszéd módszere)

Ez a legegyszerűbb hierarchikus csoportosító módszer. A megfigyelési egységek összekapcsolása a csoportok legközelebbi elemei között lévő távolság szerint történik. A módszert a következő lépésekben lehet végrehajtani:

1. a megfigyeléseket növekvő sorrendbe rendezzük. Ebben a fázisban minden megfigyelést egyelemű csoportnak tekintünk:
2. megvizsgálunk minden szomszédos (csoport) párt, és megkeressük azt a kettőt, amelyek a legközelebb állnak egymáshoz. A köztük lévő távolságnak a legközelebbi elemeik közötti távolságot tekintjük;
3. a 2. lépést mindaddig ismételjük, amíg már csak egyetlen csoport marad.

Az eredményt dendrogrammal ábrázolhatjuk, és a csoportok számát ennek alapján határozhatjuk meg.

B) Komplett összekapcsolás módszere (legtávolabbi szomszéd módszere)

Az algoritmus lényegében megegyezik az előző eljárással, a különbség csupán az, hogy a csoportok közötti távolságot a legtávolabbi elemeik közti távolságként definiáljuk.

Meg kell jegyezni, hogy a két módszer nem feltétlenül ad azonos eredményt.

C) Ward hierarchikus csoportosító módszere

Ward abból az információveszteségből indult ki, amely a megfigyeléseknek csoportokba történő összevonásából ered. Ezt az információveszteséget úgy definiálta, mint a megfigyelések csoportátlagoktól való eltérései négyzetének összegét.

Jelöljük T-vel a teljes minta eltérés-négyzetösszegét. Érvényes a következő felbontás:

$$T = B + K.$$

$$\sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij} - \overline{x})^2 = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_i)^2 + \sum_{i=1}^{g} n_i(\overline{x}_i - \overline{x})^2.$$

ahol B a csoporton belüli eltérések négyzetösszege.
K a csoportok közötti eltérések négyzetösszege.
g a csoportok száma.
n_i az i-edik csoport elemszáma.
x_{ij} az i-edik csoportba tartozó j-edik megfigyelés.
$\overline{x}_i$ az i-edik csoport átlaga.
$\overline{x}$ a teljes minta átlaga.

Adott mintánál a T egyértelműen meghatározott, B és a K a csoportosítástól függően változik. A cél olyan csoportosítás létrehozása, ahol B minimális.

Az algoritmus a következő lépésekre bontható:

a) n számú csoporttal kezdjük, akkor minden csoport egyelemű, így $B = 0$;

b) azt a két csoportot vonjuk össze, amely esetén B minimálisan növekszik;

c) a *b)* lépést mindaddig folytatjuk, amíg egyetlen csoportot nem kapunk.

A Ward-technika fogyatékossága, hogy nem adja minden esetben a B minimális értékét. Így előfordulhat, hogy a háromcsoportos felbontás minimális B értéket adó változata a következő lépésben, két csoport esetén nem lesz optimális.

A felosztás iteratív javítása

Ha kiindulunk a minta egy adott csoportosításából, a megfigyeléseknek egyik csoportból másik csoportba való átsorolásával javíthatjuk a felosztást. A Ward-féle eljárásnál szereplő B a csoportátlagokhoz viszonyított eltérések négyzetösszege. A B függvény minimalizálásánál tulajdonképpen a legkisebb négyzetek elvét alkalmazzuk.

A g csoportos felosztás esetén a feladat $g - 1$ csoportközi határ meghatározása. Egyszerű módszer a következő:

a) választunk egy kiinduló osztályközi határt (pl. a bal szélső határt);

b) megpróbáljuk a határt egy megfigyeléssel jobbfelé elmozdítani. Ha ez csökkenti a B értékét, próbáljuk meg balfelé elmozdítani. Ha bármely irányba történő elmozdítás csökkenti a B értékét, akkor folytassuk az eljárást, 1–1 megfigyeléssel tovább mozgatva a szóban forgó pontot mindaddig, amíg javulást nem tudunk elérni;

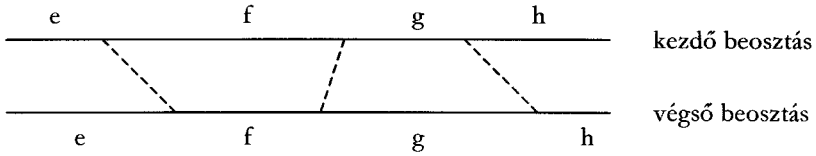
c) ismételjük meg a *b)* lépést az összes többi határ esetében is;

d) ismételjük a *b)–c)* lépéseket egészen addíg, amíg bármely megfigyelésnek bármely irányba történő elmozdítása már nem csökkenti a csoporton belüli eltérést.

Ezt az eljárást mutatja négy csoport esetén a 3.1. ábra:

Mivel az egy megfigyelést tartalmazó csoport a B összeghez nullával járul hozzá, ez az eljárás változatlanul hagyja a csoportok számát.

A módszer egyszerűsége, hogy csak egy megfigyelést vizsgál egyszerre. Ez egyben hátránya is, mivel így csak lokális minimumok meghatározására képes. Megemlíthető,



3.1. ábra. A felosztás iteratív javítása

hogy ennek a módszernek létezik egy alternatív formája is. A B minimalizálásával ekvivalens ugyanis K maximalizálása. Számítástechnikai szempontokból talán célszerű ezt az alternatívát használni. Ugyanis a

$$K = \sum_{i=1}^{g} n_i(\overline{x}_i - \overline{x})^2 \longrightarrow \max$$

iteráció alkalmazásával csak két csoportot kell vizsgálni, tudniillik egy osztályhatárt mozgatunk. Ilyenkor az

$$n_1(\overline{x}_1 - \overline{x})^2 + n_2(\overline{x}_2 - \overline{x})^2$$

összeg változását kell figyelni egy megfigyelés átsorolásánál. Ha az x megfigyelést az első csoportból átsoroljuk a másodikba, akkor nő K értéke, ha

$$\frac{(n_1\overline{x}_1 - x^0)^2}{n_1 - 1} + \frac{(n_2\overline{x}_2 - x^0)^2}{n_2 + 1} > \frac{(n_1\overline{x}_1)^2}{n_1} + \frac{(n_2\overline{x}_2)^2}{n_2}.$$

3.2.2. Áttérés intervallumskáláról nominális skálára

Lényegében hasonló módon járhatunk el, mint ordinális skálára való átalakításkor. A fenti módszerek használatakor azt kell figyelembe venni, hogy nominális skála esetén eltekintsünk a rendezéstől, így egymással nem határos osztályokat kell megadni.

Az előzőektől különböző, egyszerű módszer lehet az, amikor a megfigyeléseket az átlaghoz való viszonyuk szerint csoportosítjuk oly módon, hogy az átlag ($\overline{x}$) környezetébe esők kerülnek egy osztályba, egy következő osztályba kerülnek azok a megfigyelések, amelyek $\triangle x$-nél nagyobb, de $2 \triangle x$-nél kisebb értékkel térnek el az átlagtól (pozitív és negatív irányban egyaránt), és így tovább attól függően, hogy hány csoportot akartunk képezni. Így alakíthatjuk ki pl. az átlagos (normális), mérsékelten eltérő és szélsőséges kategóriákat.

A leírt eljárást nevezhetjük „központi szélsőséges” elvnek is.

3.2.3. Áttérés ordinális skáláról nominális skálára

Az ordinális skála rendezési tulajdonságáról is lemondhatunk a központi szélsőséges elv alkalmazásával.

Ha az osztályokat a következőképpen jelöljük:

$$A > B > C > D > E,$$

akkor a nominális skála osztályai lehetnek:

C normális,

$B \cup D$ mérsékelten eltérő,

$A \cup E$ szélsőséges.

A rendezési tulajdonságoktól eltekintő osztályokra való áttérésre általános elveket nehéz megállapítani, konkrét esetben a kutató ötletétől függ az ilyen osztályok meghatározása.

A következőkben a skálafelértékeléseket tekintjük át.

3.2.4. Áttérés nominális skálárób rdinális skálára

Ahhoz, hogy a nominális változó rendezetelen kategóriáit átalakíthassuk ordinális skálára, új információra van szükség. Eltekintve attól a két esettől, amikor *vagy* egyszerűen a kérdéses változó helyett egy vele szoros kapcsolatot mutató ordinális változót vezetünk be, *vagy* amikor a központi szélsőséges elvet megfordítva alkalmazzuk, a szélsőséges kategóriát felbontjuk egy magas és egy alacsony részre, szükségünk van egy referenciaváltozóra, amelyhez fűződő kapcsolat alapján rendezzük a nominális változó kategóriáit.

Korreláció egy intervallumváltozóval

Ha a nominális változó kategóriáihoz rendelt pontértékek felhasználásával kiszámítjuk a nominális változó és egy referenciaváltozóként szereplő intervallumváltozó közötti korrelációs együtthatót, kérdés lehet, hogy a nominális változó kategóriáihoz rendelt milyen pontértékek mellett lesz maximális a korreláció.

Később, amikor a nominális skálát intervallumskálára alakítjuk át, látni fogjuk, hogy a maximális korrelációt biztosító pontszámokat a referencia-intervallumváltozónak a nominális változó szerinti osztályátlagai adják. Ordinális skálára való transzformáláskor csak a rendezést kell biztosítanunk, így az optimális pontoknál eltekintünk a nagyságrendi különbségek értelmezésétől.

Rangkorreláció alkalmazása

Ha a referenciaváltozó ordinális (y), akkor a nominális változó (x) kategóriáihoz rendelt pontszámok felhasználásával rangkorrelációt számíthatunk az x és y változók között.

Ha a nominális változó g kategóriát tartalmaz, a *Spearman-féle rangkorreláció:*

$$r = 1 - 6 \cdot \frac{\sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_i - y_{ij})^2}{n(n^2 - 1)}.$$

Keressük az i-edik kategóriához azt az x_i rangszámot, amely mellett a rangkorreláció maximális lesz. A korreláció akkor lesz maximális, amikor a kifejezésben a számláló minimális. Minimalizálnunk kell a

$$\sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_i - y_{ij})^2 = \sum_{i=1}^{g} \left(n_i x_i^2 - 2 n_i x_i \overline{y}_i + \sum_{j=1}^{n_i} y_{ij}^2 \right)$$

kifejezést. Mivel az y referenciaváltozó rangszámai adottak, a harmadik tagot elhagyhatjuk. Így a

$$C = \sum_{i=1}^{g} n_i x_i (x_i - 2\overline{y}_i)$$

kifejezést kell minimalizálni. E függvénynek a minimuma ott lehet, ahol az x_i szerint a parciális deriváltak egyenlőek 0-val.

Eredményül azt kapjuk, hogy a rangkorreláció akkor lesz maximális, ha a nominális változó kategóriái a referenciaváltozó rangszámainak a kategóriaátlagát kapják rangszámnak, vagyis az i-edik rangszám

$$x_i = \overline{y}_i .$$

Mivel az ordinális skálán a rendezés tulajdonságát kell biztosítani, elégséges azt mondani, hogy az x_i rangszámok sorrendjének meg kell egyeznie az $\overline{y}_i$ rangszámátlagok sorrendjével.

3.2.5. Áttérés ordinális skáláról intervallumskálára

Az ordinális skálán mért változó kategóriái rangszámot kapnak. Amennyiben a rangszámok olyanok, hogy a rendezés mellett a kategóriák közötti nagyságrendi különbségeket is kifejezik, használhatjuk őket, mint egy intervallumváltozó lehetséges értékeit. Megoldhatjuk a problémát úgy is, hogy az intervallumskálán a kategóriáknak megfelelő számú pontot jelölünk ki úgy, hogy az osztályok rendezettsége megmaradjon.

A rangszámok

A rangszámok valójában felfoghatók az intervallumskála azonos hosszúságú osztályokra történő felosztásának. Ez a fordítottja annak, amit az intervallumváltozó leértékelésénél az egyenlő hosszúságú kategóriák módszerénél láttunk. A rangszámokat így közvetlenül használjuk az intervallum mérési szintre kidolgozott korrelációs együttható számításához.

Elméleti eloszlás feltételezése

Tételezzük fel, hogy a mintát valamilyen eloszlású sokaságból vettük. Ha n_k jelöli az első k osztályba eső megfigyelések számát (kummulált gyakoriság), az alapsokaság első k osztályába eső elemeinek arányát becsülhetjük az első k osztály relatív arányával:

$$p_k = n_k / n.$$

Ha a feltételezett eloszlás eloszlásfüggvénye $F(x)$, a k-adik osztály felső határának értékét, x_k-t a

$$p_k = F(x_k)$$

egyenletből számolhatjuk.

Így megkaphatjuk egy feltételezett eloszlás segítségével az osztályok alsó és felső határát. Az így kiszámított intervallumnak átlagát vagy mediánját vesszük, és azt az adott

osztályhoz rendeljük, ezen az úton az osztályok nagyságának eltéréseit is jellemző értékekhez jutunk:

$$\overline{x}_k = \frac{\int_{x_{k-1}}^{x_k} xf(x)dx}{F(x_k) - F(x_{k-1})}.$$

Standard normális eloszlás esetén

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},$$

így

$$\overline{x}_k = -\frac{1}{p_k - p_{k-1}} \int_{x_{k-1}}^{x_k} \frac{1}{\sqrt{2\pi}} t e^{-\frac{t^2}{2}} dt = \frac{1}{p_k - p_{k-1}} \left(e^{-\frac{x_{k-1}^2}{2}} - e^{-\frac{x_k^2}{2}} \right),$$

egyszerűbb formában számolható.

A számolás menete tehát a következő: Meghatározzuk az egyes osztályokhoz tartozó kumulált gyakoriságot (n_k), majd a relatív gyakoriságot ($p_k = n_k/n$). Ezután kiszámítjuk az ezekhez tartozó intervallum felső határát (x_k), azután pedig az intervallumok átlagát. Esetenként az így kapott osztályértéket standardizáljuk. Az átlag számítása helyett alternatívaként a mediánt (m_k) is számíthatjuk a következő egyenletből:

$$\frac{p_k + p_{k+1}}{2} = F(m_k).$$

Korreláció egy referenciaváltozóval

A korábbiakhoz hasonlóan a skálatranszformációhoz új információként itt is felhasználhatunk a felértékelendő változóhoz szorosan kapcsolódó referenciaváltozót. Ebben az esetben keressük azokat az optimális pontszámokat, amelyekkel a referenciaváltozó maximálisan korrelál. A következő részben be is bizonyítjuk, hogy ha a referenciaváltozó intervallum szintű, akkor az optimális pontszámok a referenciaváltozó egyes kategóriákba eső értékeinek az átlagai. Ha ezek az osztályátlagok nem felelnek meg az ordinális változó korábbi rendezésének, akkor más olyan maximális korrelációt biztosító pontszámokat kell keresni, amelyek a rendezés feltételét is kielégítik.

3.2.6. Áttérés nominális skáláról intervallumskálára

A mérési skálák közül a legkevesebb információt tartalmazó nominális skáláról a jóval több tulajdonsággal rendelkező intervallumskálára áttérni nyilvánvalóan a legnehezebb. Egyrészt biztosítanunk kell a rendezetlen kategóriák rendezését, másrészt meg kell határoznunk az osztályközöket, a rendezett kategóriák eltérését. A feladatnak ez a kettőssége érzékelhetően sejteti a kétlépcsős megoldást. Először ordinális skálára kell átalakítani a nominális skálát, és utána az ordinális skálát transzformálni intervallum jellegűre. Mivel az előzőekben többször is említettük, hogy a rangszámok használata majdnem olyan jó, mint más pontszámé, a legnagyobb gonddal éppen a rendezést kell elvégezni, és ez néha elég is.

A továbbiakban olyan módszereket tekintünk át, amelyek intervallummérési szintű pontszámokat rendelnek a nominális kategóriákhoz.

Korreláció egy intervallumváltozóval

A referenciaváltozó szerepét töltse be egy intervallumváltozó, amelynek értékkészlete legyen diszkrét, vagy soroljuk kategóriákba értékeit. Ekkor a kiinduló adatrendszer a szokásos kontingenciatáblázatba rendezhető:

	1	2	...	q	$\sum$	kategóriaérték
1	f_{11}	f_{12}	...	f_{1q}	f_{10}	x_1
2	f_{21}	f_{22}	...	f_{2q}	f_{20}	x_2
$\vdots$	$\vdots$	$\vdots$				
p	f_{p1}	f_{p2}	...	f_{pq}	f_{p0}	x_p
$\sum$	f_{01}	f_{02}	...	f_{0q}	n	
kategóriaérték	y_1	y_2	...	y_q		

Vagy mátrixaritmetikai jelölésekkel:

		$\sum$
	$\mathbf{F}$	$\mathbf{F1}$ $\mathbf{x}$
$\sum$	$\mathbf{1'F}$ $\mathbf{y'}$	

A táblázatból a következő statisztikai jellemzőket számolhatjuk ki:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{p} f_{i0}x_i, \quad \overline{y} = \frac{1}{n}\sum_{j=1}^{q} f_{0j}y_j$$

átlagok,

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{p} f_{i0}(x_i - \overline{x})^2, \quad \sigma_y^2 = \frac{1}{n}\sum_{j=1}^{q} f_{0j}(y_j - \overline{y})^2$$

szórásnégyzetek; és a korrelációs együttható:

$$r = \frac{\sum_{i=1}^{p}\sum_{j=1}^{q} f_{ij}(x_i - \overline{x})(y_j - \overline{y})}{\left\{\left[\sum_{i=1}^{p} f_{i0}(x_i - \overline{x})^2\right]\left[\sum_{j=1}^{q} f_{0j}(y_j - \overline{y})^2\right]\right\}^{1/2}}.$$

Célszerű az x és y pontszámokat standardizálni, mivel így egyszerűbb kifejezéseket kapunk. Ekkor

$$\overline{x} = \frac{1}{n}\mathbf{x}'\mathbf{F}\,\mathbf{1} = 0\,; \quad \overline{y} = \frac{1}{n}\mathbf{1}'\mathbf{F}\,\mathbf{y} = 0 \tag{3.1}$$

és

$$\sigma_x^2 = \frac{1}{n}\mathbf{x}'\,\langle\mathbf{F}\,\mathbf{1}\rangle\,\mathbf{x} = 1, \quad \sigma_y^2 = \frac{1}{n}\mathbf{y}'\,\langle\mathbf{1}'\mathbf{F}\rangle\,\mathbf{y} = 1. \tag{3.2}$$

A korreláció standardizált pontszámok esetén:

$$r = \frac{\sum_{i=1}^{p}\sum_{j=1}^{q} f_{ij}x_i y_j}{n} = \frac{1}{n}\mathbf{x}'\mathbf{F}\,\mathbf{y}. \tag{3.3}$$

Feladatunk megkeresni a maximális korrelációt biztosító x_i pontszámokat. Mivel r maximálása ekvivalens r^2 maximálásával, a Lagrange-függvény

$$L=\left[\frac{1}{n}\mathbf{x}'\mathbf{F}\,\mathbf{y}\right]^2 -\lambda_1\left(\frac{1}{n}\mathbf{x}'\mathbf{F}\,\mathbf{1}\right) -\lambda_2\left(\frac{1}{n}\mathbf{1}'\mathbf{F}\,\mathbf{y}\right) -\lambda_3\left(\frac{1}{n}\mathbf{x}'\,\langle\mathbf{F}\mathbf{1}\rangle\,\mathbf{x}-1\right) -\lambda_4\left(\frac{1}{n}\mathbf{y}'\,\langle\mathbf{1}'\mathbf{F}\rangle\,\mathbf{y}-1\right). \tag{3.4}$$

A függvény maximumát ott találjuk, ahol a parciális deriváltak egyenlőek 0-val:

$$\frac{\partial L}{\partial \mathbf{x}'} = 2\left(\frac{1}{n}\mathbf{x}'\mathbf{F}\,\mathbf{y}\right)\left(\frac{1}{n}\mathbf{F}\,\mathbf{y}\right) - \lambda_1\left(\frac{1}{n}\mathbf{F}\,\mathbf{1}\right) - \lambda_3 2\left(\frac{1}{n}\,\langle\mathbf{F}\,\mathbf{1}\rangle\,\mathbf{x}\right) = \mathbf{0}. \tag{3.5}$$

$$\mathbf{x} = \frac{1}{r}\,\langle\mathbf{F}\,\mathbf{1}\rangle^{-1}\,\mathbf{F}\,\mathbf{y}. \tag{3.6}$$

Ha az $\mathbf{y}$ vektor elemeit ismerjük és azok standardizáltak, a nominális változó kategóriáinak optimális pontszámát a (3.6) egyenlet alapján számíthatjuk ki. Ha $\mathbf{y}$ nem standardizált, a (3.6) egyenletbe y_i helyett

$$\frac{y_i - \overline{y}}{\sigma_y}$$

kifejezést kell írni. Az optimális pontszámokat adó (11.6) egyenlet akkor átalakítva:

$$x_i = \frac{1}{r}\,\frac{1}{f_{i0}}\sum_{j=1}^{q} f_{ij}\,\frac{y_j - \overline{y}}{\sigma_y} = \frac{1}{r\sigma_y}\left[\frac{1}{f_{i0}}\sum_{j=1}^{q} f_{ij}(y_j - \overline{y})\right] = x_i = \frac{\overline{y}_j - \overline{y}}{r\sigma_y},$$

ahol $\overline{y}_j$ az $\mathbf{y}$ intervallumváltozó átlaga a nominális változó j-edik kategóriájában.

Mivel a korreláció invariáns a lineáris transzformációra – és x_i az $\overline{y}_j$ lineáris transzformációja –, a korreláció értékét nem változtatja meg, ha az $x_i = \dfrac{\overline{y}_j - \overline{y}}{r\sigma_y}$ pontszám helyett egyszerűen az $x_i = \overline{y}_j$ pontszámot használjuk optimális értékként. Eszerint azt mondhatjuk, hogy egy nominális és egy intervallumváltozó között a korreláció akkor maximális, ha a nominális változó kategória-pontszámainak az intervallumváltozó kategóriaátlagait feleltetjük meg.

Kanonikus korreláció a nominális skála felértékelésére

Az előző pontban a transzformációt egy intervallummérési szintű változó ismeretében végeztük el, és az optimális megoldást a maximális korrelációt adó pontszámok adták. Változók között maximális korrelációt kereső módszer a kanonikus korreláció számítása is.

Keressünk egy p kategóriát tartalmazó nominális változóhoz referenciaváltozóként egy q kategóriát tartalmazó nominális változót. Definiáljuk a megfigyelések adott osztályba esését dichotom változóként. Eszerint x_i legyen 1, ha a megfigyelés az i-edik osztályba esik ($i = 1, \ldots, p$), 0 különben. A referenciaváltozó szerint a j-edik osztályba esést y_j jelölje.

A két nominális változóból ezzel a dichotomizálással egy p, illetve egy q számú bináris változóból álló változóhalmazt hoztunk létre. Két változó közötti korreláció számítása így átalakult két változóhalmaz közötti kanonikus korreláció számításává. Keressük a kategóriák bináris változóinak olyan súlyú lineáris kombinációját, amelyek esetén a két változóhalmaz közötti korreláció maximális. Ezeket az optimális súlyokat a nominális kategóriákhoz rendelhetjük, amivel a változó rendezetlen osztályaiból egy rendezett és a térben elhelyezett intervallummérési szintű változóhoz jutottunk.

Láttuk, hogy a bináris változók használatának nagy előnye, hogy mérési skálától függetlenül intervallumváltozóknak tekinthetők minden, a lineáris transzformációra invariáns módszer esetén. Attól az esettől eltekintve, amikor egy változót úgy dichotomizálunk, hogy a kategória számával megegyező számú, a kategóriákhoz való tartozást vagy nem tartozást kifejező bináris változókat definiálunk, a dichotomizálás a skálatranszformációk egy speciális esetét adja, így alkalmazhatók a korábbi módszerek.

Így pl. egy ordinális változó dichotomizálásánál felhasználhatjuk egy referenciaváltozó intervallumpontszámait, keresve az ordinális változónak azt a két osztályát, amely esetén a B értéke a legkisebb, vagyis amikor az intervallumváltozó csoporton belüli eltérése a legkisebb.

A módszerek elég változatos lehetőséget kínálnak különböző mérési szintű változók dichotomizálásakor, ezért ezt a problémát csak vázlatosan érintettük. Arra azért mindenképpen felhívjuk a figyelmet, hogy az ilyen jellegű transzformációnál a kutató lehetősége és egyben felelőssége még fokozottabb.

3.2.7. A skálatranszformáció alkalmazása

A gyakorlati vizsgálatoknál a feltárandó kutatási területet legtöbbször különböző mérési szintű változók együttesével tudjuk jellemezni. A gyakorlat kevert változóhalmazai és a homogén változóhalmazt feltételező sokváltozós statisztikai eljárások közötti ellentmondás feloldására a *3.2. táblázatban* összefoglalt skálatranszformáló eljárások adnak lehetőséget.

Az átalakítások stratégiája az lehet, hogy a változók közül kiválasztjuk a kívánt vagy domináns típust, és az ettől eltérő típusú változókat átalakítjuk a megfelelő skálatranszformáló eljárással. Dominánsnak azt a változótípust tekinthetjük, amelyik vagy a legnagyobb számban fordul elő, vagy az elemzés szempontjait figyelembe véve a legfontosabbnak ítélhető. A transzformáló módszerek közötti választásnál a kutató meggondolásai is szerepet kapnak, amit nem ítélhetünk feltétlenül károsnak, mindenesetre rutinszerű felhasználásról itt nem nagyon lehet szó. Így, ha egy nominális változó kategóriáihoz akarunk intervallummérési szintű értékeket adni, a referenciaváltozó megválasztása meghatározza ezen lehetséges értékeket. Más referenciaváltozó figyelembevétele természetszerűen más kategóriaértékeket eredményez. Lehet, hogy egy nominális változó több referenciaváltozó szerinti értékeit célszerű megtartani, és ezzel a változók számát növelni. Ezeket a lehetőségeket mindig a konkrét feladat körültekintő megítélése szerint kell megvalósítani.

A skálatranszformáció a kevert változóhalmazok kezelésének egyik lehetséges módja. Említettük már azt az utat is, hogy a kevert változóhalmazt a mérési skála típusa szerint csoportokra bontjuk; így már homogén változóhalmazokra külön-külön végezzük el az elemzést.

A harmadik megoldás azt a lehetőséget használja fel, hogy a különböző típusú változók közötti kapcsolatok mérésére mutatók állnak rendelkezésünkre. Eszerint minden

Mérési skálák	Intervallum	Ordinális	Nominális
Intervallum		1. Rangszámok 2. Elméleti eloszlás feltételezése 3. Korreláció egy referenciaváltozóval	1. Korreláció egy intervallumvátozóval 2. Kanonikus korreláció 3. Kétlépéses transzformáció
Ordinális	1. Nézőpont megváltoztatása 2. Helyettesítés 3. Egyenlő hosszúságú kategóriák 4. Azonos tagszámú kategóriák 5. Egydimenziós hierarchikus módszerek 6. A felosztás iteratív javítása 7. Lineáris diszkriminánsfüggvény	1. Korreláció egy intervallumváltozóval 2. Rangkorreláció	
Nominális	1. A fenti módszerek, eltekintve a rendezéstől 2. Központi szélsőséges elv	1. Központi szélsőséges elv 2. Egyéb intuitív módszerek	

3.2. táblázat.

különböző típuskombinációra különböző számítási módon határozzuk meg a mértékeket. Ezek feldolgozását viszont már együttesen végezzük.

A javasolt mértékek különböző változók sztochasztikus kapcsolatának mérésére pl. a következők lehetnek:

– két intervallumváltozó esetén használjuk a Pearson-féle korrelációs együtthatót;
– egy nominális és egy intervallumváltozó esetén a korrelációt optimális kategóriaértékekre számítjuk, ahogyan a skálatranszformációnál láthattuk, de számíthatunk point-biseriális korrelációt is, ha a nominális változó két kategóriából áll;
– két nominális változó esetén a kanonikus korrelációt alkalmazhatjuk.

Az ordinális változót azért nem vettük külön figyelembe, mivel a rangszámokkal intervallumváltozóként kezelhetjük. A dichotom változót akár nominális, akár intervallumváltozóként kezelhetjük, ahogyan korábban ezt bemutattuk.

PÉLDA

A nominális skála felértékelésére a kanonikus korrelációt használhatjuk, egy referenciaváltozó segítségével. Mindkét változót bináris változókkal helyettesítjük, amelyek

egy-egy kategóriához tartozást vagy nem tartozást fejeznek ki. A kanonikus korrelációt az így kapott két bináris változóhalmaz között számítjuk, és a kategóriák súlyozását használjuk fel a kategóriák értékeként.

Példánkban nominális változónak a foglalkozást, referenciaváltozónak a legmagasabb iskolai végzettséget választottuk. Az adatok az MTA Szociológiai Intézet 1968-as kérdőíves felméréséből származnak. A minta a falusi népességet reprezentálta. Elemszáma 1974 fő. A számításokat elvégezve a következő eredményhez jutottunk.

A kanonikus korreláció: 0,72 (χ^2 = 2078,62, sz. fok = 72)

Foglalkozási kategóriák	Kanonikus együtthatók	Traszformált értékek
– értelmiségi, irányító beosztás	0,927	106
– irodai, adminisztratív dolgozó	0,302	43
– ipari szakmunkás	0,073	20
– ipari segédmunkás	−0,098	3
– mezőgazdasági, fizikai	0,128	0
– nyugdíjas	−0,114	1
– háztartásbeli	−0,068	6
– egyéb (önálló stb.)	−0,006	12

Eszerint az értelmiség 106-os értékével szemben a legkisebb értéket a mezőgazdasági fizikai foglalkozási kategória (0) adja.

Közöljük a referenciaváltozó súlyozását is:

Legmagasabb iskolai végzettség	Kanonikus együtthatók
– nem végzett semmilyen iskolát	−0,0187
– 3 osztálynál kevesebb	−0,023
– 4–5 osztály	−0,0088
– 6–7 osztály	0,011
– 8 általános, (illetve 4 polgári)	−0,206
– szakiskolák, bejezetlen középiskola	−0,199
– érettségi	−0,683
– befejezetlen főiskola, egyetem	−0,312
– főiskola, egyetem	−0,773

II. rész

Többváltozós módszerek

4. fejezet

Szóráselemzés

A szóráselemzés egy vagy több folytonos (legalább intervallummérési szintű) változó és egy kategorikus (nominális mérési szintű) változó közötti kapcsolat elemzésére ad módszert. A folytonos változó(k) megfigyelt értékeit a nominális változó lehetséges kategóriáiba soroljuk, és arra vagyunk kíváncsiak, hogy az így kialakított csoportok (amelyeket részmintáknak is tekinthetünk) között mutatkozik-e szignifikáns különbség. Más szóval függ(nek)-e az intervallumváltozó(k) a kategóriákat megadó nominális változótól, vagy tekinthetjük-e az egész mintát hasonlónak.

4.1. Egyváltozós szóráselemzés

Egy folytonos és egy kategorikus változó esetén az adatokat a következő táblázatba rendezzük:

	Minták (alminták vagy csoportok)				Együtt
	1.	2.	...	k.	
Adatok (egyes megfigyelések)	x_{11}	x_{12}	...	x_{1k}	
	x_{21}	x_{22}	...	x_{2k}	
	$\vdots$	$\vdots$		$\vdots$	
	$x_{n_1 1}$	$x_{n_2 2}$	...	$x_{n_k k}$	
Együtt	G_1	G_2	...	G_k	G
Átlag	$\overline{x}_1$	$\overline{x}_2$	...	$\overline{x}_k$	
Szórásnégyzet	S_1^2	S_2^2	...	S_k^2	
Szabadságfok	v_1	v_2	...	v_k	v

$v_j = n_j - 1 \quad (j = 1, \ldots, k), \quad v = n - k.$

Először azt a hipotézist vizsgáljuk, hogy a minták szórásnégyzetei egyenlőek a populáció varianciájával, így egyenlőek egymással is:

$$H_0: E(S_1^2) = \ldots = E(S_k^2) = V.$$

A próbafüggvényt Bartlett dolgozta ki, feltételezve, hogy a változók normális eloszlásúak:

$$\chi^2 = BC^{-1},$$

ahol $B = \sum_{j=1}^{k} v_j \ln S^2 - \sum_{j=1}^{k} v_j \ln S_j^2,$

$$C = 1 + \{1/[3(k-1)]\} \left\{ \sum_{j=1}^{k} 1 \Big/ v_j - 1 \Big/ \sum_{j=1}^{k} v_j \right\},$$

$$S^2 = \left(\sum_{j=1}^{k} v_j S_j^2 \right) \Big/ \left(\sum_{j=1}^{k} v_j \right).$$

A χ^2 szabadságfoka $k - 1$.
Az alternatív hipotézis egyoldali.

4.1. PÉLDA

Induljunk ki a következő táblázatból:

Minták		
1	2	3
45	23	32
40	24	48
16	31	63
28	25	42
39	48	39
38	22	57
25	35	39
31	21	54
–	18	49
–	16	–

4.1. táblázat.

A becslések:

$$\bar{x}_1 = 32{,}75, \qquad \bar{x}_2 = 26{,}3, \qquad \bar{x}_3 = 47{,}0,$$
$$S_1^2 = 90{,}79, \qquad S_2^2 = 89{,}79, \qquad S_3^2 = 98{,}5,$$
$$v_1 = 7, \qquad v_2 = 9, \qquad v_3 = 8.$$

A próbához szükséges mennyiségek:

$$S^2 = [7(90{,}79) + 9(89{,}79) + 8(98{,}5)]/(7 + 9 + 8) = 2231{,}64/24 = 92{,}985,$$

$$B = 24 \ln 92{,}985 - [7 \ln 90{,}79 + 9 \ln 89{,}79 + 8 \ln 98{,}5] = 0{,}021,$$

$$C = 1 + \frac{1}{6}\left[\frac{1}{7} + \frac{1}{9} + \frac{1}{8} - \frac{1}{24}\right] = 1 + \frac{1}{6}[0{,}379 - 0{,}042] = 1{,}056.$$

A próba:

$$\chi^2 = BC^{-1} = 0{,}021/1{,}056 = 0{,}0198.$$

Mivel

$$(\chi^2 = 0{,}0198) < (\chi^2_{0{,}05,\ 2} = 5{,}991),$$

így a nullhipotézist elfogadjuk.

Ha elfogadtuk a varianciák várható értékei egyenlőségére vonatkozó hipotézist, megvizsgálhatjuk az átlagok várható értékei egyenlőségére vonatkozó hipotézist.
A nullhipotézis:

$$H_0: E(\overline{x}_1) = \ldots = E(\overline{x}_k) = \mu.$$

Feltételezzük, hogy bármelyik megfigyelt adathoz lineáris additív modellt illeszthetünk:

$$x_{ij} = \mu + \tau_j + e_{ij} = \mu_j + e_{ij},$$

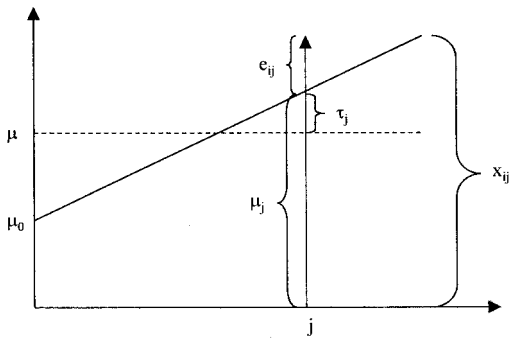
ahol x_{ij} a megfigyelt egyedek a j-edik mintában,

μ az átlagos egyed az összes mintában,

τ_j a j-edik minta átlagának és a teljes átlagnak a különbsége,

e_{ij} normális eloszlású hibatag, várható értéke 0, varianciája V_c.

Minden mintában feltételezzük, hogy az x változó azonos szórású normális eloszlást követ. A modellt a 4.1. ábrával szemléltetjük.



4.1. ábra. Lineáris additív modell

A mintából becsült mennyiségek:

$$\overline{x} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij} \Big/ \sum_{j=1}^{k} n_j = G/n\ ,$$

$$\overline{x}_j = n_j^{-1} \sum_{i=1}^{n_j} x_{ij} = G_j/n_j.$$

A τ_j becslése:

$$T_j = \overline{x}_j - \overline{x}.$$

Egyéb összefüggések:

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij}-\overline{x})=0,$$

$$\sum_{i=1}^{n_j}(x_{ij}-\overline{x}_j)=0 \quad (\forall\ j-re),$$

$$\sum_{j=1}^{k} n_j T_j = 0.$$

A teljes eltérésnégyzetek összegét felbonthatjuk két részre:

$$Q=\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij}-\overline{x})^2,$$

$$Q = Q_T + Q_E,$$

ahol

$$Q_T=\sum_{j=1}^{k} n_j(\overline{x}_j-\overline{x})^2,$$

$$Q_E=\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij}-\overline{x}_j)^2.$$

A Q_T a hipotetikus hatást, a Q_E pedig a hibahatást fejezi ki.
Az ezekkel összefüggő varianciák:

$$S^2 = Q/(n-1),$$

$$S_T^2 = Q_T/(k-1),$$

$$S_E^2 = Q_E/(n-k).$$

A varianciák várható értékei:

$$E(S^2) = V + (k-1)^{-1}\sum_{j=1}^{k} n_j(\mu_j-\mu)^2,$$

$$E(S_E^2) = V.$$

Ha az átlagok egyenlőségére vonatkozó H_0 hipotézis igaz, akkor a varianciák várható értékei egyenlők:

$$E(S_T^2) = E(S_E^2) = V_T = V.$$

A nullhipotézis:

$$H_0\colon E(S_T^2) = E(S_E^2) = V$$

és az alternatív hipotézis:

$$H_1\colon E(S_T^2) > E(S_E^2) = V.$$

A H_0 hipotézist elvetjük, ha a minták közötti eltérések szignifikánsan nagyobbak, mint a mintákon belüli eltérések.
A próba függvénye:

$$F = S_T^2/S_E^2$$

F-eloszlású változó $v_T = k - 1$ és $v_E = n - k$ szabadságfokkal (feltételezve azt, hogy a populáció normális eloszlású).

A H_0 hipotézist elutasítjuk akkor, ha a számított érték nagyobb vagy egyenlő, mint a választott szignifikanciaszint és adott szabadságfok melletti elméleti érték

$$F \geq F_{\alpha, v_T, v_E}.$$

A szóráselemzés eredményeit ún. szórásfelbontó táblázatba szokás rendezni. Ezt mutatja a 4.2. táblázat.

A szórás	Négyzetösszeg	Szabadságfok	A variancia becslései	F
A minták között	Q_T	$k-1$	S_T^2	S_T^2/S_E^2
A mintán belül	Q_E	$n-k$	S_E^2	
Teljes	Q	$n-1$	S^2	

4.2. táblázat.

4.2. PÉLDA

A 4.1. táblázat adataiból a következő számításokat végezzük:

$$Q = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \overline{x})^2 = \sum_j\sum_i x_{ij}^2 - \left(\sum_j\sum_i x_{ij}\right)^2 /n = 4324{,}7,$$

$$Q_E = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \overline{x}_j)^2 = \sum_{j=1}^{k}\left[\sum_{i=1}^{n_j} x_{ij}^2 - \left(\sum_{i=1}^{n_j} x_{ij}\right)^2 /n_j\right] = 2231{,}6,$$

$$Q_T = Q - Q_E = 4324{,}7 - 2231{,}6 = 2093{,}1,$$

$$S_T^2 = Q_T/v_T = 2093{,}1/2 = 1046{,}5,$$

$$S_E^2 = Q_E/v_E = 2231{,}6/24 = 92{,}98.$$

Az eredményeket a 4.3. táblázat összesíti.

A szórás eredete	Négyzetösszeg	Szabadságfok	A szórásnégyzet becslése	F
Szisztematikus	2093,1	2	1046,5	11,26
Hiba	2231,6	24	92,98	
Teljes	4324,7	26		

4.3. táblázat.

Az elméleti eloszlás kritikus értéke:

$$F_{0{,}05,2,24} = 3{,}40.$$

Mivel

$$(F = 11{,}26) > (F_{\alpha, v_T, v_E} = 3{,}40),$$

a H_0 hipotézist elvetjük. A három minta átlaga nem tekinthető a populációbeli μ várható érték közös becslésének.

4.2. Többváltozós szóráselemzés

Feltételezzük, hogy a k számú mintát egy normális eloszlású sokaságból véletlenszerűen vettük. Mindegyik mintában p számú változót figyelünk meg, a mintát a p számú változó kovarianciamátrixával jellemezzük. Azt vizsgáljuk, hogy a minták kovarianciamátrixai ($\mathbf{S}_j$) egyenlőknek tekinthetők-e.

A nullhipotézis:

$$H_0: E(\mathbf{S}_1^2) = \ldots = E(\mathbf{S}_k^2).$$

A próbafüggvény:

$$\chi^2 = BC^{-1}$$

χ^2-eloszlású $v = (k-1)p(p+1)/2$ szabadságfokkal, ahol B a kovarianciamátrixok determinánsainak a függvénye

$$B = \sum_{j=1}^{k}(n_j - 1)\ln(|\mathbf{S}|/|\mathbf{S}_j|)$$

(ahol $\mathbf{S} = \sum_{j=1}^{k}(n_j - 1)\mathbf{S}_j \Big/ \sum_{j=1}^{k}(n_j - 1)$), C^{-1} a minták elemszámának a függvénye

$$C^{-1} = 1 - (2p^2 + 3p - 1)\left[\sum_{j=1}^{k}(n_j - 1)^{-1} - 1 \Big/ \sum_{j=1}^{k}(n_j - 1)\right] \Big/ [(p+1)6(k-1)].$$

Az alternatív hipotézis egyoldali.

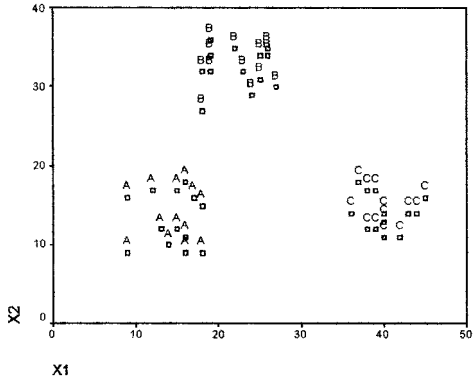
4.3. Példa

Tekintsünk három mintában két változót (4.4. táblázat).

Megfigyelések	Minták					
	A		*B*		*C*	
	x_1	x_2	x_1	x_2	x_1	x_2
1	16	11	18	27	43	14
2	15	12	19	34	38	17
3	18	9	23	32	37	18
4	12	17	19	36	39	12
5	18	15	25	31	44	14
6	9	9	24	29	40	13
7	15	17	22	35	42	11
8	9	16	27	30	40	13
9	16	18	25	34	36	14
10	18	9	19	32	39	12
11	17	16	18	32	39	17
12	13	12	26	35	45	16
13	16	9	25	31	40	14
14	15	17	26	34	40	11
15	14	10	26	34	38	12

4.4. táblázat.

Ezeket az adatokat az 4.2. ábra szemlélteti.



4.2. ábra. Három alminta két dimenzióban

A kovarianciamátrixok:

$$\mathbf{S}_A = \begin{bmatrix} 8{,}49524 & -0{,}961905 \\ -0{,}961905 & 12{,}40950 \end{bmatrix},$$

$$\mathbf{S}_B = \begin{bmatrix} 11{,}0285 & 0{,}657143 \\ 0{,}657143 & 6{,}25714 \end{bmatrix},$$

$$\mathbf{S}_C = \begin{bmatrix} 6{,}42857 & -0{,}642857 \\ -0{,}642857 & 4{,}98095 \end{bmatrix},$$

$$\mathbf{S} = [3(15-1)]^{-1}[14 \cdot \mathbf{S}_A + 14 \cdot \mathbf{S}_B + 14 \cdot \mathbf{S}_C] = \begin{bmatrix} 8{,}65079 & -0{,}315873 \\ -0{,}315873 & 7{,}88254 \end{bmatrix},$$

$$|\mathbf{S}_A| = 104{,}497,$$

$$|\mathbf{S}_B| = 68{,}5755,$$

$$|\mathbf{S}_C| = 31{,}6071,$$

$$|\mathbf{S}| = 68{,}0904,$$

$$B = 14\ln(68{,}0904/104{,}497) + 14\ln(68{,}0904/68{,}5755) + 14\ln(68{,}0904/31{,}6071) = 4{,}64853,$$

$$C^{-1} = 1 - (8+6-1)\cdot(0{,}2143 - 1/42)/[(3)6(2)] = 0{,}931217,$$

$$\chi^2 = BC^{-1} = 4{,}329.$$

Az $\alpha = 0{,}05$ és $v = 0{,}5(2)(2)(3) = 6$ mellett a kritikus érték:

$$\chi^2_{0{,}05,\ 6} = 12{,}592.$$

Mivel

$$(\chi^2 = 4{,}329) < (\chi^2_{0{,}05,\ 6} = 12{,}592),$$

a nullhipotézist elfogadjuk.

Ha a minták összehasonlításakor elsősorban a relatív helyük és nem alakjuk, kiterjedésük érdekel bennünket, a kovarianciák helyett az átlagokat vetjük össze. Ekkor a

nullhipotézisünk:

$$H_0\colon E\begin{bmatrix}\overline{x}_{11}\\ \overline{x}_{12}\\ \vdots\\ \overline{x}_{1p}\end{bmatrix} = \ldots = E\begin{bmatrix}\overline{x}_{k1}\\ \overline{x}_{k2}\\ \vdots\\ \overline{x}_{kp}\end{bmatrix}.$$

A próbafüggvény:

$$\Theta = \lambda/(1+\lambda),$$

ahol λ a $\mathbf{HE}^{-1}$ mátrix legnagyobb sajátértéke.

A $\mathbf{H}$ és $\mathbf{E}$ mátrix a keresztszorzat- és négyzetösszeg-mátrix két összetevője:

$$\mathbf{T} = \mathbf{H} + \mathbf{E},$$

ahol

$$t_{ij} = \sum_{g=1}^{k}\sum_{f=1}^{n_g} x_{fgi}x_{fgj} - x_{00i}x_{00j}/n \quad (i, j = 1, \ldots, p),$$

$$h_{ij} = \sum_{g=1}^{k} x_{0gi}x_{0gj}/n_g - x_{00i}x_{00j}/n,$$

$$e_{ij} = \sum_{g=1}^{k}\left[\sum_{f=1}^{n_g} x_{fgi}x_{fgj} - x_{0gi}x_{0gj}/n_g\right],$$

a h_{ij} a minták közötti eltéréseket,

az e_{ij} a mintán belüli eltéréseket fejezi ki,

x_{fgi} az i-edik változó értéke a g-edik csoport f-edik megfigyelési egységénél,

x_{0gi} az i-edik változó összesített értéke a g-edik csoportban,

x_{00i} az i-edik változó teljes összege,

n_g a g-edik csoport elemszáma,

$$n = \sum_{g=1}^{k} n_g.$$

Feltételezve, hogy a populáció normális eloszlású, a Θ eloszlása ismert, ha a nullhipotézis igaz. Az alternatív hipotézis egyoldali.

A H_0 hipotézist elfogadjuk, ha

$$\Theta < \Theta_{\alpha,s,m,r}.$$

A kritikus $\Theta_{\alpha,s,m,r}$ értéket a Heck-féle táblázatból keressük meg, ahol

$s = \min(k-1, p)$,

$m = (|k-p-1| - 1)/2$,

$r = (n-k-p-1)/2$.

Ha a H_0 hipotézist elvetjük, érdekes lehet tudni, hogy a várhatóérték-vektorok közül melyek azok, amelyek nem becslései a populáció várható értékének.

A nullhipotézis:

$$H_0\colon E\left[\sum_{i=1}^{p}\sum_{g=1}^{k} a_i c_g \overline{x}_{gi}\right] = 0,$$

ahol $\mathbf{a} = [a_1, \ldots, a_p]'$ egy „design" vektort jelöl 0, 1 elemekkel, aszerint, hogy melyik változót vizsgáljuk (pl. $\mathbf{a} = [1, 0, 1]$ azt jelöli, hogy az első és harmadik változó átlagait vetjük össze);

$\mathbf{c} = [c_1, \ldots, c_k]$ a „contrast coefficient", amely azt jelöli ki, hogy mely átlagvektorokat hasonlítjuk össze.

A c_j-k összege nullával egyenlő:

$$\sum_{j=1}^{k} c_j = 0.$$

A nullhipotézist elfogadjuk, ha az $1 - \alpha$ valószínűségű intervallum tartalmazza a nullát. A konfidenciasáv:

$$\sum_{i=1}^{p} \sum_{g=1}^{k} a_i c_g \overline{x}_{gi} \pm \left[\Theta_{\alpha,s,m,r} (1 - \Theta_{\alpha,s,m,r})^{-1} \mathbf{a}' \mathbf{E} \mathbf{a} \sum_{g=1}^{k} c_g^2 / n_g \right]^{1/2}.$$

4.4. PÉLDA

A 4.3. példát vizsgáljuk. A három átlagvektor

$$\overline{\mathbf{x}}_A = \begin{bmatrix} 14{,}7 \\ 13{,}1 \end{bmatrix}, \quad \overline{\mathbf{x}}_B = \begin{bmatrix} 22{,}8 \\ 32{,}4 \end{bmatrix}, \quad \overline{\mathbf{x}}_C = \begin{bmatrix} 40{,}0 \\ 13{,}9 \end{bmatrix}.$$

A próbafüggvényhez szükséges mátrixok:

$$\mathbf{H} = \begin{bmatrix} 4996{,}57 & -724{,}133 \\ -724{,}133 & 3576{,}13 \end{bmatrix},$$

ahol $h_{12} = [221(197)/15 + 342(486)/15 + 600(208)/15] - [1163(891)/45] = -724{,}133$.

$$\mathbf{E} = \begin{bmatrix} 363{,}333 & -13{,}2667 \\ -13{,}2667 & 331{,}067 \end{bmatrix},$$

ahol $e_{12} = [2889 - 221(197)/15 + 11090 - 342(486)/15 + 8311 - 600(208)/15] = -13{,}26$.

Az $\mathbf{E}$ mátrix inverze:

$$\mathbf{E}^{-1} = \begin{bmatrix} 0{,}00275633 & 0{,}00011045 \\ 0{,}00011045 & 0{,}00302497 \end{bmatrix}.$$

A $\mathbf{H}\mathbf{E}^{-1}$ szorzatmátrix inverze:

$$\mathbf{H}\mathbf{E}^{-1} = \begin{bmatrix} 13{,}6922 & -1{,}6386 \\ -1{,}6386 & 10{,}7377 \end{bmatrix}.$$

A $\mathbf{H}\mathbf{E}^{-1}$ szorzatmátrix sajátértékét úgy számítjuk, hogy megoldjuk a következő egyenletet:

$$|\mathbf{H}\mathbf{E}^{-1} - \lambda \mathbf{I}| = 0,$$

ahol $\mathbf{I}$ az egységmátrixot jelöli.

A determinánst kifejtve a következő egyenlethez jutunk:

$$\lambda^2 - 24{,}4299\lambda + 144{,}3953 = 0.$$

A két gyök:

$$\lambda_1 = 14{,}4071,$$

$$\lambda_2 = 10{,}0228.$$

A próbafüggvény:

$$\Theta = \lambda_1/(1 + \lambda_1) = 14{,}4071/(1 + 14{,}4071) = 0{,}935.$$

Az elméleti érték paraméterei:

$$s = \min(2, 2) = 2,$$

$$m = (3 - 2 - 1 - 1)/2 = -0{,}5,$$

$$r = (45 - 3 - 2 - 1)/2 = 19{,}5.$$

Mivel

$$(\Theta = 0{,}935), \quad Z(\Theta_{0{,}05,s,m,r} = 0{,}185),$$

a H_0 hipotézist elvetjük, vagyis az átlagvektorok szignifikánsan különböznek.

Érdekes lehet tudni, hogy mely változók mely mintabeli átlagai járultak szignifikánsan hozzá a nullhipotézis elvetéséhez. Vizsgáljuk meg azt, hogy az A és B minta első és második változójának mi volt a szerepe a nullhipotézis elutasításában.
Ekkor

$$\mathbf{c} = [1, -1{,}0]',$$

$$\mathbf{a} = [1{,}1]'.$$

A konfidenciaintervallum

$$\mathbf{a}'\mathbf{E}\mathbf{a} = 667{,}867,$$

$$\sum_{g=1}^{3} c_g^2/n_g = 0{,}133333,$$

$$\sum_{i=1}^{2}\sum_{g=1}^{3} a_i c_g \overline{x}_{gi} = -27{,}3333.$$

A konfidenciaintervallum felső határa:

$$L_1 = -22{,}8374,$$

alsó határa:

$$L_2 = -31{,}8293.$$

Mivel a két határ nem tartalmazza a nullát, a H_0 hipotézist a fenti összehasonlításban is elvetjük.

Ha

$$\mathbf{c} = [0, 1, -1],$$

és mindkét változót vizsgáljuk, az

$$L_1 = 5{,}82928,$$

$$L_2 = -3{,}16261$$

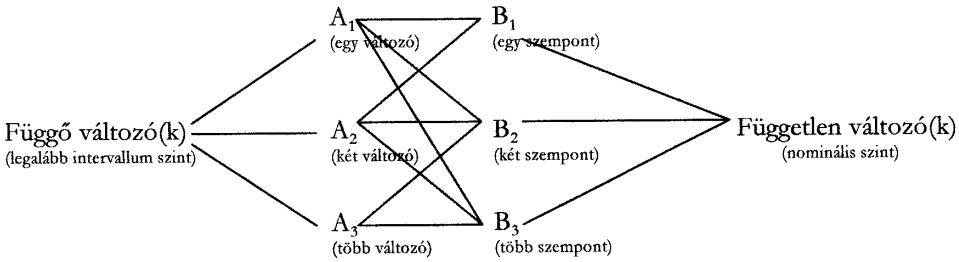
határok közé esik a nulla is, ezért azt állapíthatjuk meg, hogy a B és C minta nem különbözik szignifikánsan.

4.3. A szóráselemzésről általában

A látottak szerint a szóráselemzés hipotézisvizsgálati módszer, érdemes külön is felhívni a figyelmet sajátosságaira:

1. lényegesen eltérő mérési szintű (egyrészt intervallum- vagy arányskálán, másrészt nominális szinten mért változók sztochasztikus kapcsolatának egzisztenciájára vonatkozik a hipotézis).
2. Az elemi egyváltozós esetből kiindulva, a mindkét oldalon többváltozós rendszerek vizsgálatáig, egyre szerteágazóbban, további érdekes kérdésfeltevésekkel együtt használható.
3. A gyakorlati alkalmazások szempontjából az eljárás annál érdekesebb, minél több változó szerepel az „intervallum" oldalon, és annál számolásigényesebb, mondhatjuk, hogy bonyolultabb, minél több szempont szerepel a „nominális" oldalon.

Használjuk fel a továbbiakban az alábbi vázlatos sémát:



4.3. ábra

1. A_1B_1: (egyváltozós), egyszempontos szóráselemzés (4.1. fejezet)
2. $\left.\begin{matrix}A_1B_2\\A_1B_3\end{matrix}\right\}$: két- vagy többszempontos szóráselemzés (4.4. fejezet) (egy változóval)
3. $\left.\begin{matrix}A_2B_1\\A_3B_1\end{matrix}\right\}$: (egyszempontos) többváltozós $\longrightarrow$ szorosan kapcsolódik hozzá a később szóráselemzés (4.2. fejezet) tárgyalandó:

   $\left\{\begin{matrix}\text{kovarianciaelemzés (7.4. fejezet)}\\\text{diszkriminanciaelemzés (7. fejezet)}\end{matrix}\right.$
4. $\left.\begin{matrix}A_2B_2\\A_3B_2\end{matrix}\right\}$: faktoriális diszkriminanciaelemzés (több változó, több szempont) (7.5. fejezet)
5. A_3B_3:

A fenti sémából és a hozzá kapcsolódó szituációk szerteágazó tárgyalásából jól látható, hogy bár az alaphipotézis végig ugyanaz, az egész problémakör mégis eléggé összetett. Az 4.1. és 4.2. fejezetek tömören és egyszerű numerikus példákkal tárgyalják a leggyakrabban alkalmazott eseteket. Az egy változó – egy szempont, a klasszikus alapeset, a több változó – egy szempont szituáció pedig nem túl bonyolult gondolatmenet mellett, új, érdekes kérdésfeltevésekre ad alkalmat. A szempontok számának növelésével előálló összetettebb helyzeteket s azok speciális eseteit az 4.4. fejezetben külön tárgyaljuk.

Bár könyvünk e fejezetében számos példát talál az olvasó a szóráselemzés alkalmazására, alábbiakban mégis megadunk néhány további hazai irodalmi hivatkozást, ahol különböző területekről származó, további, nagyobb lélegzetű példák találhatók. Úgy gondoljuk, az alkalmazók ma még nem használják ki a szóráselemzés eljárásában rejlő lehetőségeket.

1. Vincze István: Statisztikai minőségellenőrzés. Bp.: KJK, 1958. 107. old. Ipari alkalmazás.
2. Vincze István: Matematikai statisztika ipari alkalmazásokkal. Bp.: Műszaki Kiadó, 1968. 187. old. Ipari alkalmazás. Ebben a könyvben benne van a 3 szempontú osztályozás felbontó táblája!
3. Füstös–Meszéna–Simonné: A sokváltozós adatelemzés statisztikai módszerei. Bp.: Akadémiai Kiadó, 1986. 136. old. Példa a többváltozós szóráselemzésre szociológiai témában. (A családfők intergenerációs mobilitása.)
4. Lásd a 3. pontbeli könyvet: 441. old. a teljes 3. esettanulmány. Közgazdasági probléma vizsgálata a nagyberuházások jellemzői közötti kapcsolatok elemzésével.

4.4. A két- és többszempontos (egyváltozós) szóráselemzés

Az 4.3. ábrában és a hozzá fűzött magyarázatban ezt a fejezetet is elhelyeztük a szóráselemzés egész rendszerében. Megindokoltuk a fejezet külön történő tárgyalását is. Gondolkodhatnánk úgy is, hogy csak az intervallumváltozók számának növekedését tartjuk „sokváltozós statisztikai" problémának, s a szempontok (szintek) oldalán csak egy esetet veszünk figyelembe. Ekkor azonban ez a könyv sem foglalkozna a szóráselemzéssel a maga teljességében, s ezt szeretnénk elkerülni.

A szóráselemzés alapvető gondolatmenete összhangban az előzőekkel – de azért kicsit átfogalmazva – a következő: a hatótényezők különböző állapotait – amelyeket ezentúl szinteknek nevezünk – figyelembe véve, mintákat veszünk a szóban forgó valószínűségi változóra. Ha több tényezőt vizsgálunk, akkor a tényezők szintjeinek kombinációjára veszünk mintákat. A mintavétel történhet úgy is, hogy minden lehetséges szintkombincióra egy-egy elemű, vagy mindegyikre azonos elemszámú (több elemű) mintát veszünk. (Vannak olyan szóráselemzési modellek is, ahol nincs kikötve az azonos elemszám, sőt pl. a „latin négyzet" módszernél nem is veszünk minden lehetséges szintkombinációra mintát.)

Ha pl. egy valószínűségi változó kialakításában két – az A és B – tényező játszik szerepet, és az A tényezőnek p számú szintje (állapota) van, a B tényezőnek pedig q számú, akkor a két tényező szintjeinek kombinációját az $A_i \cap B_j$ szorzat jelöli ($i = 1, \ldots, p$ és $j = 1, \ldots, q$), ahol az első tényező i-edik, és a második tényező j-edik szintjének fennállása mellett vesszük az előírt számú mintát.

A gyakorlat oldaláról fogalmazzuk meg az alábbi kérdésfeltevéseket.

1. Mezőgazdasági kísérleteknél vizsgáljuk a különböző fajtájú termények terméseredményét. Nyilvánvalóan a terméseredményt a vetőmag fajtáján kívül számos más tényező is befolyásolja, amelyek hatását (pl. a kísérleti földdarab talajának különbözőségét) azonban szeretnénk kiszűrni, mivel csak arra a hatásra vagyunk kíváncsiak, amit az eltérő vetőmag okozott.

2. Valamilyen közgazdasági jellemzőre, pl. a „megélhetési indexre” ismerjük néhány szakértő becslését az adott területi egységekre. Ekkor elsősorban két kérdés merülhet fel (más-más szempontból érdekesek):
   a) Van-e lényeges különbség a különböző szakértők között a becslés eredményét illetően? (pl. van-e olyan szakértő, aki általában kisebbre becsüli minden terület esetében az illető jelzőszámot stb.)
   b) Van-e lényeges különbség a különböző területi egységek között a megélhetési indexet illetően?

Az a) esetben nyilván a területi egységek különbözősége okozta hatást szeretnénk kiszűrni, a b)-ben pedig éppen ellenkezőleg a területi egységek különbözőségét vizsgáljuk, így a szakértők becslései közötti különbséget szeretnénk kiküszöbölni.

Ezek után az ismeretlen, közös szórásra több, egymástól független becslést készítünk. Ha a becslések lényegesen különböznek egymástól, arra következtetünk, hogy a véletlenen kívül más tényező is befolyásolja becsléseinket. Mivel adataink pl. a feltétel szerint csak egyetlen *ismert* tényező hatását tartalmazhatták, következtetésünk: az illető tényező befolyása az eredményre szignifikáns.

4.4.1. A szóráselemzés alaptétele: A Fisher–Cochran-tétel

Mindenekelőtt tisztázandó a szabadságfok fogalma.

Legyenek $X_1, \ldots, X_n$ független $N(0, \sigma)$ eloszlású valószínűségi változók. Legyenek továbbá az

$$Y_1, \ldots, Y_m \quad \text{az } X_i\text{-ik lineáris függvényei:}$$
$$Y_i = a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{in}X_n; \quad (i = 1, \ldots, m).$$

Tegyük fel, hogy az Y_i-k között k darab lineáris összefüggés írható fel:

$$b_{11}Y_1 + b_{12}Y_2 + \ldots + b_{1m}Y_m = 0$$
$$\vdots \qquad\qquad\qquad \vdots$$
$$b_{k1}Y_1 + b_{k2}Y_2 + \ldots + b_{km}Y_m = 0.$$

Legyen a

$$\mathbf{B} = [b_{ij}] = \begin{bmatrix} b_{11} & \ldots & b_{1m} \\ \vdots & & \vdots \\ b_{k1} & \ldots & b_{km} \end{bmatrix}$$

mátrix rangja r, ekkor az $Y_1, \ldots, Y_m$ rendszer szabadságfokán az $(m - r)$ számot értjük.

Az Y_i lineáris kifejezésekből képzett

$$Q = Y_1^2 + Y_2^2 + Y_m^2$$

négyzetösszeg szabadságfokán az $(Y_1, \ldots, Y_m)$ rendszer szabadságfokát értjük.

A χ^2-eloszlás bevezetésénél szó volt arról, hogy az $X_1, \ldots X_n$ független, $N(0, 1)$ eloszlású valószínűségi változók négyzetösszege

$$Q = X_1^2 + X_2^2 + \ldots + X_n^2$$

χ^2-eloszlást követ n-szabadságfokkal. Ekkor nyilván az is igaz, hogy egy n-szabadságfokú χ^2-eloszlású valószínűségi változó előállítható n darab független, $N(0, 1)$ eloszlású valószínűségi változó négyzetének összegeként.

Ezek után nézzük a *Fisher–Cochran-tételt* – amely két állítást tartalmaz: a χ^2-*addíciós tételt*, és a χ^2-*particiós tételt*.

χ^2-ADDÍCIÓS TÉTEL

Ha $Q_1, \ldots, Q_k$ független, rendre $f_1, \ldots, f_k$ szabadságfokú, χ^2-eloszlású valószínűségi változók, akkor a

$$Q = Q_1 + Q_2 + \ldots + Q_k$$

is χ^2-eloszlású lesz $\sum f_i = f$ szabadságfokkal.

χ^2-PARTÍCIÓS TÉTEL

Legyen a független $N(0, 1)$ eloszlású X_i valószínűségi változókból álló $\sum_{i=1}^{n} X_i^2$ négyzetösszeg k darab tagra felbontva:

$$\sum_{i=1}^{n} X_i^2 = Q_1 + Q_2 + \ldots + Q_k,$$

ahol Q_j-k az X_i-k lineáris kifejezéseinek négyzetösszegei f_j szabadságfokkal ($j = 1, \ldots k$).

Ekkor annak szükséges és elégséges feltétele, hogy a Q_j-k függetlenek, és f_j szabadságfokú χ^2-eloszlásúak legyenek, az, hogy fennálljon:

$$\sum_{j=1}^{k} f_j = n.$$

4.4.2. Egyszeres osztályozás

Legyenek az egyes osztályokba tartozó valószínűségi változók $X_1, \ldots X_p$, rendre $N(m_1, \sigma), \ldots, N(m_p, \sigma)$ eloszlásúak. Ekkor nullhipotézisünk:

$$H_0: m_1 = m_2 = \ldots = m_p \quad \text{és} \quad \sigma \quad \text{ismeretlen.}$$

A szórásokról az azonosságot tételezzük fel, értéküket nem ismerjük.

A hipotézisünk megvizsgálásához vegyünk az $X_1, \ldots, X_p$ változókra r elemű mintákat.

Az i-edik változóra vett k-adik mintaelemet X_{ik}-val jelöljük.

Alkalmazni fogjuk még a következő jelölési módot is: ha egy többindexes változó valamilyen index szerinti összegét akarjuk jelölni, akkor az illető index helyére pontot teszünk, ha pedig valamelyik index szerinti átlagát, akkor az index helyére pontot teszünk és felülvonással jelöljük az átlagolást. Pl.:

$$X_{ijk} \quad \text{esetén} \quad X_{i\cdot k}$$

a j szerinti összeget jelöli (vagyis ha $j = 1, \ldots, q$), akkor

$$X_{i\cdot k} = \sum_{j=1}^{q} X_{ijk},$$

$\overline{X}_{i\cdot k}$ pedig a j-szerinti átlagot jelenti.

$$\left(\text{vagyis} \quad \overline{X}_{i\cdot k} = \frac{\sum_{j=1}^{q} X_{ijk}}{q} \right).$$

A modell alapvető tartalma a következő:

$$X_{ik} = \mu + \tau_i + \varepsilon_{ik}.$$

A feltételek:

1. Az X_i változók normális eloszlásúak,

$$\underset{k}{E}(X_{ik}) = \mu + \tau_i \quad \text{várható értékkel}$$

és

$$\underset{k}{D}(X_{ik}) = \sigma^2 \quad (\text{minden } i\text{-re azonos}) \text{ szórással.}$$

2. Minden mintaelem független a többitől.
3. A hatások összegére érvényes, hogy

$$\sum_i \tau_i = 0.$$

A μ és τ_i paraméterek becslését a

$$Q = \sum_i \sum_k (X_{ik} - \widehat{\mu} - \widehat{\tau}_i)^2$$

négyzetösszeg minimalizálásával kapjuk meg, ahol $\widehat{\mu}$ a μ becslése és $\widehat{\tau}_i$ a τ_i becslése. Ekkor

$$\widehat{\mu} = \frac{\sum_i \sum_k X_{ik}}{rp} = \overline{X}_{\cdot\cdot}$$

$$\widehat{\tau}_i = \frac{\sum_k X_{ik}}{r} - \widehat{\mu} = \overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot}$$

$$\varepsilon_{ik} = X_{ik} - \underset{i,k}{E}(X_{ik}) = X_{ik} - \mu - \tau_i$$

ε_{ik} becslése $\widehat{\varepsilon}_{ik}$, ahol $\widehat{\varepsilon}_{ik} = X_{ik} - \widehat{X}_{ik} = X_{ik} - \overline{X}_i$.

$$X_{ik} = \widehat{\mu} + \widehat{\tau}_i + \widehat{\varepsilon}_{ik}.$$

Mindkét oldalát négyzetre emelve és összegezve,

$$\sum_i \sum_k X_{ik}^2 = \sum_i \sum_k \widehat{\mu}^2 + \sum_i \sum_k \widehat{\tau}_i^2 + \sum_i \sum_k \widehat{\varepsilon}_{ik}^2 +$$
$$+ 2\widehat{\mu} \sum_i \sum_k \widehat{\tau}_i + 2\widehat{\mu} \sum_i \sum_k \widehat{\varepsilon}_{ik} + 2 \sum_i \sum_k \widehat{\tau}_i \widehat{\varepsilon}_{ik}.$$

Mivel

$$\sum_i \sum_k \widehat{\tau}_i = 0, \quad \text{és} \quad \sum_i \sum_k \widehat{\varepsilon}_{ik} = 0$$

$$\sum_i \sum_k X_{ik}^2 = \sum_i \sum_k \widehat{\mu}^2 + \sum_i \sum_k \widehat{\tau}_i^2 + \sum_i \sum_k \widehat{\varepsilon}_{ik}^2.$$

A jobb oldal első tagját átvíve a másik oldalra, valamint elvégezve a helyettesítéseket:

$$\sum_i \sum_k (X_{ik} - \overline{X}_{..})^2 = r \sum_i (\overline{X}_{i.} - \overline{X}_{..})^2 + \sum_i \sum_k (X_{ik} - \overline{X}_{i.})^2.$$

Így a teljes eltérésnégyzetek összegét felbontottuk két olyan tagra, ahol az első a csoportok közötti eltérésnégyzet, a másik pedig a csoportokon belüli eltérésnégyzet.

Így a Q felbontható

$$Q = Q_1 + Q_{\text{rez}}$$

alakba, ahol

$$Q_{\text{rez}} = \sum_{i=1}^{p} \sum_{k=1}^{r} (X_{ik} - \overline{X}_{i.})^2$$

$$Q_1 = r \sum_{i=1}^{r} (\overline{X}_{i.} - \overline{X}_{..})^2.$$

A Q_{rez} elemeinek vizsgálata során, mivel a zárójelben egy valószínűségi változó átlagától való eltérései szerepelnek:

$$E(X_{ik} - \overline{X}_{i.}) = E(X_{ik}) - E(\overline{X}_{i.}) = m_i - m_i = 0.$$

ami mindig teljesül, akár igaz a H_0 nullhipotézis, akár nem. A Q_{rez} a csoporton belüli ingadozást mutatja, s ezt az ingadozást a véletlen okozza.

A Q_1 a zárójelében lévő kifejezés várható értéke

$$E(\overline{X}_{i.} - \overline{X}_{..}) = E(\overline{X}_{i.}) - E(\overline{X}_{..}) = m_i - \frac{\sum_{s=1}^{p} m_s}{p}.$$

Ez a várható érték csak akkor lesz minden i-re 0, ha teljesül a H_0 nullhipotézis, vagyis minden m_i $(i = 1, \ldots, p)$ egyenlő.

Nézzük meg, hogy a Q, Q_1, Q_{rez} négyzetösszegeknek mennyi a szabadságfoka?

A Q-t alkotó négyzetösszegek elemei között egy lineáris reláció, mégpedig a

$$\sum_{i=1}^{p} \sum_{k=1}^{r} (X_{ik} - \overline{X}_{..}) = 0$$

reláció áll fenn, így a Q szabadságfoka $(pr - 1)$.

A Q_{rez}-t alkotó négyzetösszegek között a

$$\sum_{k=1}^{r} (X_{ik} - \overline{X}_{i.}) = 0 \quad (i = 1, \ldots, p)$$

p db független lineáris összefüggés írható fel. Mivel Q_{rez} rp-számú tag négyzetösszegéből áll, ennek megfelelően Q_{rez} szabadságfoka $rp - p = p(r - 1)$. A Q_1 kifejezésében

p tag szerepel, ezek között egyetlen lineáris reláció írható fel

$$\sum_{i=1}^{p}(\overline{X}_{i.} - \overline{X}_{..}) = 0,$$

így Q_1 szabadságfoka $p - 1$.

Ekkor a $Q = Q_1 + Q_{\text{rez}}$ kapcsolatnak megfelelő

$$f = f_1 + f_2$$

összefüggés a szabadságfokok között:

$$pr - 1 = (p - 1) + p(r - 1).$$

A χ^2-partíciós tételt alkalmazva kapjuk, hogy Q_1 és Q_{rez} függetlenek, χ^2-eloszlásúak $f_1 = p - 1$, illetve $f_2 = p(r - 1)$ szabadságfokkal.

A Q_1 és Q_2 négyzetösszegekből megkaphatjuk a korrigált empirikus szórásnégyzeteket, mégpedig az:

S_1^{*2} a csoportok közötti szórásnégyzet:

$$S_1^{*2} = \frac{Q_1}{p - 1}.$$

Az S_{rez}^{*2} a csoporton belüli szórásnégyzet

$$S_{\text{rez}}^{*2} = \frac{Q_{\text{rez}}}{p(r - 1)}.$$

Ha a H_0 hipotézis fennáll, akkor S_1^{*2} és S_{rez}^{*2} várható értéke is az elméleti szórásnégyzetet, σ^2-t adja

$$E(S_1^{*2}) = E(S_{\text{rez}}^{*2}) = \sigma^2.$$

Két normális eloszlású sokaságból származó valószínűségi változó szórásnégyzetének megegyezésére az F-próbát alkalmazhatjuk, mégpedig $[(p - 1, p(r - 1)]$ szabadságfok mellett.

Ha az

$$F_{\text{emp}} = \frac{S_1^{*2}}{S_{\text{rez}}^{*2}} \quad \text{hányados}$$

nagyobb az ε-szintű és $(p - 1)$, $p(r - 1)$ szabadságfokú táblabeli elméleti F-értéknél:

$$F_{\text{emp}} > F_{\text{elm}},$$

akkor ez azt jelenti, hogy a csoportok között lényeges eltérés van; $[S_{\text{rez}}^{*2}$ csak a véletlentől függ, míg az S_1^{*2} a csoportok közötti eltéréseket is tükrözi; lásd az F-próbánál mondottakat!]

Ha elvetjük a H_0-nullhipotézist, akkor tulajdonképpen elfogadjuk, hogy az S_1^{*2}-ben szereplő

$$(\overline{X}_{i.} - \overline{X}_{..})$$

kifejezések nem mindegyikének nulla a várható értéke, azaz van az $m_i = \underset{k}{E}(X_{ik}) =$ $= E(\overline{X}_{i.})$ értékek között olyan, amely nem egyenlő a többivel.

A könnyebb áttekinthetőség céljából a kiszámítandó mennyiségeket az 4.5. szórásfelbontó táblában foglalhatjuk össze:

Szóródás oka	Négyzetösszeg	Szabadságfok
Csoportok között	$Q_1 = r\sum_{i=1}^{p}(\overline{X}_{i.} - \overline{X}_{..})^2$	$p-1$
Csoportokon belül	$Q_{\text{rez}} = Q_2 = \sum_{i=1}^{p}\sum_{k=1}^{r}(X_{ik} - \overline{X}_{i.})^2$	$p(r-1)$
Teljes	$Q = \sum_{i=1}^{p}\sum_{k=1}^{r}(X_{ik} - \overline{X}_{..})^2$	$pr-1$

4.4.3. Kétszeres osztályozás szintkombinációnként egy-egy elemű minták alapján

Esetünkben legyen az egyik tényező szintjeinek száma p $(i = 1, \ldots, p)$, a másik tényező szintjeinek száma q $(j = 1, \ldots, q)$.

Ekkor a modellünk alakja:

$$X_{ij} = \mu + \tau_i + \gamma_j + \varepsilon_{ij}.$$

A modell alkalmazásához fenn kell állni a következő feltételnek:

1. X_{ij}-k normális eloszlásúak,

$$E(X_{ij}) = \mu + \tau_i + \gamma_j, \quad \text{és}$$

$$D(X_{ij}) = \sigma.$$

2. Minden X_{ij} mintaelem független a többitől.

3. $\sum_i \tau_i = 0$ és $\sum_j \gamma_j = 0$.

4. Mindkét tényező hatása additív.

Ha γ_j becslése $\widehat{\gamma}_j$, akkor a paraméterek kiszámításához a következő négyzetösszeget kell minimalizálni:

$$Q = \sum_i \sum_j (X_{ij} - \widehat{\mu} - \widehat{\tau}_i - \widehat{\gamma}_j)^2 \longrightarrow \min.$$

Ekkor a

$$Q = Q_1 + Q_2 + Q_{\text{rez}}$$

összefüggés:

$$\sum_i \sum_j (X_{ij} - \overline{X}_{..})^2 = q\sum_i (\overline{X}_i - \overline{X}_{..})^2 + p\sum_j (\overline{X}_j - \overline{X}_{..})^2 + \sum_i \sum_j (X_{ij} - \overline{X}_{i.} - \overline{X}_{.j} - \overline{X}_{..})^2.$$

Megvizsgálhatjuk a

$$\tau_1 = \tau_2 = \ldots = \tau_i = \ldots = \tau_p = 0$$

és

$$\gamma_1 = \gamma_2 = \ldots = \gamma_j = \ldots = \gamma_q = 0$$

hipotéziseket.

A Q_1, Q_2 és Q_{rez} négyzetösszegekből a már ismert módon képezhetjük a megfelelő szórásnégyzetek becslését:

$$S_1^{*2} = \frac{Q_1}{p-1}$$

$$S_2^{*2} = \frac{Q_2}{q-1}$$

$$S_{\text{rez}}^{*2} = \frac{Q_{\text{rez}}}{(p-1)(q-1)}.$$

Az $\tau_i = 0$ (minden $i = 1, \ldots, p$) hipotézist az S_1^{*2} és az S_{rez}^{*2} összevetésével vizsgálhatjuk, ugyanis az

$$F_{\text{emp}} = \frac{S_1^{*2}}{S_{\text{rez}}^{*2}}$$

valószínűségi változó $[(p-1), (p-1)(q-1)]$ szabadságfokú F-eloszlással rendelkezik. Így, ha $(1-\varepsilon)100\%$-os szinten akarjuk vizsgálni a szóban forgó hipotézist, akkor

$$F_{\text{emp}} \leq F_{\text{elm}}$$

esetén elfogadjuk a nullhipotézist (ahol F_{elm} a megfelelő szabadságfok és valószínűségi szinthez tartozó táblabeli érték),

$$F_{\text{emp}} > F_{\text{elm}}$$

esetén pedig elutasítjuk a tényező hatástalanságára vonatkozó nullhipotézist.

Hasonló módon járunk el a

$$\gamma_j = 0 \quad (\text{minden} \quad j = 1, \ldots q)$$

nullhipotézis vizsgálatakor is, csak az S_2^{*2} és S_{rez}^{*2} értékeket vetjük össze, és ekkor az:

$$F = \frac{S_2^{*2}}{S_{\text{rez}}^{*2}}$$

$[(q-1), (p-1)(q-1)]$ szabadságfokú valószínűségi változóhoz keressük ki a megfelelő szabadságfokú és valószínűségi szintű F_{elm} táblabeli értéket, majd a döntést az előzőekhez hasonlóan hozzuk meg.

4.4.4. Az interakció

Amennyiben a tényezők (szempont) hatásainak additivitása nem teljesül, akkor azt mondjuk, a tényezők közötti interakció, vagy más szóval kölcsönhatás áll fenn. Jelentése a tényezők egymást erősítő, vagy egymást gyengítő hatásában mutatkozik meg. Tekintsünk egy kéttényezős modellt, ahol mindkét tényezőnek két-két szintje van (4.6. táblázat:)

		B tényező	
		B_1	B_2
A tényező	A_1	X_{11}	X_{12}
	A_2	X_{21}	X_{22}

4.6. táblázat.

Ebben az esetben az interakció úgy jelentkezhet például, hogy az A tényező a B_1 szinten szignifikánsan különböző értéket ad, mint a B_2 szinten.

A B_1 szinten ez a különbség $d_1 = X_{11} - X_{21}$, a B_2 szinten pedig $d_2 = X_{12} - X_{22}$.

Az A és B tényező közötti interakciót AB-vel jelöljük, és az interakció mértékének a

$$D = d_1 - d_2$$

különbséget tekinthetjük.

Ha az interakció szignifikáns voltát vizsgáljuk, és a közös σ szórásra valamilyen más forrásból becslésünk van, akkor a becslés és a $\frac{D}{2}$ érték összehasonlításával (mivel a $\frac{D}{2}$ is a σ torzítatlan becslése!) eldönthetjük, hogy az interakció nagysága lényegesnek tekintendő-e valamilyen $(1-\varepsilon)100\%$-os szinten. Az utóbbi problémát t-próbával kezelhetjük.

Ha az egyes szintkombinációkra nem egyetlen elemű mintát veszünk, hanem minden szintkombinációra r eleműt, akkor az ismeretlen, közös σ szórásra vonatkozó becslést a

$$\frac{\sqrt{r}D}{2}$$

értékkel vetjük össze, ahol a különböző szintkombinációkat az $\overline{X}_{ij}$ értékek (vagyis az illető r számú mintaelem átlaga) képviseli.

4.4.5. Kétszeres osztályozás (interakcióval) többelemű minták alapján

A modell a következő:

$$X_{ijk} = \mu + \tau_i + \gamma_j + \lambda_{ij} + \varepsilon_{ijk}$$

Teljesülnie kell a következő három feltételnek:

1. X_{ijk} értékek eloszlása normális $\underset{k}{E}(X_{ijk}) = \mu + \tau_i + \gamma_j + \lambda_{ij}$

   $D(X_{ijk}) = \sigma$ (minden i-re és minden j-re),

2. X_{ijk} értékek függetlenek egymástól,

3. $\sum_i \tau_i = \sum_j \gamma_j = \sum_j \lambda_{ij} = 0.$

A paraméterek becslését a

$$Q = \sum_i \sum_j \sum_k (X_{ijk} - \widehat{\mu} - \widehat{\tau}_i - \widehat{\gamma}_j - \widehat{\lambda}_{ij})^{2\cdot} \longrightarrow \min$$

kifejezésből kapjuk, ahol $\widehat{\lambda}_{ij}$ a λ_{ij} becslését jelöli.

A megfelelő

$$Q = Q_1 + Q_2 + Q_{12} + Q_{\text{rez}}$$

eltérésnégyzet-felbontás:

$$\sum_i \sum_j \sum_k (X_{ijk} - \overline{X}\ldots)^2 = rq \sum_i (\overline{X}_{i..} - \overline{X}\ldots)^2 + rp \sum_j (\overline{X}_{.j.} - \overline{X}\ldots)^2 +$$
$$+ r \sum_i \sum_j (\overline{X}_{ij} - \overline{X}_{i..} - \overline{X}_{.j.} + \overline{X}\ldots)^2 + \sum_i \sum_j \sum_k (X_{ijk} - \overline{X}_{ij})^2.$$

A vonatkozó „szabadságfok-egyenlet”:

$$f = f_1 + f_2 + f_3 + f_4$$

$$\underset{Q}{pqr-1} = \underset{Q_1}{(p-1)} + \underset{Q_2}{(q-1)} + \underset{Q_{12}}{(p-1)(q-1)} + \underset{Q_{raz}}{pq(r-1)}\,.$$

Látható, hogy

$$f = pqr - 1$$
$$f_1 = p - 1$$
$$f_2 = q - 1$$
$$f_3 = (p-1)(q-1)$$
$$f_4 = pq(r-1).$$

Vezessük be a nullhipotézisek jelölésére

$$H_0^{(1)}: \quad \tau_1 = \ldots = \tau_i = \ldots = \tau_p = 0$$
$$H_0^{(2)}: \quad \gamma_1 = \ldots = \gamma_i = \ldots = \gamma_q = 0$$
$$H_0^{(12)} := \lambda_{11} = \ldots = \lambda_{ij} = \ldots = \lambda_{pq} = 0.$$

A τ_i-k eltérésének mérésére vezessük be a

$$D^2(\tau_i) = \frac{\sum_i \tau_i^2}{p-1} = \sigma_1^2$$

jelölést.

Hasonló értelemben használjuk a

$$D^2(\gamma_j) = \frac{\sum_i \gamma_j^2}{q-1} = \sigma_2^2$$

és a

$$D^2(\lambda_{ij}) = \frac{\sum_i \sum_j \lambda_{ij}^2}{(p-1)(q-1)} = \sigma_{12}^2.$$

Ha nullhipotéziseink fennállnak, akkor

$$\sigma_1^2 = \sigma_2^2 = \sigma_{12}^2.$$

Ezeknek az elméleti szórásnégyzeteknek az empirikus becsléseit a következőképpen kaphatjuk:

$$\sigma_1^2 \approx S_1^{*2} = \frac{Q_1}{p-1}$$
$$\sigma_2^2 \approx S_2^{*2} = \frac{Q_2}{q-1}$$
$$\sigma_{12}^2 \approx S_{12}^{*2} = \frac{Q_{12}}{(p-1)(q-1)}.$$

A Cochran–Fisher-tétel partíciós részét alkalmazva kapjuk, hogy Q_1, Q_2, Q_{12} négyzetösszegek a $H_0^{(1)}$, $H_0^{(2)}$, $H_0^{(12)}$ nullhipotézisek fennállása esetén függetlenek, χ^2-eloszlásúak, rendre f_1, f_2, f_3 szabadságfokkal, továbbá Q_{rez} χ^2-eloszlású: f_4 szabadságfokkal.

A Q_1, Q_2, Q_{12} értékekre mondottak érvényesek a belőlük számított

$$S_1^{*2},\ S_2^{*2},\ S_{12}^{*2} = S_3^{*2}$$

értékekre is.

Mivel az S_1^{*2}, S_2^{*2}, S_3^{*2}, valamint az $S_{rez}^{*2} = \dfrac{Q_{rez}}{pq(r-1)}$ a nullhipotézisek fennállása esetén a σ^2 torzítatlan becslései, F-próbával hasonlíthatjuk értékeiket össze.

Először az interakció létével kapcsolatos $H_0^{(12)}$ hipotézist vizsgáljuk meg annak figyelembevételével, hogy $H_0^{(12)}$ fennállása esetén S_3^{*2} és S_{rez}^{*2} hányadosa

$$F_{\text{emp}} = \frac{S_3^{*2}}{S_{\text{rez}}^{*2}}$$

$[(p-1)(q-1)]; [pq(r-1)]$ szabadságfokú F-eloszlással rendelkezik.

Így, ha hipotéziseinket $(1-\varepsilon)100\%$-os szinten akarjuk vizsgálni, és a megfelelő szabadságfokú táblabeli érték F_{elm}, akkor

$$F_{emp} \leq F_{elm}$$

esetén elfogadjuk a $H_0^{(12)}$ nullhipotézist, amely szerint a két tényező között nincs kölcsönhatás.

$$F_{emp} > F_{elm}$$

esetén elutasítjuk a $H_0^{(12)}$ nullhipotézist, és azt a következtetést vonjuk le, hogy a két tényező között kölcsönhatás áll fenn. Ha elutasítjuk a $H_0^{(12)}$ hipotézist, akkor nyilván nem teljesülhetnek a $H_0^{(1)}$ és $H_0^{(2)}$ hipotézisek sem, mivel csak akkor jelentkezhet kölcsönhatás, ha a tényezők kombinációinak nem mindegyik szintjén azonos a várható érték. Ebben az esetben úgy folytathatjuk a vizsgálatot, hogy megnézzük, az egyes szintek között mekkora a várható eltérés (lásd a következő alpontot).

Ha az F-próba alapján elfogadjuk a $H_0^{(12)}$ hipotézist, akkor a következő módon folytatjuk a próbát:

Mivel Q_{12} négyzetösszeg $[(p-1)(q-1)]$ szabadságfokú, és független a $Q_{rez}[pq(r-1)]$ szabadságfokú χ^2-eloszlású valószínűségi változótól, és mindkettő a σ^2 torzítatlan becslése, összegük is a σ^2 torzítatlan becsléseként használható:

$$Q'_{rez} = \frac{Q_{12} + Q_{rez}}{pqr - p - k + 1},$$

ahol Q'_{rez} a χ^2-addíciós tétel alapján $(pqr - p - k - 1)$ szabadságfokú χ^2-eloszlású valószínűségi változó.

Így eljutottunk a

$$Q = Q_1 + Q_2 + Q_{rez}$$

típusú felbontáshoz, amelynek az

$$X_{ijk} = \mu + \tau_i + \gamma_j + \varepsilon_{ijk}$$

modell felel meg.

Ezek után a $H_0^{(1)}$ nullhipotézis ellenőrzését az

$$F_1 = \frac{S_1^{*2}}{S_{rez}^{*2}},$$

a $H_0^{(2)}$ nullhipotézis ellenőrzését pedig az

$$F_2 = \frac{S_2^{*2}}{S_{rez}^{*2}}$$

értékek segítségével végezhetjük el, ahol F_1 egy $[(p-1), (pqr-q-p+1)]$ szabadságfokú, F_2 pedig egy $[(q-1), (pqr-p-q+1)]$ szabadságfokú F-eloszlás aktuális értéke, ha fennáll a $H_0^{(1)}$, illetve F_2 esetén a $H_0^{(2)}$ nullhipotézis.

4.4.6. A „latin négyzet" módszer

A „latin négyzet" módszert alkalmazhatjuk, ha

1. három hatótényezőt emelünk ki;
2. mindegyik hatótényezőt ugyanannyi számú szinten vizsgáljuk;
3. nincs interakció.

A latin négyzet módszernek nagy előnye, hogy nem szükséges a tényezők szintjeinek minden kombinációjára mintaelemeket vennünk és vizsgálnunk, hanem már lényegesen kevesebb szintkombináció is elegendő.

Hátránya viszont, hogy igen nehezen teljesíthető az a feltétel, hogy minden tényezőt ugyanolyan számú szinten vizsgáljunk.

Ha a tényezők szintjeinek száma mindegyik tényezőnél r, akkor az előbb ismertetett módszerek alkalmazásához $rrr = r^3$ szintkombinációt kellene elemeznünk, a latin négyzet módszerénél ugyanakkor csak r^2 szintkombinációval dolgozunk.

A latin négyzet modellje

$$X_{ijk} = \mu + \tau_i + \gamma_j + \delta_k + \varepsilon_{ijk},$$

ahol $\sum \tau_i = \sum \gamma_j = \sum \delta_k = 0$.

Jelöljük a hatótényezők szintjeit A_i, B_j és C_k-val.

$$(i = 1, \ldots r; \quad j = 1, \ldots, r; \quad k = 1, \ldots, r).$$

Ha $r = n$, akkor a latin négyzet sémája pl.

	A_1	A_2	A_3	A_4
B_1	C_1	C_2	C_3	C_4
B_2	C_2	C_3	C_4	C_1
B_3	C_3	C_4	C_1	C_2
B_4	C_4	C_1	C_2	C_3

alakban írható fel. A latin négyzetek szerkesztési szabálya az, hogy minden sorban és minden oszlopban pontosan egyszer szerepeljen minden C_k szint.

Így csak azokra a szintkombinációkra kell mintát vennünk, amelyet a séma kijelöl. Pl. a második oszlop (A_2) harmadik sorában (B_3) levő C_4 azt jelenti, hogy erre a helyre az $A_2B_3C_4$ szintkombinációk melletti mintavétel eredménye kerül.

A latin négyzet módszerrel is a

$$H_A: \tau_i = 0 \quad i = 1, \ldots, r$$

$$H_B: \gamma_j = 0 \quad j = 1, \ldots, r$$

$$H_C: \delta_k = 0 \quad k = 1, \ldots, r$$

nullhipotéziseket vizsgálhatjuk.

A

$$Q = Q_1 + Q_2 + Q_3 + Q_{rez}$$

eltérésnégyzetek összegének felbontása ekkor

$$\sum_{i=1}^{r}\sum_{j=1}^{r}(X_{ijk}-\overline{X}\ldots)^2=\sum_{i=1}^{r}(\overline{X}_{i..}-\overline{X}\ldots)^2+\sum_{j=1}^{r}(\overline{X}_{.j.}-\overline{X}\ldots)^2+$$

$$\sum_{k=1}^{r}(\overline{X}_{..k}-\overline{X}\ldots)^2+\sum_{i=1}^{r}\sum_{j=1}^{r}(X_{ijk}-\overline{X}_{i..}-\overline{X}_{.j.}-\overline{X}_{..k}+2\overline{X}\ldots)^2.$$

A megfelelő szabadságfok egyenlet:

$$f=f_1+f_2+f_3+f_4$$

$$\underset{(Q)}{r^2-1}=\underset{(Q_1)}{(r-1)}+\underset{(Q_2)}{(r-1)}+\underset{(Q_2)}{(r-1)}+\underset{(Q_{raz})}{(r^2-3r+2)}\,.$$

Így a Cochran–Fisher-féle tétel szerint, ha H_A, H_B és H_C nullhipotézis igaz, akkor Q_i-k $(i=1,2,3)$ f_i szabadságfokú χ^2-eloszlással rendelkeznek, egymástól függetlenek, valamint a Q_{rez} – a H_A, H_B, H_C hipotézisek teljesülésétől függetlenül – f_4 szabadságfokú, a Q_1, Q_2, Q_3 változóktól független, χ^2-eloszlású lesz.

A nullhipotézisek ellenőrzéséhez szükséges a

$$S_1^{*2}=\frac{Q_1}{r-1}$$

$$S_2^{*2}=\frac{Q_2}{r-1}$$

$$S_3^{*2}=\frac{Q_3}{r-1}$$

és az

$$S_{rez}^{*2}=\frac{Q_{rez}}{r^2-3r+2}$$

empirikus szórásnégyzetek ismerete.

Ezek segítségével a

$$H_A\colon \tau_i=0\quad (i=1,\ldots r)$$

nullhipotézist az

$$F=F_{[(r-1,(r^2-3r+2)]}=\frac{S_1^{*2}}{S_{rez}^{*2}}$$

$[(r-1),(r^2-3r+2)]$ szabadságfokú F-eloszlású valószínűségi változó, a

$$H_B\colon \gamma_j=0\quad (j=1,\ldots,r)$$

nullhipotézist az

$$F=F_{[(r-1),(r^2-3r+2)]}=\frac{S_2^{*2}}{S_{rez}^{*2}}$$

$[(r-1),(r^2-3r+2)]$ szabadságfokú F-eloszlású valószínűségi változó, a

$$H_C\colon \delta_k=0\quad (k=1,\ldots,r)$$

nullhipotézist az

$$F=F_{[(r-1),(r^2-3r+2)]}=\frac{S_3^{*2}}{S_{rez}^{*2}}$$

$[(r-1),(r^2-3r+2)]$ szabadságfokú F-eloszlású valószínűségi változó segítségével a következő módon végezhetjük el:

ha az F érték kisebb az $1-\varepsilon$ szinthez, és $[(r-1), (r^2-3r+2)]$ szabadságfokhoz tartozó táblázatbeli

$$F_{elm} = F^{(\varepsilon)}_{[(r-1),(r^2-3r+2)]} \quad \text{értéknél,}$$

$$F \leq F_{elm},$$

akkor elfogadjuk az illető nullhipotézist (azt mondjuk, hogy a szóban forgó tényező különböző szintjei nincsenek hatással a bekövetkező eredményre), ellenkező

$$F > F_{elm}$$

esetben elvetjük a szóban forgó nullhipotézist (vagyis azt mondjuk, hogy az illető tényező különböző szintjei befolyásolják a bekövetkező eredményt), és esetleg további vizsgálatként a várható értékek közötti eltérést becsüljük (lásd később).

Ha megbízhatóbb megállapításokat akarunk tenni, akkor nem elégedhetünk meg a séma szerint előírt szintkombinációk egyetlen megfigyelt értékével (vagyis egyelemű mintával), hanem két lehetőség között választhatunk:

a) teljesen megismételjük a mintavételt (természetesen ismét csak az előírt szintkombinációkra vonatkozóan)

b) nagyobb méretű latin négyzetek esetén csak néhány szintkombinációra ismételjük meg a mintavételt.

Eljárásunk lényege a következő; a σ^2_{rez} elméleti szórásnégyzet becslésére az $S^{*2}_{\text{rez}} = \dfrac{Q_{rez}}{r^2-3r+2}$ korrigált empirikus szórásnégyzeten kívül egy másik becslést is adunk, s ezzel ellenőrizni tudjuk, hogy az S^{*2}_{rez} nem túl nagy-e. Ugyanis ha S^{*2}_{rez} indokolatlanul nagy, akkor elfedi a tényezők hatását (mivel $F = \dfrac{S^{*2}_i}{S^{*2}_{rez}}$ kisebb lesz az indokoltnál, és így nehezebben mutatható ki szignifikáns különbség.) Annak, hogy S^{*2}_{rez} túl nagy, igen gyakori oka, hogy nem teljesül a feltétel, hogy interakció, így az interakció hatása is a reziduális szórásnégyzetben halmozódik fel.

Tegyük fel, hogy h-számú szintkombináció esetén ismételjük meg a kísérletet, így rendelkezésre áll h-számú $X^{(1)}_{ijk}$ és h-számú $X^{(2)}_{ijk}$ mintaelem (ahol a felső index az ismétlésre utal). Ekkor a

$$D_{rez} = \frac{\sum \left[X^{(1)}_{ijk} - X^{(2)}_{ijk}\right]^2}{2}$$

érték is torzítatlan becslése a σ^2_{rez}-nek.

Így, ha D_{rez} és S^{*2}_{rez} jelentősen eltér, akkor azt a következtetést vonjuk le, hogy nem teljesülnek a latin négyzet alkalmazásának feltételei (mivel S^{*2}_{rez}-ben szerepel az interakció okozta eltérés is). Ebben az esetben a teljes elrendezésű szóráselemzési módszerhez kell folyamodnunk.

Az irodalomban ismert az ún. „görög–latin négyzetek” módszere, amely négy tényező hatását vizsgálja hasonló elrendezésben, mint a latin négyzet módszer.

4.4.7. A 2^n-tényezős modellek

Ha a feladatunk olyan modell felállítása, amelyben két, három, vagy négy tényező hatását vizsgáljuk, és az egyes tényezőknek csak 2–2 változatuk (szintjük) van, akkor a 2^n-tényezős modelleket alkalmazhatjuk. A 2^n-tényezős modellek használatához – ellentétben pl. a latin négyzet módszerrel – nem kell az interakció hiányát feltételezni. [Lásd pl. Meszéna–Ziermann: Valószínűségelmélet és matematika statisztika. KJK, 1981. V.7. fejezet.]

5. fejezet

Kereszttábla-elemzés és loglineáris modell

A kereszttábla-elemzés és a loglineáris modell kategorikus változók gyakoriságtábláinak elemzésére ad módszert. Ebben a fejezetben két- vagy többdimenziós táblázatokban vizsgáljuk a változók függetlenségét, és ismertetjük a sztochasztikus kapcsolat szorosságának különböző mérőszámait. Vizsgáljuk továbbá, hogy a változók együttes eloszlásait a változók milyen kapcsolódása (milyen hatások) alakítják.

5.1. Kereszttábla-elemzés

Kereszttáblában ábrázolható két (vagy több) kategorikus (nominális és ordinális mérési szintű) változó együttes valószínűségeloszlása, vagy a mintában megfigyelt együttes gyakoriságok nagysága.

Ha az A változónak I lehetséges kategóriája, a B-nek pedig J számú kategóriája van, akkor a kétdimenziós kereszttábla $I \times J$ méretű (5/a és 5/b táblázat).

	B_1	B_2	...	B_J	Perem
A_1	p_{11}	p_{12}	...	p_{1J}	p_{10}
A_2	p_{21}	p_{22}		p_{2J}	p_{20}
⋮	⋮				⋮
A_I	p_{I1}	p_{I2}		p_{IJ}	p_{I0}
Perem	p_{01}	p_{02}		p_{0J}	1

5.1/a. táblázat. Együttes valószínűségek

	B_1	B_2	...	B_J	Összesen
A_1	f_{11}	f_{12}	...	f_{1J}	f_{10}
A_2	f_{21}	f_{22}		f_{2J}	f_{20}
⋮	⋮				⋮
A_I	f_{I1}	f_{I2}		f_{IJ}	f_{I0}
Összesen	f_{01}	f_{02}		f_{0J}	$f_{00} = n$

5.1/b. táblázat. Együttes gyakoriságok

Az együttes valószínűségek sorainak és oszlopainak összegei a peremvalószínűségek. Hasonlóan képezzük a peremgyakoriságokat A és B változók egyes kategóriáira:

$$p_{i0} = \sum_{j=1}^{J} p_{ij}\,; \qquad f_{i0} = \sum_{j=1}^{J} f_{ij}, \tag{5.1}$$

$$p_{0j} = \sum_{i=1}^{I} p_{ij}\,; \qquad f_{0j} = \sum_{i=1}^{I} f_{ij}. \tag{5.2}$$

Az együttes valószínűségek teljes összege egyet, az együttes gyakoriságok kettős szummája pedig a teljes minta elemszámát adja:

$$\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij} = 1\,; \qquad \sum_{i=1}^{I}\sum_{j=1}^{J} f_{ij} = f_{00} = n. \tag{5.3}$$

Kétdimenziós kereszttáblák valószínűségeloszlásai

A kétdimenziós kereszttáblák két (A és B) változójának együttes gyakoriságait modellezve a modell függő változói az f_{ij} gyakoriságok. A gyakoriságok alakulását leíró folyamat modellezésére leggyakrabban az alábbi valószínűségeloszlásokat tételezzük fel:

– Együttes Poisson-eloszlás
– Polinomiális eloszlás
– Szorzat-polinomiális eloszlás

Az eloszlások felhasználásával a cellagyakoriságokra maximum likelihood becslés adható. A becslést és az eloszlások főbb jellemzőit foglaljuk össze a következőkben.

a) *Poisson-eloszlás*

Ha feltételezzük, hogy az f_{ij} gyakoriságok független Poisson-eloszlású valószínűségi változók λ_{ij} paraméterrel, akkor együttes eloszlásuk (együttes sűrűségük) felírható:

$$f(f_{11}, f_{12}, \ldots, f_{IJ}) = \prod_{i=1}^{I}\prod_{j=1}^{J} \frac{\lambda_{ij}^{f_{ij}} \cdot e^{-\lambda_{ij}}}{f_{ij}!} \tag{5.4}$$

A cellagyakoriságok kölcsönös függetlensége miatt az együttes eloszlás paraméterei:

$$E(f_{ij}) = \lambda_{ij} \quad \text{és} \quad Var(f_{ij}) = \lambda_{ij}, \quad Cov(f_{ij}, f_{k\ell}) = 0$$
$$\text{ha} \quad i \neq k \quad \text{és} \quad j \neq \ell,$$
$$\text{miközben} \quad i, k = 1, \ldots, I \quad \text{és} \quad j, \ell = 1, \ldots, J.$$

A λ_{ij} paraméter maximum likelihood becslése a megfigyelt gyakoriság:

$$\widehat{\lambda}_{ij} = f_{ij}.$$

A független Poisson-eloszlású változók additív tulajdonsága miatt a tábla összege $\left(\sum_i \sum_j f_{ij} = n\right)$ is Poisson-eloszlású, paramétere:

$$\lambda = \sum_i \sum_j \lambda_{ij}.$$

A Poisson-folyamat feltételezi, hogy előre nem ismerjük vagy nem rögzítjük a minta elemszámát. A Poisson-eloszlás feltételezésével felírt gyakoriságok fix n mellett polinomiális eloszlást követnek.

b) *Polinomiális*[1] *eloszlás*

Ha a végtelen alapsokasából vett minta elemszáma (n) előre adott, akkor az $I \times J$ számú cellába rendezett polinomiális eloszlású gyakoriságok együttes eloszlásának alakja

$$f(f_{11}, f_{12}, \ldots, f_{IJ}) = n! \prod_{i=1}^{I} \prod_{j=1}^{J} \frac{p_{ij}^{f_{ij}}}{f_{ij}!} = \frac{n!}{e^{-\lambda} \cdot \lambda^n} \prod_{i=1}^{I} \prod_{j=1}^{J} \frac{\lambda_{ij}^{f_{ij}} \cdot e^{-\lambda_{ij}}}{f_{ij}!}, \qquad (5.5)$$

ahol $p_{ij} = \dfrac{\lambda_{ij}}{\lambda}$, és $0 \le p_{ij} \le 1$; $\sum_i \sum_j p_{ij} = 1$ fennáll.

Az eloszlás paraméterei:

$E(f_{ij}) = n \cdot p_{ij}$ és $\operatorname{var}(f_{ij}) = n \cdot p_{ij}(1 - p_{ij})$, továbbá

$\operatorname{cov}(f_{ij}, f_{k\ell}) = -n \cdot p_{ij} \cdot p_{k\ell}$, ahol $i \neq k$ és $j \neq \ell$

valamint $i, k = 1, \ldots, I$ és $j, \ell = 1, \ldots, J$.

A cella valószínűségek maximum likelihood becslései a mintabeli arányok: $\widehat{p}_{ij} = f_{ij}/n$.

A polinomiális elosztást követő valószínűségi változók összege is polinomiális eloszlást követ, az összeg paraméterei a paraméterek összegeként határozhatók meg. A polinomiális eloszlás speciális esete a binomiális eloszlás, ha $I = 2$ és $J = 1$, azaz összesen csak két lehetséges kategóriát tételezünk fel:

$$f(f_{11}, f_{21}) = n! \frac{p_{11}^{f_{11}}}{f_{11}!} \cdot \frac{p_{21}^{f_{21}}}{f_{21}!} = \frac{n!}{f_{11}! f_{21}} \cdot p_{11}^{f_{11}} \cdot p_{21}^{f_{21}} = \binom{n}{f_{11}} \cdot p_{11}^{f_{11}} (1 - p_{11})^{n - f_{11}}$$

mert $p_{11} + p_{21} = 1$ és $f_{11} + f_{21} = n$

c) *Szorzat-polinomiális eloszlás*

Két vagy több független polinomiális eloszlású kereszttábla együttes eloszlásaként állítható elő szorzat-polinomiális eloszlás.

Kétdimenziós kereszttáblában rögzíthetjük a sorok marginálisait, azaz a *sorösszegeket*, és ezek mindegyikéről feltételezzük, hogy külön-külön polinomiális eloszlást

[1] Egyes szerzők multinomiális eloszlásnak nevezik.

követő sokaságból származnak. Az I számú sor együttes sűrűsége a sorok sűrűségének szorzata:

$$f(f_{11}, f_{12}, \ldots, f_{IJ}) = \prod_{i=1}^{I} \left[\frac{f_{i0}!}{\prod_j f_{ij}!} \prod_{j=1}^{J} \left(\frac{p_{ij}}{p_{i0}} \right)^{f_{ij}} \right] \tag{5.6}$$

A szorzat-polinomiális eloszlás a sorok mintanagyságára, f_{i0}-ra felírt feltételes eloszlás, ahol

$$\sum_j p_{ij} = 1, \quad \text{ha} \quad i = 1, \ldots, I.$$

Szorzat-polinomiális eloszlást felírhatunk úgy is, hogy az *oszlopokban* szereplő mintanagyságokat (f_{0j}) rögzítjük.

A három eloszlás alkalmazásának lehetséges eseteként tételezzük fel, hogy N számú ember válaszolta meg a kérdőívben a lakóhelyre és a jövedelem-kategóriára vonatkozó kérdést. A kétdimenziós táblában N_{ij} válaszoló jut az (i, j) cellába.

Ha úgy veszünk mintát, hogy a minta elemszáma (n) nem rögzített, akkor az egyes cellákba eső válaszolók száma Poisson-eloszlást követ.

Ha a válaszolók közül *n-elemű mintát* veszünk, és az (i, j) cellába esők száma elég nagy, akkor a mintában az f_{ij} gyakoriságok polinomiális eloszlást követnek. Ha a cellákban a megfigyelésszám kicsi, akkor hipergeometriai eloszlás tételezhető fel.

Vizsgálhatjuk a válaszolókat úgy is, hogy az *egyik változó* (pl. a lakóhely) kategóriái szerinti mintanagyságot rögzítjük, és így veszünk véletlen mintát az egyes alcsoportokból. A lakóhely- típusok szerinti mintanagyság mint feltétel mellett kiválasztott megfigyelések szorzat-polinomiális eloszlást követnek.

A kereszttáblák elemzésének több útja lehetséges. Egy alapvető kérdés – a változók függetlensége – azonban mindig felmerül. Ezt vizsgáljuk a továbbiakban.

Az A és B változó függetlenségének tesztelése

Függetlenség esetén annak valószínűsége, hogy a mintából egy esetet véletlenszerűen kiválasztva az a tábla i-edik sor j-edik oszlopába (vagyis az (i, j) cellába) esik:

$$p_{ij} = p_{i0} p_{0j} \quad (i = 1, \ldots, I; \quad j = 1, \ldots, J). \tag{5.7}$$

ahol a peremvalószínűségek a következőképpen írhatók:

$$p_{i0} = \sum_{j=1}^{J} p_{ij}; \quad p_{0j} = \sum_{i=1}^{I} p_{ij} \quad \text{és} \quad \sum_{i=1}^{I} \sum_{j=1}^{J} p_{ij=1} \tag{5.8}$$

Ha a p_{ij} elméleti valószínűségeket nem ismerjük, értékét a mintából becsülhetjük:

$$\widehat{p}_{ij} = f_{ij}/f_{00} = f_{ij}/n. \tag{5.9}$$

Ennek alapján a tábla várható gyakoriságai a függetlenség feltételezésével felírhatók:

$$F_{ij} = (f_{i0} f_{0j})/f_{00}. \tag{5.10}$$

A függetlenség tesztelése a χ^2-próbával történik:

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} [(f_{ij} - F_{ij})^2 / F_{ij}]. \tag{5.11}$$

A χ^2-statisztika $n \longrightarrow \infty$ esetén aszimptotikusan χ^2-eloszlású $(I - 1)(J - 1)$ szabadságfokkal. A χ^2-próba azon a feltevésen alapul, hogy a sokaság $(I \cdot J)$ cellát tartalmazó

polinomiális eloszlást követ. Nagy minta esetén a peremvalószínűségek becsült értékei ($\widehat{p}_{i0} = f_{i0}/n$ és $\widehat{p}_{0j} = f_{0j}/n$) normális eloszlást követnek.

A tesztfüggvény alternatív formája a likelihood aránypróba (L^2), amely szintén $(I - 1)(J - 1)$ szabadságfokú χ^2-eloszlást követ:

$$L^2 = 2 \sum_{i=1}^{I} \sum_{j=1}^{J} f_{ij} \cdot \ln \left(\frac{f_{ij} \cdot n}{f_{i0} \cdot f_{0j}} \right) \tag{5.12}$$

Nagy minták esetén a χ^2-próba és az L^2-próba egymáshoz közeli eredményeket ad a függetlenségi hipotézis vizsgálata során. Ha *szorzat-polinomiális eloszlást* tételezünk fel, akkor az f_{i0} sorösszegeket vagy az f_{0j} oszlopösszegeket előre rögzítjük. A nem rögzített peremvalószínűségek és a függetlenség feltételezésével várt gyakoriságok maximum likelihood becslései megegyeznek a három valószínűségeloszlás esetén.

A függetlenség feltételezésével rögzített peremvalószínűségek esetén lényegében a sor- vagy oszloparányok *homogenitását* teszteljük. A várt gyakoriságok becslésére felírt (5.10) képlet átrendezésével kapott arány: $\frac{F_{ij}}{f_{i0}} = \frac{f_{0j}}{n}$ azt fejezi ki, hogy a várt sorarányok minden egyes j oszlopra azonosak, azaz az I számú sorban az arányok homogének.

5.1.1. Függetlenség és asszociáció 2×2-es táblában

Legyen két változónk. Mindkét változó legyen dichotom (kétértékű, bináris). A két-két kategóriát jelölje A_1 és A_2, illetve B_1 és B_2. A megkérdezetteket (a megfigyelési egységeket) így négy típusra tudjuk bontani aszerint, hogy az (A_1, B_1), (A_1, B_2), (A_2, B_1), (A_2, B_2) kategóriák melyikébe estek.

Jelölje f_{ij} azoknak a számát, akik az (A_i, B_j) kategóriába estek. Ezeket a gyakoriságokat a következő táblázatba rendezzük (5.2. táblázat).

	B_1	B_2	összesen
A_1	f_{11}	f_{12}	f_{10}
A_2	f_{21}	f_{22}	f_{20}
összesen	f_{01}	f_{02}	$f_{00} = n$

5.2. táblázat. 2×2-es gyakorisági tábla

Az A és B változó függetlenségének tesztelése

Ha A és B függetlenek egymástól, akkor ez azt jelenti, hogy A kategóriájának ismerete semmiféle információt nem ad B kategóriájára nézve, azaz a B_1 kategóriába tartozóknak azon aránya, akik az A_1 kategóriába tartoznak, meg kell hogy egyezzen a B_2 kategóriába tartozók azon arányával, akik az A_1 kategóriához tartoznak:

$$f_{11}/f_{01} = f_{12}/f_{02} \tag{5.13}$$

vagy a sorokra felírva:

$$f_{11}/f_{10} = f_{21}/f_{20} \tag{5.14}$$

A függetlenség teszteléséhez vissza kell térnünk a kétváltozós valószínűségeloszláshoz (5.3. táblázat).

	B_1	B_2	peremvsz.
A_1	p_{11}	p_{12}	p_{10}
A_2	p_{21}	p_{22}	p_{20}
peremvsz.	p_{01}	p_{02}	$p_{00} = 1$

5.3. táblázat. Elméleti valószínűségeloszlás 2×2-es táblára

A p_{ij} elméleti valószínűség annak a valószínűségét adja, hogy véletlenszerűen választva egy megfigyelési egységet, az éppen az (i, j) cellába tartozik.

A peremeloszlásokat (5.8) alapján a következőképpen írhatjuk:

$$p_{i0} = \sum_{j=1}^{2} p_{ij}, \quad p_{0j} = \sum_{i=1}^{2} p_{ij}, \quad p_{00} = \sum_{i} \sum_{j} p_{ij} = 1. \tag{5.15}$$

Függetlenség esetén (5.13) felírható a valószínűségek hányadosaként:

$$\frac{p_{11}}{p_{01}} = \frac{p_{12}}{p_{02}} = \frac{p_{10}}{p_{00}} = p_{10}, \tag{5.16}$$

vagy az első és utolsó tag alapján:

$$p_{11} = p_{10} p_{01}. \tag{5.17}$$

Az A kategóriára akkor nincs hatással a B_1 feltétel, ha (5.14) szerint

$$\frac{p_{11}}{p_{10}} = \frac{p_{21}}{p_{20}} = p_{01}, \tag{5.18}$$

vagyis

$$p_{11} = p_{01} p_{10}. \tag{5.19}$$

Általánosságban is kimondhatjuk, hogy A és B független, ha

$$p_{ij} = p_{i0} p_{0j} \quad (i, j = 1,2). \tag{5.20}$$

Ha az (5.16) egyenlet nevezőivel beszorzunk, és figyelembe vesszük az (5.15) alapösszefüggéseket, akkor

$$p_{11}(p_{12} + p_{22}) = p_{12}(p_{11} + p_{21}). \tag{5.21}$$

Az egyenletet rendezve az esélyhányadost kapjuk, amelynek értéke függetlenség esetén 1:

$$\frac{p_{11}}{p_{21}} : \frac{p_{12}}{p_{22}} = \frac{p_{11} \cdot p_{22}}{p_{12} \cdot p_{21}} = 1. \tag{5.22}$$

A p_{11}/p_{21} hányadost *esélynek* (odds) nevezzük. Annak esélyét fejezi ki, hogy egy B_1 kategóriához tartozó személy az A_1 kategóriába esik inkább, mint az A_2 kategóriába.

Hasonlóan, p_{12}/p_{22} annak az esélye, hogy egy B_2 kategóriához tartozó egyén A_1-ben található.

A gyakorlatban sem az esélyhányadosokat, sem a tényleges valószínűségeket nem ismerjük, azok becslését állítjuk elő.

A várható gyakoriság, A és B függetlenségét feltételezve (5.10) szerint írható fel:

$$F_{ij} = f_{00} \widehat{p}_{ij} = \frac{f_{i0} f_{0j}}{f_{00}}.$$

A tényleges f_{ij} és várható gyakoriság F_{ij} alapján a χ^2-függvény tapasztalati értékét kiszámíthatjuk és összehasonlíthatjuk az elméleti mértékével, amit a χ^2 táblázatból adott

α szignifikanciaszint és a szabadságfok mellett kereshetünk meg. Amennyiben a mintából számított χ^2 értéke nagyobb, mint az elméleti érték, az adott valószínűségi szinten elvetjük a függetlenségre vonatkozó hipotézisünket.

A χ^2-próba értéke 2×2-es tábla esetén a gyakoriságból közvetlenül kiszámítható

$$\chi^2 = f_{00}(f_{11}f_{22} - f_{12}f_{21})^2/(f_{10}f_{20}f_{01}f_{02}), \tag{5.23}$$

és a szabadságfok $(2-1)(2-1) = 1$.

Tekintsünk példát az elmondottakra.

		Önértékelés		$\sum$
		1–5	6–9	
Iskolai végzettség	9–11 év	280	198	478
	12 vagy több év	141	215	356
	$\sum$	421	413	834

5.4. táblázat. Mintapélda (2×2)-es táblára

Az A változó a megfigyelt egyéni iskolai végzettsége két kategóriából áll: 9–11 évet végzettek, 12 évet vagy többet iskolába jártak csoportja. A *másik szempont* (B változó) az emberek önértékelése a hasznosság szerint, amelyet kilencfokú skálán mértek, majd két kategóriára bontottak; az egyik csoportba az 5 vagy kevesebb, a másikba az 5-nél több pontot elérők kerültek.

$$\chi^2 = \frac{834(280 \cdot 215 - 141 \cdot 198)^2}{478 \cdot 356 \cdot 421 \cdot 413} = \frac{86{,}9}{2{,}96} = 29{,}36.$$

A χ^2-táblázat alapján

$$(\chi^2_{0{,}05,1} = 3{,}841) < (\chi^2_{\text{emp}} = 29{,}36);$$

a minta alapján elvethetjük a két változó függetlenségére vonatkozó hipotézisünket.

Az asszociáció mérése 2×2-es tábla esetén

Ha a függetlenség hipotézisét elvetjük, felmerül a kapcsolat szorosságának a mérése. Számos mutatót dolgoztak ki az asszociáció mérése. Ezek közül a legáltalánosabban használtakat röviden áttekintjük. Az a)–c) mutatókat közvetlenül a gyakoriságból, a d) és e) mutatót pedig a χ^2-próba alapján számítjuk.

a) *Yule-féle Q*

Yule (1900) javasolta a következő *asszociációs együtthatót*:

$$Q = (f_{11}f_{22} - f_{12}f_{21})/(f_{11}f_{22} + f_{12}f_{21}). \tag{5.24}$$

Az együttható a (–1, 1) intervallumban ingadozik. Ha f_{00} elég nagy, Q közelítően normális eloszlású, aszimptotikus standard hibája (ASE):

$$ASE_Q = 1/2 \cdot (1 - Q^2)\left(\frac{1}{f_{11}} + \frac{1}{f_{12}} + \frac{1}{f_{21}} + \frac{1}{f_{22}}\right)^{1/2} \tag{5.25}$$

Így a Q asszociációs együtthatóra konfidenciaintervallum adható $(Q \pm ASE)$, vagy tesztelhetjük az együttható értékét a z-próbával, ahol

$$z = \frac{Q}{ASE},$$

és z aszimptotikusan (0, 1) paraméterű normális eloszlású változó.

Az iskolai végzettség és az önértékelés közötti Yule-féle asszociációs együttható értéke:

$$Q = \frac{280 \cdot 215 - 198 \cdot 141}{280 \cdot 215 + 198 \cdot 141} = 0{,}3663,$$

a standard hibája 0,0618. Így a konfidenciaintervallum alsó határa 0,3045 és felső határa 0,4281, vagyis nem tartalmazza a nullát. A z-próba értéke [$z = 0{,}3663/0{,}0618 = 5{,}927$] is megerősíti azt a következtetést, hogy a két változó közötti asszociáció szignifikáns.

b) *Esélyhányados*

A (5.22)-beli esélyhányados mintabeli becslése alkalmas az asszociáció mérésére:

$$\alpha = (f_{11} f_{22} / f_{12} f_{21}), \tag{5.26}$$

aszimptotikus standard hibája:

$$ASE_\alpha = \alpha \cdot \left[\frac{1}{f_{11}} + \frac{1}{f_{12}} + \frac{1}{f_{21}} + \frac{1}{f_{22}} \right]^{1/2} \tag{5.27}$$

Az esetleg előforduló nulla gyakoriságú cella miatt az esélyhányados korrigált formáját szokták használni:

$$\alpha = \left[\left(f_{11} + \frac{1}{2} \right) \left(f_{22} + \frac{1}{2} \right) \right] \Big/ \left[\left(f_{12} + \frac{1}{2} \right) \left(f_{21} + \frac{1}{2} \right) \right]. \tag{5.28}$$

Az α a $(0, \infty)$ intervallumban veszi fel értékeit, és $\alpha = 1$ esetén nincs asszociáció. Mivel a mutatók között szokatlan ez az értelmezés, az α logaritmusát vesszük, így $\ln(\alpha)$ értéke $(-\infty, \infty)$ között mozog, és a 0 jelenti a „nincs asszociáció” esetét. Példánkban $\alpha = 2{,}156$ ($ASE_\alpha = 0{,}328$), és $\ln \alpha = 0{,}768$.

A Yule-féle Q kifejezhető az esélyhányadossal is:

$$Q = (\alpha - 1)/(\alpha + 1). \tag{5.29}$$

c) *Goodman és Kruskal-féle τ mérték*

A Goodman és Kruskal-féle τ mérték 2×2-es tábla esetén megegyezik a χ^2 statisztikával, ha a χ^2-et elosztjuk f_{00}-lal:

$$\tau = (f_{11} f_{22} - f_{12} f_{21})^2 / (f_{10} f_{20} f_{01} f_{02}). \tag{5.30}$$

Példánkban $\tau = 0{,}035$.

A függetlenség tesztelésére használt χ^2-statisztika alkalmasnak látszik az asszociáció mérésére is. A probléma csak az, hogy a χ^2 értéke a $(0, \infty)$ intervallumba esik. Ezért különböző transzformációkkal juthatunk a (0, 1) intervallumban mozgó mutatókhoz:

d) *PHI-együttható*

A PHI-együttható lehetővé teszi azt, hogy különböző megfigyelésszámú mintákban mért asszociációt hasonlítsunk össze:

$$\phi = (\chi^2 / n)^{1/2} \tag{5.31}$$

2×2 tábla esetén $\phi^2 = \tau$ áll fenn, így példánkban $\phi = 0{,}187$.

e) *C kontingenciaegyüttható*

A χ^2-próbából számítható a Pearson-féle (1901) kontingenciaegyüttható is.

$$C = [\chi^2/(\chi^2 + n)]^{1/2}. \tag{5.32}$$

Példánkban a kontingenciaegyüttható értéke: $C = 0{,}1844$. Fontos megjegyezni, hogy a C maximuma $1/\sqrt{2}$, azaz kisebb, mint egy. Az eltérő értékkészlet miatt az asszociációs mérőszámok egymással nem hasonlíthatók össze.

5.1.2. Függetlenség és asszociáció $I \times J$-s táblában

Ha a függetlenség hipotézisét elvetettük, akkor az asszociáció mérésénél a változók két mérési típusát különböztetjük meg: a nominális változókat és az ordinális változókat. Először a nominális változók közötti asszociációs mértékeket vesszük sorra.

A. *Asszociáció mérése nominális változók esetén*

a) *Guttman-féle szimmetrikus és aszimmetrikus λ mértékek*

Vegyünk ki véletlenszerűen a mintából egy személyt, és becsüljük meg, hogy B melyik kategóriájába esik, ha

(a) nincs több információnk róla, vagy ha

(b) adott a személy A szerinti kategóriaértéke.

Ha A és B között semmiféle kapcsolat nincs, akkor a (b) eset nem ad több információt, mint (a). A λ_B a hibavalószínűség relatív csökkenését méri egy megfigyelés B változó szerinti besorolásánál, ha az A szerinti kategória ismert:

$$\lambda_B = \left(\sum_{i=1}^{I} f_{im} - f_{0m} \right) / (f_{00} - f_{0m}), \tag{5.33}$$

ahol f_{im} a tábla i-edik sorának legnagyobb gyakorisága,

f_{0m} az oszlopösszegek közül a legnagyobb érték.

A λ_B mutató kiszámításának előfeltétele az, hogy

$$\max_j f_{ij} \neq \max_j f_{0j}.$$

A λ_B tulajdonságai az alábbiak:

i) λ_B értéke nem határozható meg, ha a megfigyelések egyetlen oszlopban helyezkednek el.

ii) Ha A és B függetlenek, akkor $\lambda_B = 0$, de ez az állítás nem megfordítható. A mutató felveszi a minimumát, azaz a nullát akkor is, ha a sormaximumok azonos oszlopban helyezkednek el.

iii) A $\lambda_B = 1$ maximális értéke azt jelzi, hogy az A változó teljesen meghatározza B-t.

iv) A mutató invariáns a sorok és oszlopok felcserélésére.

A számítást az 5.5. táblázatban szereplő mintán mutatjuk be. Az A változó jelölje az egyén szakmáját, a B pedig azt, hogy az egyén melyik idegen nyelvet tudja legjobban.

	B_1	B_2	B_3	B_4	Összesen
A_1	10	5	18	20	53
A_2	8	16	5	13	42
A_3	11	7	3	4	25
Összesen	29	28	26	37	120

5.5. táblázat. Szakma és nyelvtudás szerinti együttes gyakoriságok

Ebben az esetben a λ_B azt méri, hogy az egyén szakmáját ismerve hány százalékkal kisebb hibával becsülhető az, hogy melyik nyelvet tudja jobban az illető.

$$\lambda_B = [(20 + 16 + 11) - 37]/(120 - 37) = 10/83 = 0{,}12$$

Ha a B kategóriát ismerjük, akkor az A kategória becslésére a λ_A mértéket használjuk.

A statisztika:

$$\lambda_A = \left(\sum_{j=1}^{J} f_{mj} - f_{m0} \right) / (f_{00} - f_{m0}), \tag{5.34}$$

ahol f_{mj} a tábla j-edik oszlopának legnagyobb gyakorisága,

f_{m0} a sorösszegek közül a legnagyobb.

Ha a kapcsolat szorosságának mérésénél nem emeljük ki egyik változót sem, a szimmetrikus asszociációt a következőképpen mérhetjük:

$$\lambda = \left[\left(\sum_{i=1}^{I} f_{im} - f_{0m} \right) + \left(\sum_{j=1}^{J} f_{mj} - f_{m0} \right) \right] / (2f_{00} - f_{m0} - f_{0m}). \tag{5.35}$$

A λ értéke mindig λ_A és λ_B között van.

Az 5.5. táblázat adatai alapján

$$\lambda_A = [(11 + 16 + 18 + 20) - 53]/(120 - 53) = 12/67 = 0{,}18,$$

azaz a nyelv ismeretében 18%-kal kisebb hibával becsülhető a szakmai besorolás.

Ha a két ismérv kölcsönös kapcsolatát tételezzük fel, akkor

$$\lambda = (12 + 10)/(67 + 83) = 0{,}15.$$

A legnagyobb gyakoriságok ismerete tehát 15%-kal csökkenti a besorolásnál elkövetett hibát.

b) *Goodman és Kruskal-féle τ mérték $I \times J$-s táblára*

A τ asszociációs mértéknek is van szimmetrikus és nem szimmetrikus változata, és értéke éppúgy 0 és 1 közé esik, mint a λ mutatónak. Ugyanakkor a τ mértékek minden egyes cella megfigyelt gyakoriságait felhasználva számíthatók ki, ezért értékük csak kivételesen egyezik meg a megfelelő λ mértékkel. A τ mérték aszimptotikusan χ^2-eloszlású, szabadságfoka $(I - 1)(J - 1)$. Ha az A szerinti kategóriába sorolás ismert, akkor τ_B-t

számoljuk, amely a B oszlopváltozó varianciájában bekövetkező csökkenést fejezi ki.

$$\tau_B = \frac{f_{00}\sum_{i=1}^{I}\sum_{j=1}^{J}(f_{ij}^2/f_{i0}) - \sum_{j=1}^{J} f_{0j}^2}{f_{00}^2 - \sum_{j=1}^{J} f_{0j}^2} \tag{5.36}$$

Az 5.5. táblázat példájában:

$$\tau_B = \frac{120 \cdot (36{,}057) - 3670}{14400 - 3670} = 0{,}0612$$

Tehát a szakmai kategória ismerete csak 6%-kal csökkenti az oszlopváltozó (nyelvtudás) szórásnégyzetét.

A τ_A és a τ (5.36)-hoz hasonlóan definiálható.

A közvetlenül χ^2-en alapuló mértékek közül a 2×2-es táblánál már említett PHI és C kontingenciaegyütthatók alkalmazhatók $I \times J$-s táblák esetében is. Itt két további mérőszámot mutatunk be, amelyek figyelembe veszik a kereszttábla méretét is.

c) *Csuprov-féle T-mutató*

A Csuprov-féle T-mutató a sorok és oszlopok számát is figyelembe veszi a minta nagysága mellett:

$$T = \left\{\chi^2/n \cdot \sqrt{(I-1)(J-1)}\right\}^{1/2} \tag{5.37}$$

d) *Cramer-féle V-mutató*

Cramer mutatószáma csak a tábla kisebbik méretét használja. A $q = \min(I; J)$ bevezetésével a mutató képlete:

$$V = \{\chi^2/n \cdot (q-1)\}^{1/2} \tag{5.38}$$

A V-mutató értékének minimuma 0, ha nincs kapcsolat A és B között, és maximuma 1, ha tökéletes asszociáció van a két változó között. A Cramer-V csak négyzetes táblák esetén egyezik meg a Csuprov-féle T-vel, egyébként mindig nagyobb annál. Ha $I = J = 2$, akkor $V = \phi$ is teljesül.

A Cramer-féle V standard hibája:

$$ASE(V) = [n(q-1)]^{-1/2} \tag{5.39}$$

és így a H_0: $E(V) = 0$ hipotézis tesztelhető.

A χ^2-próba elvégzéséhez előállítjuk a mintapélda (5.10) szerinti várható gyakoriságait (5.6. táblázat).

	B_1	B_2	B_3	B_4	$\sum$
A_1	12,8	12,4	11,5	16,3	53
A_2	10,15	9,8	9,1	12,95	42
A_3	6,05	5,8	5,4	7,75	25
$\sum$	29	28	26	37	120

5.6. táblázat. Várható gyakoriságok

A próbafüggvény értéke: $\chi^2 = \frac{(10-12{,}8)^2}{12{,}8} + \frac{(5-12{,}4)^2}{12{,}4} + \ldots + \frac{(4-7{,}75)^2}{7{,}75} =$
$= 22{,}95$, szabadságfok: $(3-1)(4-1) = 6$.

Mivel

$$(\chi^2_{0{,}05,0} = 12{,}592) < (\chi^2 = 22{,}95),$$

elvetjük a függetlenség hipotézisét.

Az asszociációs kapcsolat szorosságát a különböző mutatókkal mérve eltérő értékeket kapunk:

$$V = \sqrt{\frac{22{,}95}{120 \cdot 2}} = 0{,}309$$

$$T = \left\{\frac{22{,}95}{120 \cdot \sqrt{6}}\right\}^{1/2} = 0{,}279$$

$$\phi = \left\{\frac{22{,}95}{120}\right\}^{1/2} = 0{,}437$$

$$C = \left\{\frac{22{,}95}{120 + 22{,}95}\right\}^{1/2} = 0{,}401$$

B. *Az asszociáció mérése információelméleti alapon*

Ha a nominális változók lehetséges kategóriái nem rendezettek, akkor az entrópia-mérőszámok segítségével fejezhetjük ki a valószínűségeloszlás rendezetlenségének mértékét.

Egy változó kategóriái szerinti rendezetlenség mérésére több mutató is használható. Témánk szempontjából elegendő a Shannon–Wiener-féle entrópia-indexet említeni:

$$H = -\sum_{i=1}^{I} p_i \ln p_i \tag{5.40}$$

Két változó esetén az együttes „bizonytalanságot” a Kullback-féle kölcsönös entrópia méri, amely az együttes valószínűségek logaritmusait véve határozható meg.

$$H_{12} = -\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij} \cdot \ln[p_{ij}/p_{i0}p_{0j}] \tag{5.41}$$

A valószínűségeket a relatív gyakoriságokkal becsülve a kölcsönös entrópia mérőszáma az alábbi alakot ölti:

$$H_{12} = -\sum_{i=1}^{I}\sum_{j=1}^{J} \frac{f_{ij}}{f_{00}} \cdot \ln\left[\frac{f_{ij} \cdot f_{00}}{f_{i0} \cdot f_{0j}}\right]. \tag{5.42}$$

Ha figyelembe vesszük azt az arányt, amivel az egyik – (B) – változó kategóriájának ismerete csökkenti a másik – (A) – változó szerinti besorolás bizonytalanságát, akkor a

bizonytalansági együtthatót[2] kapjuk, amely az entrópia relatív mérőszáma:

$$U_A = \frac{H_{12}}{H_A} = \frac{\sum_i \sum_j \frac{f_{ij}}{f_{00}} \cdot \ln\left(\frac{f_{i0} \cdot f_{0j}}{f_{ij} \cdot f_{00}}\right)}{\sum_i \frac{f_{i0}}{f_{00}} \cdot \ln\left(\frac{f_{i0}}{f_{00}}\right)}$$

U_A értéke 0 és 1 között változik. A nulla azt fejezi ki, hogy a sorváltozó varianciája nem csökkent attól, hogy az oszlopváltozó kategóriája ismert. Ha $U_A = 1$, akkor az oszlopkategória ismerete teljesen megszüntette a sorváltozó szerinti besorolás bizonytalanságát. Az asszimptotikus standard hiba segítségével konfidenciaintervallum írható fel a bizonytalansági együtthatóra. Az asszociáció hiányát feltételező nullhipotézis tesztelésekor arra a tényre építünk, hogy U az $(I-1)(J-1)$ szabadságfokú χ^2 eloszláshoz konvergál. A bizonytalansági együtthatót az 5.5. táblázat megfigyelt gyakoriságait és az 5.6. táblázat várt gyakoriságait felhasználva számítjuk ki.

$$U_A = \frac{10 \cdot \ln \frac{12{,}8}{10} + 5 \cdot \ln \frac{12{,}4}{5} + \ldots + 4 \cdot \ln \frac{7{,}75}{4}}{53 \cdot \ln \frac{53}{120} + 42 \cdot \ln \frac{42}{120} + 25 \cdot \ln \frac{25}{120}} = \frac{-11{,}604}{-126{,}612} = 0{,}092$$

A nyelvtudás ismeretében 9%-kal kisebb bizonytalansággal következtethetünk az egyén szakmájára.

Fordított hatást feltételezve 7%-kal csökken a bizonytalanság, mivel a szakma ismeretében vizsgálva a nyelvtudás szerinti besorolást:

$$U_B = \frac{H_{12}}{H_B} = \frac{-11{,}604}{29 \cdot \ln \frac{29}{120} + 28 \cdot \ln \frac{28}{120} + 26 \cdot \ln \frac{26}{120} + 37 \cdot \ln \frac{37}{120}} = \frac{-11{,}604}{-165{,}231} = 0{,}07$$

adódik.

C. *Az asszociáció mérése ordinális változók esetén*

A következőkben olyan táblákkal foglalkozunk, amelyekben az A és B változó kategóriái rendezettek. Ez azt jelenti, hogy ha valaki az A változó i-edik kategóriájába kerül, magasabbra rangsorolt (preferáltabb), mint aki a k-adik kategóriába kerül.

Tekintsük a megfigyelt személyek egy általános párosítását. Az egyik személy tartozzon az (i, j) cellához, vagyis az A változó i-edik kategóriájához és a B változó j-edik kategóriájához. A másik személy pedig kerüljön a (k, ℓ) cellába.

Az asszociáció ordinális mértéke a következő négy mennyiségnek a függvénye:

– S = a személyek (megfigyelési egységek) azon párjainak a teljes száma, amelyekre vagy $i > k$ és $j > \ell$, vagy $i < k$ és $j < \ell$ teljesül.

$$S = \sum_{i=1}^{I} \sum_{j=1}^{J} f_{ij} \left(\sum_{k>i} \sum_{\ell>j} f_{k\ell} \right) \tag{5.43}$$

– D = a megfigyelések azon párjainak a teljes száma, amelyekre vagy $i > k$ és $j < \ell$, vagy $i < k$ és $j > \ell$ teljesül.

$$D = \sum_{i=1}^{I} \sum_{j=1}^{J} f_{ij} \left(\sum_{k>i} \sum_{\ell<j} f_{k\ell} \right) \tag{5.44}$$

[2] A bizonytalansági együttható a likelihood-arány teszthez hasonló tartalmú mérőszám.

– T_A = a megfigyelések azon párjainak a teljes száma, amelyekre $i = k$ teljesül.

– T_B = a megfigyelések azon párjainak a teljes száma, amelyekre $j = \ell$ teljesül.

Az A és B változó erős pozitív asszociációja esetén S értéke nagy és D értéke kicsi lesz.

Így természetes, hogy az asszociációt S és D különbségével, ennek a különbségnek valamilyen standardizálásával célszerű mérni.

a) *Goodman és Kruskal-féle* γ

Goodman és Kruskal (1954) javasolta a következő mértéket két változó szimmetrikus kapcsolatának mérésére:

$$\gamma = (S - D)/(S + D). \tag{5.45}$$

A γ mértéknek valószínűségi értelmezése is van. Annak a valószínűségéből, hogy a mintából véletlenszerűen kiválasztott két megfigyelés hasonlóan rendezett, kivonjuk annak a valószínűségét, hogy nem hasonlóan rendezett, eltekintve azoktól a pároktól, amelyek valamelyik változó azonos kategóriájába esnek.

A γ a $(-1,1)$ intervallumban veheti fel értékét. Ha A és B függetlenek, γ értéke nulla, azonban az állítás nem megfordítható. A $\gamma = 1$, ha $D = 0$, azaz a megfigyelések mindkét index szerint hasonlóan rendezettek. A $\gamma = -1$, ha $S = 0$. Ekkor az egyik változó szerinti preferáltság a mások szerinti kedvezőtlenebb besorolással jár együtt, azaz a megfigyelések ellentétesen rendezettek.

A γ értéke nem határozható meg, ha a megfigyelések a kereszttábla egyetlen sorában vagy oszlopában koncentrálódnak.

A γ eloszlása közelítőleg normális. A γ-mérték 2×2-es tábla esetén megegyezik a Yule-féle Q-val.

Tekintsük az 5.7. táblázatban szereplő két ordinális változót,[3] és jellemezzük kapcsolatukat a γ-mutatóval.

		Önértékelés, hasznosság		
		1–5	6–9	Összesen
Iskolai végzettség	9–11 év	280	198	478
	12 év	37	44	81
	13 vagy több év	104	171	275
	Összesen	421	413	834

5.7. táblázat. Kétszempontos gyakorisági táblázat

A mutató kiszámításához szükségünk van S és D értékére:

$$S = 280(44 + 171) + 37 \cdot 171 = 66\,527$$

és $$D = 198(37 + 104) + 44 \cdot 104 = 32\,494$$

A Goodman–Kruskal-mutató értéke tehát:

$$\gamma = \frac{66\,527 - 32\,494}{66\,527 + 32\,494} = \frac{34\,033}{99\,021} = 0{,}3437,$$

amely szerint közepesen szoros együttjárás tapasztalható az iskolai végzettség és az önértékelés között.

[3] Az 5.4. táblázatban ugyanez a példa szerepel, de itt a sorváltozót 3 kategóriára bontva adjuk meg.

b) *Kendall és Stuart-féle τ mértékek*

Kendall (1962) mértéke figyelembe veszi az azonos kategóriába eséseket mindkét változó esetén:

$$\tau = \frac{S - D}{\left[\frac{1}{2}(S + D + T_A) \cdot \frac{1}{2}(S + D + T_B)\right]^{1/2}}, \tag{5.46}$$

és standard hibája:

$$ASE(\tau) = \left[\frac{4n + 10}{9(n^2 - n)}\right]^{1/2}. \tag{5.47}$$

A τ és a γ mutató azonos minta esetén csak a nevezetes értékekre (0 és ± 1) egyezik meg. A közbülső tartományokon a γ abszolút értékben magasabb a τ-nál.

A számítógépes programcsomagok τ_b jelöléssel megkülönböztetik a négyzetes ($I = J$) táblára számított Kendall-féle τ-t a Stuart-féle τ_c-től, amely $I \neq J$ esetén számítandó.

Ha $q = \min(I, J)$, akkor Stuart (1967) mértéke az alábbi:

$$\tau_c = \frac{2q(S - D)}{n^2(q - 1)} \tag{5.48}$$

Példánkban $\tau_c = \dfrac{2 \cdot 2 \cdot 34\,033}{834^2 \cdot (2 - 1)} = \dfrac{136\,132}{695\,556} = 0{,}1957$; azaz jóval alacsonyabb, mint a γ.

c) *Somer-féle aszimmetrikus d mutató*[4]

Somer (1962) mutatóját számolva figyelembe vehető a változók közötti függőség.

Ha a B a függő változó, akkor a mutató:

$$d_B = \frac{S - D}{S + D + T_B}. \tag{5.49}$$

A d_B kiszámításánál azt tételezzük fel, hogy az A változó szerint nincsenek kategória-egyezések ($T_A = 0$). A mutató nem számítható ki, ha a megfigyelések egyetlen sorban vagy oszlopban sűrűsödnek. A $d_B = 0$, ha a két változó független, de az állítás nem megfordítható.

d) *Spearman-féle rangkorreláció*

Ordinális változók szimmetrikus kapcsolatának szorosságát méri a Spearman-féle rangkorrelációs együttható:

$$r_S = \frac{12 \cdot \sum_i \sum_j f_{ij} \left[\frac{f_{i0}}{2} - \frac{n}{2} + \sum_{k<i} f_{i0}\right] \left[\frac{f_{0j}}{2} - \frac{n}{2} + \sum_{\ell<j} f_{0j}\right]}{\left\{\left[n^3 - n - \sum_i (f_{i0}^3 - f_{i0})\right] \left[n^3 - n - \sum_j (f_{0j}^3 - f_{0j})\right]\right\}^{1/2}} \tag{5.50}$$

[4] Ha nominális változókra számítjuk ki a Somer-féle d mutatót, akkor a nominális változó n kategóriája $n!$-féleképpen rendezhető. Így $n!$ számú Somer-féle d-t kapunk. Ezek közül a maximális a Freeman-együttható.

A mutató standard hibája:

$$ASE(r_S) = \left[\frac{1 - r_S^2}{n - 2}\right]^{1/2}, \tag{5.51}$$

így a $z = \dfrac{r_S}{ASE(r_S)}$ mérőváltozóval tesztelhetjük a kapcsolat hiányát feltételező $H_0 : r_S = 0$ hipotézist.

A Spearman-féle rangkorreláció $(-1; +1)$ között mér, jelezve ezzel nemcsak a kapcsolat erejét, de irányát is.

A Spearman-féle rangkorreláció akkor is méri két ordinális változó között a kapcsolat szorosságát, ha nem rendezzük kereszttáblába a megfigyeléseket, csak a két változó szerinti sorrendet ismerjük. Ekkor a mutató az alábbi képlettel számítható ki:

$$r_S = 1 - \frac{6 \cdot \sum_{i=1}^{n} di^2}{n^3 - n}, \tag{5.52}$$

ahol di az i-edik egyed rangszámainak különbsége.

Az $r_S = 1$, ha a két változó szerinti sorrend teljesen megegyezik, és $r_S = -1$, ha a két sorrend teljesen fordított.

Ez a mutató a Pearson-féle lineáris korreláció speciális esetének tekinthető.

5.1.3. Az asszociációs mértékek közötti választás problémája

Az asszociációs mértéktől azt várjuk, hogy egyszerű numerikus értékével megmondja, két változó milyen szorosan kapcsolódik egymáshoz. Ha A és B asszociációs mértéke 0,4, és a C és D közötti asszociáció 0,6, akkor a C és D változó szorosabb kapcsolatban van egymással, mint A és B. Azonban, ha egy másik mutatóval mérjük ugyanezt, lehet, hogy A és B között mérjük a 0,6, és C és D között a 0,4 asszociációs mértéket. Így a válasz pontosan a fordítottja az előzőnek.

Ez azért fordulhat elő, mivel mindegyik mutató a kapcsolódás valamelyik aspektusára figyel, ezek pedig különböznek, így a mutatók között nincs „legjobb". Kruskal (1958) véleménye szerint egyszerre több különböző típusú asszociációs mértéket kell számítani a közülük történő választás helyett.

Ugyanakkor az asszociációs mértékek mintabeli becslései helyett helyesebb lenne azok konfidenciaintervallumait megadni, amely intervallum valamilyen magas (például 95%-os) valószínűségi szinten tartalmazza a becsülni kívánt asszociációs mértéket.

A számítógépes programcsomagok a λ, a Goodman–Kruskal-féle τ, a bizonytalansági együttható, a Kendall-féle τ, a γ és a Somer-féle d mutatók értékei mellé megadják ezek standard hibáit is. Ezen értékekből felírható a konfidenciaintervallum, vagy hipotézisvizsgálat végezhető arra a nullhipotézisre, hogy a két változó között nincs asszociáció.

A mutatók több szempont szerinti csoportosítása segítségünkre lehet a megfelelő mérték kiválasztásában.

Az asszociációs mérőszámokat négyféleképpen csoportosíthatjuk:

1. Melyik az a legalacsonyabb mérési skála, amelyre alkalmazható?

• nominális skálán mér: α esélyhányados, ϕ, C kontingenciaegyüttható, Guttman-féle λ-k, Yule-féle Q, Goodman–Kruskal τ mértékei, Cramer-V, Csuprov-féle T, bizonytalansági együttható.

• ordinális skálán mér: Goodman–Kruskal-féle γ, Kendall-féle τ_b és Stuart-féle τ_c, Somer d_B-je, Spearman rangkorrelációja.

• intervallumskálán mér: Pearson korrelációs együtthatója.

2. Jelzi-e a kapcsolat irányát a mutató előjele?

• Nem jelzi, azaz 0 és 1 között mér: ϕ, C, V, λ, Goodman–Kruskal τ, bizonytalansági együttható.

• Nem jelzi, de más a maximuma: Csuprov T $(0; 1/\sqrt{2})$ között, α esélyhányados $(0; \infty)$.

• Jelzi, –1 és+1 között mér: Yule-Q, γ, τ_b, τ_c, d_B, rangkorreláció, korreláció.

3. Milyen módon számítható?

• csak a legnagyobb gyakoriságok figyelembevételével: Guttman-féle λ

• a kereszttábla minden celláját közvetlenül felhasználva: α, Q, Goodman–Kruskal τ, γ, τ_b, τ_c, d_B, rangkorreláció, bizonytalansági együttható.

• a χ^2 értékéből közvetetten: ϕ, C, V, T.

4. A változók szerepét megkülönbözteti?

• Szimmetrikus mérték, nincs különbség a változók között: λ, τ, τ_b, τ_c, α, Q, ϕ, C, V, T, γ.

• Aszimmetrikus mérték, függőségi viszonyt feltételez a változók között: λ_A, λ_B, τ_A, τ_B, d_B, bizonytalansági együttható.

5.1.4. Függetlenség és asszociáció sokdimenziós táblában

A következőkben több mint két változó szerint soroljuk a megfigyeléseket közös cellákba, csoportokba. Háromdimenziós (háromutas) táblában három változó (A, B és C) szerint rendezzük a megfigyeléseket.

A három változó kategóriáinak a száma I, J és K.

Jelölések a háromutas táblában

A megfigyelési egységeket $I \times J \times K$ cella valamelyikébe helyezzük. Az (i, j, k) cellába kerülő megfigyelések száma f_{ijk}. A megfelelő valószínűségeket p_{ijk} jelöli. A háromutas táblában két különböző peremgyakoriságunk van. Ha egy változót pl. az A kategóriái szerint összegzünk, akkor:

$$f_{0jk} = \sum_{i=1}^{I} f_{ijk}, \tag{5.53}$$

ha A-ra és B-re is összegzünk, akkor

$$f_{00k} = \sum_{j=1}^{J} f_{0jk} = \sum_{i=1}^{I} f_{i0k} = \sum_{i} \sum_{j} f_{ijk} \tag{5.54}$$

a peremgyakoriság.

A megfelelő valószínűségek jelölése rendre p_{0jk} és p_{00k}.

Háromutas táblában a függetlenséget többféleképpen értelmezhetjük.

Két változó függetlenségét (5.7) mondja ki. Ennek közvetlen kiterjesztése három változó esetére a kölcsönös függetlenség:

$$p_{ijk} = p_{i00}p_{0j0}p_{00k}, \quad i = 1, \ldots, I; \quad j = 1, \ldots, J; \quad k = 1, \ldots, K. \tag{5.55}$$

Az A, B és C változók *kölcsönösen függetlenek*, ha a fenti egyenlet minden cellára fennáll.

A és B *feltételesen függetlenek* a C változó adott kategóriájára, ha

$$p_{ijk} = p_{i0k}p_{0jk}/p_{00k}. \tag{5.56}$$

A C *többszörösen független* A és B-től, ha az A és B változók a C változó minden kategóriájában azonos asszociációt mutatnak, vagyis C nincs hatással (AB)-re. Matematikailag:

$$p_{ijk} = p_{ij0}p_{00k}. \tag{5.57}$$

Tekintsünk a függetlenség értelmezésére és tesztelésére egy példát[5] *a háromutas táblában.*

65 hallgató vizsgázott algebrából (A), analízisből (B) és statisztikából (C). Akik valamely vizsgán átmentek, az A_1, B_1 és C_1 kategóriába tartoznak. A_2, B_2 és C_2 jelöli a bukást az adott tárgyból.

Az együttes gyakoriságok tehát a három tárgyból elért eredmény-kombináció előfordulását fejezik ki. Az f_{111} azok száma, akik három sikeres vizsgát tettek. Az f_{121}, f_{112} és f_{211} gyakoriságok egy-egy bukást és két sikeres vizsgát fejeznek ki. Az együttes előfordulásokat a 5.8. táblázat mutatja.

			Analízis		
			B_1	B_2	Összesen
Algebra	A_1 Statisztika	C_1	16	3	19
		C_2	8	4	12
		Összesen	24	7	31
	A_2 Statisztika	C_1	12	6	18
		C_2	9	7	16
		Összesen	21	13	34
		Mindösszesen	45	20	65

5.8. táblázat. Háromutas tábla gyakoriságai

A tábla szabadsági foka $(I-1)(J-1)(K-1) = 1$.

Először vizsgáljuk meg a három változó *kölcsönös függetlenségét*. Ekkor a várt gyakoriságokat az

$$F_{ijk} = n \cdot \frac{f_{i00}}{n} \cdot \frac{f_{0j0}}{n} \cdot \frac{f_{00k}}{n} \tag{5.58}$$

alapján becsüljük. Az F_{111} értéke tehát a függetlenség feltételezésével a következő:

$$\frac{31 \cdot 45 \cdot 37}{65^2} = 12{,}22$$

[5] A példa eredeti változata megtalálható Nguyen–Rogers [1989] könyvének 386. oldalán.

és F_{222} becslése:

$$\frac{34 \cdot 20 \cdot 28}{65^2} = 4{,}51$$

A χ^2-próba értéke:

$$\chi^2 = \frac{(16 - 12{,}22)^2}{12{,}22} + \ldots + \frac{(7 - 4{,}51)^2}{4{,}51} = 4{,}08$$

A χ^2 táblában 1 szabadságfok és 5%-os valószínűségi szint[6] mellett a kritikus érték (3,84) alacsonyabb, mint a mintabeli érték, ezért a háromutas kölcsönös függetlenséget elvetjük.

Két-két tantárgy vizsgaeredményeinek *páronkénti függetlenségét* is vizsgálhatjuk.

Az algebrát és az analízist vizsgálva az alábbi összevont táblán számoljuk a χ^2_{AB} értékét.

		Analízis		
		B_1	B_2	
Algebra	A_1	24	7	31
	A_2	21	13	34
		45	20	65

$$\chi^2_{AB} = \frac{(24 - 21{,}46)^2}{21{,}46} + \ldots + \frac{(13 - 10{,}46)^2}{10{,}46} = 1{,}866.$$

Mivel a szabadságfok itt is 1, a függetlenség hipotézisét nem vetjük el.

Az algebra és a statisztika esetében

$$\chi^2_{AC} = 0{,}461,$$

míg az analízis és statisztika vizsgákon

$$\chi^2_{BC} = 1{,}675$$

adódik.

Így a páronkénti függetlenség hipotézisét a két utóbbi esetben sem vetjük el.

A példa fontos tanulsága, hogy a páronkénti függetlenségből nem következik a kölcsönös függetlenség fennállása.

Simpson-féle paradoxon

Simpson (1951) mutatott rá először arra, hogy egy sokdimenziós táblában, ha egy változó kategóriái szerinti táblákat összesítjük, a többi változó asszociációja ellenkezőjére is változhat. Például a C_1 és C_2 kategóriák szerinti kétdimenziós táblákban A és B asszociációja pozitív lehet, míg összesítve a táblát (kategóriát), A és B asszociációja negatívvá válhat.

Általánosan is megállapítható, hogy A, B és C változókra felírt háromdimenziós táblában csak akkor vonhatjuk össze a C változó kategóriáit, ha C független legalább A-tól vagy B-től. Ellenkező esetben a C változó kiküszöbölésével az interakciók megváltoznak, és ezért lép fel a Simpsonról elnevezett ellentmondás.

[6] Fogalmazhatunk úgy is, hogy a 4,08-hoz tartozó empirikus szignifikanciaszint kisebb, mint 5%. Ezért a nullhipotézist elvetjük $\alpha = 0{,}05$ mellett.

Tekintsünk példát az ellentmondás fellépésére:

Egy egyetem két karára jelentkező és felvételt nyerő hallgatókat nemek szerinti bontásban is vizsgáljuk.[7] Az 5.9. táblázatban látható, hogy a felvett lányok aránya alacsonyabb összességében, mint a fiúké. Ha karonkénti bontásban vizsgáljuk a nemek arányát, akkor éppen az ellenkező arányt tapasztaljuk; mindkét karon a fiú hallgatók kerültek kisebb arányban felvételre.

	lány	fiú
Közgazdasági kar		
Jelentkezők száma	160	60
Felvettek száma	40	12
Felvettek aránya	0,25	0,20
Mérnöki kar		
Jelentkezők száma	40	140
Felvettek száma	26	84
Felvettek aránya	0,65	0,60
Karok együtt		
Jelentkezők száma	200	200
Felvettek száma	66	96
Felvettek aránya	0,33	0,48

5.9. táblázat. A felvételi arány nemek szerinti vizsgálata

Ha tehát együtt elemezzük a két kart (azaz összevonjuk a kategóriákat), akkor eltűnik az a jelentős különbség, ami a jelentkezők összetételében karonként kimutatható.

A jelentkezők száma alapján nem független a hallgató neme és a választott kar. (Zárójelben a várt gyakoriságok szerepelnek.)

Karok	Lány	Fiú	
Közgazdasági	160	60	220
	(110)	(110)	
Műszaki	40	140	180
	(90)	(90)	
	200	200	400

$\chi^2 = 101{,}0$ szf: 1

$\phi = \sqrt{\frac{\chi^2}{n}} = 0{,}5025$

A felvettek száma alapján sem független a hallgatók neme és az egyetem kara (zárójelben itt is a várt gyakoriságok kerekített értékei szerepelnek.)

Karok	Lány	Fiú	
Közgazdasági	40	12	52
	(21)	(31)	
Műszaki	26	84	110
	(45)	(62)	
	66	96	162

$\chi^2 = 34{,}1$ szf: 1

$\phi = 0{,}44588$

[7] A példa eredeti változata J. D. Jobson (1992) 46. oldalán szerepel.

Ha eltekintünk a karok szerinti bontástól, akkor a jelentkezettek-felvettek és a nemek kapcsolata nem szignifikáns:

Karok	Lány	Fiú	
Jelentkezett	200	200	400
	(189)	(211)	
Felvett	66	96	162
	(77)	(85)	
	266	296	562

$\chi^2 = 4{,}20$ szf: 1
Empirikus szignifikancia szint:
$p = 0{,}04$

A fenti példában a karok szerinti elemzés kimutatta az asszociáció létét, míg a kari adatok összevonása elfedte az összefüggést. Ha nem nominális, hanem ordinális mérési szintű adatokat elemzünk, akkor a kapcsolat előjelének megfordulására is számíthatunk.

5.2. A loglineáris modell

A gyakorisági táblák vizsgálatára Birch (1963) javasolta először a loglineáris modellt. Goodman (1964, 1979), Haberman (1974), Bishop, Fienberg és Holland (1975) a legismertebbek a későbbi szerzők közül.

A többdimenziós kereszttáblákra nagyszámú hipotézist fogalmazhatunk meg, amelyek a változók különböző kapcsolódásait, kapcsolatrendszerét írják le. Így feltételezhetjük, hogy a változók függetlenek egymástól, vagy hogy páronkénti kapcsolódásaik (interakcióik) függetlenek a többi változótó stb. A hipotézisek szerint ezen feltételezett hatások szorzata eredményezi az egyes cellák megfigyelt gyakoriságait. A loglineáris modell a cellák gyakoriságainak logaritmusait veszi, így a modellt a hatások lineáris függvényeként írja fel. A loglineáris modell általában nem tesz különbséget a változók között függőségi viszony szerint, csak a változók kapcsolatrendszerének struktúráját elemzi. A fejezet végén ismertetésre kerülő logit-modell lehetőséget ad arra, hogy a magyarázó változóknak (factors) egy függő (response) változóra gyakorolt hatását vizsgáljuk.

5.2.1. Loglineáris modellek étdimenziós kereszttáblára

Jelöljük a kereszttábla két változóját A-val és B-vel. Az A változó kategóriáinak a számát jelölje I ($i = 1, \ldots, I$), a B változó kategóriáinak a számát pedig J ($j = 1, \ldots, J$). A $P(A = i)$ jelentse annak a valószínűségét, hogy véletlenszerűen választva egyet a megfigyelések közül, a kiválasztott egyed éppen az A változó i-edik kategóriájába esik. Az elméleti valószínűségek becsléseit a kontingenciatáblázat megfigyelt gyakoriságaiból számítjuk. Ha f_{i0} az A változó peremgyakoriságait (marginálisait) jelenti, akkor $P(A = i) = f_{i0}/n$, ahol n a megfigyelések száma (a minta elemszáma). Hasonlóan, $P(B = j) = f_{0j}/n$, ahol f_{0j} jelöli a B változó peremgyakoriságait. Az f_{ij} a mintában az $A = i$ és $B = j$ együttes előfordulásainak a száma. Ha a két változó független, akkor annak a valószínűsége, hogy egy megfigyelés éppen az (i, j) cellába fog esni:

$$p_{ij} = P(A = i)P(B = j) = (f_{i0}/n)(f_{0j}/n). \tag{5.59}$$

A loglineáris modellt legkönnyebben a függetlenségi hipotézisből kiindulva írhatjuk fel. A χ^2-statisztika[8] a kereszttábla megfigyelt gyakoriságai és a függetlenség hipotézisének feltételezésével számolt várható gyakoriságok közötti különbséget méri:

$$\chi^2 = \sum_i^I \sum_j^J \frac{(f_{ij} - F_{ij})^2}{F_{ij}}, \tag{5.60}$$

ahol $F_{ij} = E(f_{ij}) = n \cdot p_{ij} = n \frac{f_{i0}}{n} \frac{f_{0j}}{n}$. Vegyük mindkét oldal természetes alapú logaritmusát, hogy lineáris alakot kapjunk:

$$\ln F_{ij} = \ln f_{i0} + \ln f_{0j} - \ln n. \tag{5.61}$$

Szummázzuk az (5.61) egyenletet először i, utána j, majd mindkét index szerint, hogy A és B változók hatását kifejezhessük:

$$\sum_i \ln F_{ij} = \sum_i \ln f_{i0} + I \ln f_{0j} - I \ln n,$$

ebből

$$\ln f_{0j} = \frac{1}{I} \sum_i \ln F_{ij} - \frac{1}{I} \sum_i \ln f_{i0} + \ln n. \tag{5.62}$$

$$\sum_j \ln F_{ij} = J \ln f_{i0} + \sum_j \ln f_{0j} - J \ln n,$$

ebből

$$\ln f_{i0} = \frac{1}{J} \sum_j \ln F_{ij} - \frac{1}{J} \sum_j \ln f_{0j} + \ln n. \tag{5.63}$$

$$\sum_i \sum_j \ln F_{ij} = J \sum_i \ln f_{i0} + I \sum_j \ln f_{0j} - IJ \ln n,$$

ebből

$$-\ln n = \frac{1}{IJ} \sum_i \sum_j \ln F_{ij} - \frac{1}{I} \sum_i \ln f_{i0} - \frac{1}{J} \sum_j \ln f_{0j}. \tag{5.64}$$

Helyettesítsük vissza (5.62), (5.63) és (5.64) egyenleteket az (5.61) egyenletbe:

$$\ln F_{ij} = \frac{1}{J} \sum_j \ln F_{ij} + \frac{1}{I} \sum_i \ln F_{ij} - \frac{1}{IJ} \sum_i \sum_j \ln F_{ij} \tag{5.65}$$

Az egyenletet bővítve:

$$\ln F_{ij} = \frac{1}{IJ} \sum_i \sum_j \ln F_{ij} + \left(\frac{1}{J} \sum_j \ln F_{ij} - \frac{1}{IJ} \sum_i \sum_j \ln F_{ij} \right) +$$

$$+ \left(\frac{1}{I} \sum_i \ln F_{ij} - \frac{1}{IJ} \sum_i \sum_j \ln F_{ij} \right),$$

és a következő jelöléseket bevezetve:

$$u = \frac{1}{IJ} \sum_i \sum_j \ln F_{ij}, \tag{5.66}$$

[8] Az (5.11)-ben bevezetett értelmezéssel megegyező módon.

$$u_{A(i)} = \frac{1}{J}\sum_j \ln F_{ij} - u, \tag{5.67}$$

$$u_{B(j)} = \frac{1}{I}\sum_i \ln F_{ij} - u \tag{5.68}$$

a *két változó függetlenségének hipotézisére felírt loglineáris modell* általános alakját kapjuk:

$$\ln F_{ij} = u + u_{A(i)} + u_{B(j)}. \tag{5.69}$$

Eszerint az (i, j) cellában a várható gyakoriságok logaritmusa lineáris függvénye a várható gyakoriságok logaritmusai átlagának, az A változó i-edik kategóriahatásának és a B változó j-edik kategóriahatásának. A loglineáris modellt összevetve a varianciaelemzés logikájával, az u jelenti az átlagos hatást, az $u_{A(i)}$ és $u_{B(j)}$ pedig a főhatásokat. Mivel a főhatásokat az átlagtól való eltérésekkel mérjük, a szóráselemzéshez (ANOVA) hasonlóan feltételezzük, hogy

$$\sum_i u_{A(i)} = \sum_j u_{B(j)} = 0. \tag{5.70}$$

Az (5.69) modell abból a feltételezésből indult ki, hogy az A és B változó független, és így az (i, j) cella valószínűsége a peremvalószínűségek szorzataival egyenlő. Azonban az (AB) táblához illeszthetünk más modellt is.

A legegyszerűbb modell a *minimális modell*:

$$\ln F_{ij} = u, \tag{5.71}$$

amely azt fejezi ki, hogy a tábla minden eleme egyenlő az átlagos hatással. Ennél a modellnél az a hipotézisünk, hogy $p_{ij} = 1/IJ$ és $F_{ij} = n/IJ$.

A varianciaelemzéshez hasonlóan bevonhatjuk az (5.69) modellbe a két változó interakcióját (kölcsönhatását) is:

$$\ln F_{ij} = u + u_{A(i)} + u_{B(j)} + u_{AB(ij)}, \tag{5.72}$$

ahol $u_{AB(ij)}$ reziduális tagként fejezhető ki:

$$u_{AB(ij)} = \ln F_{ij} - u - u_{A(i)} - u_{B(j)}. \tag{5.73}$$

vagy másképpen

$$u_{AB(ij)} = u - \frac{1}{J}\sum_j \ln F_{ij} - \frac{1}{I}\ln\sum_i F_{ij} + \ln F_{ij} \tag{5.74}$$

Az (5.72) modell a *telített modell*, mivel minden hatást tartalmaz, ezért tökéletesen illeszkedik az adatokhoz. A telített modell azt a hipotézist fejezi ki, hogy

$$p_{ij} = f_{ij}/n, \quad \text{azaz} \quad F_{ij} = f_{ij}.$$

A telített modellben a kölcsönhatások sorösszege, illetve oszlopösszege a szóráselemzéshez hasonlóan ki kell, hogy elégítse a

$$\sum_i u_{AB(ij)} = \sum_j u_{AB(ij)} = 0 \tag{5.75}$$

ANOVA-feltételeket.

Összefoglalva, a kétdimenziós kereszttáblához illesztett telített loglineáris modell az alábbi hatásokat tartalmazza:

$$\begin{aligned} y_{ij} = \ln F_{ij} = {} & u && \text{zérórendű hatás,} \\ & + u_{A(i)} + u_{B(j)} && \text{elsőrendű hatások,} \\ & + u_{AB(ij)} && \text{másodrendű hatások.} \end{aligned}$$

Ez az additív modell vektor- és mátrixjelölésekkel is felírható:

$$\mathbf{y} = \mathbf{X} \cdot \mathbf{u}, \tag{5.76}$$

ahol $\mathbf{u}$: a hatásokat leíró paraméterek P elemű vektora, ahol $P \leq (1 + I + J + I \cdot J)$;
$\mathbf{X}$: az együtthatómátrix ($C \times P$ elemű), ahol $C = I \cdot J$, és

$$x_{ij} = \begin{cases} 1, & \text{ha az } i\text{-edik cella becslésénél a modellben szerepel a } j\text{-edik paraméter} \\ 0, & \text{különben.} \end{cases}$$

$\mathbf{y}$: a várt gyakoriságok logaritmusai (sorfolytonosan elhelyezve a C elemű vektorban).
A paraméterek becslése a legkisebb négyzetek módszere értelmében:

$$\widehat{\mathbf{u}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{5.77}$$

Azonban az $\mathbf{y} = \mathbf{X}\,\mathbf{u}$ egyenlet paramétereit így nem tudjuk becsülni, mivel az $\mathbf{X}'\mathbf{X}$ mátrix szinguláris. A modell ebben a formájában túlparametrizált. Ahhoz, hogy a paraméterek becsléseit megkaphassuk, figyelembe kell venni az (5.70) és (5.75) szerinti ANOVA-feltételeket.

A 2×2-es táblára a telített modell mátrix alakban a következő:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} \ln F_{11} \\ \ln F_{12} \\ \ln F_{21} \\ \ln F_{22} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ u_{A(1)} \\ u_{A(2)} \\ u_{B(1)} \\ u_{B(2)} \\ u_{AB(11)} \\ u_{AB(12)} \\ u_{AB(21)} \\ u_{AB(22)} \end{bmatrix}.$$

Az (5.70) és (5.75) feltételek beépítése után a 2×2-es tábla a következőképpen írható:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} \ln F_{11} \\ \ln F_{12} \\ \ln F_{21} \\ \ln F_{22} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix}, \quad \text{azaz} \quad \mathbf{y} = \mathbf{X}_r\mathbf{w}$$

ahol

$$w_1 = u,$$
$$w_2 = u_{A(1)} = -u_{A(2)},$$
$$w_3 = u_{B(1)} = -u_{B(2)},$$
$$w_4 = u_{AB(11)} = -u_{AB(12)} = -u_{AB(21)} = u_{AB(22)}.$$

Az $\mathbf{X}_r$ redukált együtthatómátrix most ortogonális (bármely két oszlopvektorának a szorzata nullát ad eredményül), így a w paraméterek becsléseiből kiszámíthatjuk az u hatások becsléseit.

Az $\mathbf{X}_r'\mathbf{X}_r = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}$ inverz mátrixa

$$(\mathbf{X}_r'\mathbf{X}_r)^{-1} = \begin{bmatrix} 1/4 & 0 & 0 & 0 \\ 0 & 1/4 & 0 & 0 \\ 0 & 0 & 1/4 & 0 \\ 0 & 0 & 0 & 1/4 \end{bmatrix} \text{ és } (\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{X}_r' = 1/4 \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$$

felhasználásával (5.77) alapján végezzük a becslést:

$$\widehat{w}_1 = \widehat{u} = \frac{1}{4}[\ln F_{11} + \ln F_{12} + \ln F_{21} + \ln F_{22}] = \frac{1}{4}\ln[F_{11} \cdot F_{12} \cdot F_{21} \cdot F_{22}]$$

$$\widehat{w}_2 = \widehat{u}_{A(1)} = -\widehat{u}_{A(2)} = \frac{1}{4}[\ln F_{11} + \ln F_{12} - \ln F_{21} - \ln F_{22}] = \frac{1}{4}\ln\left[\frac{F_{11} \cdot F_{12}}{F_{21} \cdot F_{22}}\right]$$

$$\widehat{w}_3 = \widehat{u}_{B(1)} = -\widehat{u}_{B(2)} = \frac{1}{4}[\ln F_{11} - \ln F_{12} + \ln F_{21} - \ln F_{22}] = \frac{1}{4}\ln\left[\frac{F_{11} \cdot F_{21}}{F_{12} \cdot F_{22}}\right]$$

$$\widehat{w}_4 = \widehat{u}_{AB(11)} = -\widehat{u}_{AB(12)} = -\widehat{u}_{AB(21)} = \widehat{u}_{AB(22)} =$$

$$= \frac{1}{4}[\ln F_{11} - \ln F_{12} - \ln F_{21} + \ln F_{22}] = \frac{1}{4}\ln\left[\frac{F_{11} \cdot F_{22}}{F_{12} \cdot F_{21}}\right]$$

Az u_{AB} hatások becslésekor a várható gyakoriságok szerepelnek az esélyhányados (5.26) képletében, így $\widehat{u}_{AB} = \frac{1}{4}\ln\alpha$ írható. Ha $\alpha = 1$, azaz a két változó független, akkor az u_{AB} kölcsönhatások becsült értéke minden cellában zérus.

A hatások értelmezése

A loglineáris modell a gyakoriságok logaritmusait bontja additív tagokra. A becsült hatások értelmezhetők úgy is, ha felírjuk a telített modell multiplikatív megfelelőjét:

$$F_{ij} = t \cdot t_{A(i)} \cdot t_{B(j)} \cdot t_{AB(ij)} \tag{5.78}$$

Az additív modell (5.72) szerinti hatásai a multiplikatív modell paramétereinek logaritmusai, és egymásból kölcsönösen előállíthatók:

$$u = \ln t \quad \text{és} \quad t = e^u,$$

ahol t a várt gyakoriságok geometriai átlaga;

$$u_{A(i)} = \ln t_{A(i)} \quad \text{és} \quad t_{A(i)} = e^{u_{A(i)}},$$

ahol t_A a várt sorgyakoriságok és a várt gyakoriságok geometriai átlagának aránya, az i-edik kategória előfordulásának esélye.

$$u_{B(j)} = \ln t_{B(j)} \quad \text{és} \quad t_{B(j)} = e^{u_{B(j)}},$$

ahol t_B a várt oszlopgyakoriságok és a várt gyakoriságok arányának geometriai átlaga, a j-edik kategória előfordulásának esélye.

$$u_{AB(ij)} = \ln t_{AB(ij)} \quad \text{és} \quad t_{AB(ij)} = e^{u_{AB(ij)}},$$

ahol t_{AB} a várt gyakoriság és az előző három multiplikatív tényező szorzatának aránya.

A multiplikatív modell feltételei:

$$\prod_i t_{A(i)} = \prod_j t_{B(j)} = \prod_i t_{AB(ij)} = \prod_j t_{AB(ij)} = 1. \tag{5.79}$$

A t-hatások (2×2)-es kereszttáblában (5.79) alapján egymás reciprokai, és felírhatók az esélyhányados segítségével is, pl.:

$$t_{AB(11)} = e^{u_{AB(11)}} = (F_{11}F_{22}/F_{12}F_{21})^{1/4} = \alpha^{1/4}. \tag{5.80}$$

A telített modellben annak az esélye, hogy egy megfigyelés az A változó 1. kategóriájába esik inkább, mint a 2. kategóriájába, feltéve hogy a B változó j-edik kategóriájába esik, az (5.78) egyenlet alapján:

$$\frac{F_{1j}}{F_{2j}} = \frac{t_{A(1)}t_{AB(1j)}}{t_{A(2)}t_{AB(2j)}}. \tag{5.81}$$

Az A változó teljes esélye két tényezőtől függ, egyrészt a $t_{A(1)}/t_{A(2)}$ hányadostól, az A változó specifikus hatásától, másrészt a két változó interakciójától, a $t_{AB(1j)}/t_{AB(2j)}$ hányadostól. Ebből látszik, hogy a függetlenséget feltételező modellben a teljes hatás egyenlő a specifikus hatással (mivel $t_{AB(1j)}/t_{AB(2j)} = 1$). A null-modellben (minimális modellben) az $F_{11}/F_{21} = 1$, azaz annak az esélye, hogy egy megfigyelés melyik kategóriába esik, minden kategóriára egyformán 1.

A modell illeszkedése

A loglineáris modellt a következő modellépítési stratégiával alkalmazzuk. Kiindulunk egy egyszerű, kevés hatást feltételező modellből, és megvizsgáljuk, hogy a modell milyen pontosan illeszkedik az adatokhoz.

Az illeszkedés jóságát két statisztikával mérhetjük:

a) Az (5.11) szerinti Pearson-féle χ^2-statisztika

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(f_{ij} - F_{ij})^2}{F_{ij}}. \tag{5.82}$$

b) A likelihood hányados statisztika

$$G^2 = 2 \cdot \sum_{i=1}^{I} \sum_{j=1}^{J} f_{ij} \cdot \ln \frac{f_{ij}}{F_{ij}}. \tag{5.83}$$

Sorszám	Hatások (u-tagok)	Független paraméterek száma	Hipotézis	A modell jellemzője és elnevezése
1.	u	1	$p_{ij} = 1/IJ$ $F_{ij} = n/IJ$	Egyenlő cella-valószínűségek; Minimális null-modell
2.	$u + u_{A(i)}$	$1 + (I - 1) = I$	$p_{ij} = p_{i0}/J$ $F_{ij} = f_{i0}/J$	Konstans oszlop-valószínűségek modellje
3.	$u + u_{B(j)}$	$1 + (J - 1) = J$	$p_{ij} = p_{0j}/I$ $F_{ij} = f_{0j}/I$	Konstans sor-valószínűségek modellje
4.	$u + u_{A(i)} + u_{B(j)}$	$1 + (I - 1) + (J - 1) = I + J - 1$	$p_{ij} = p_{i0}p_{0j}$ $F_{ij} = n\frac{f_{i0}}{n}\frac{f_{0j}}{n}$	Független változókra felírt modell
5.	$u + u_{A(i)} + u_{B(j)} + u_{AB(ij)}$	$1 + (I - 1) + (J - 1) + (I - 1)(J - 1) = IJ$	$p_{ij} = f_{ij}/n$ $F_{ij} = f_{ij}$	Kölcsönhatások is szerepelnek; Telített modell

5.10. táblázat. Loglineáris modellek kétdimenziós kereszttáblákra

Ha a minta elemszáma elég nagy, akkor mindkét statisztika közelítőleg χ^2-eloszlású $S = I \cdot J - P$ szabadságfokkal, ahol P az illesztett paraméterek (hatások) száma. Ha a likelihood hányadost átalakítjuk és a várt gyakoriságok helyébe az illesztett modellt írjuk, akkor a G^2 képletben a hatások összeadódnak, azaz a függvény additív. A kétdimenziós tábla null-modelljének illeszkedését (5.83) alapján mérve és G_0^2-tel jelölve:

$$G^2 = 2 \sum_i \sum_j f_{ij} [\ln f_{ij} - \ln F_{ij}]$$

$$G_0^2 = 2 \sum_i \sum_j f_{ij} [\ln f_{ij} - u],$$

míg a telített modell esetében a képlet további paraméterekkel bővül, és a G_t^2 értéke csökken, ahogy az illeszkedés javul:

$$G_t^2 = 2 \sum_i \sum_j f_{ij} [\ln f_{ij} - (u + u_{A(i)} + u_{B(i)} + u_{AB(ij)})].$$

Ha az illeszkedés nem elég jó, akkor további paramétereket adunk a modellhez addig, amíg kielégítő illeszkedést érünk el. A lehetséges modelleket az 5.10. táblázat foglalja össze.

5.2.2. Loglineáris modell három- és többdimenziós kereszttáblára

Ha A és B mellé bevezetünk egy új változót, C-t[9] (kategóriáinak száma K), akkor háromdimenziós táblába rendezhetjük az együttes gyakoriságokat.

A telített modell

A kétdimenziós gyakoriságtáblához illesztett loglineáris modellhez hasonlóan írható fel a telített modell:

$$\begin{aligned} y_{ijk} = \ln F_{ijk} = & \qquad (5.84) \\ = u + & \quad \text{zérórendű hatás,} \\ + u_{A(i)} + u_{B(j)} + u_{C(k)} + & \quad \text{elsőrendű hatások,} \\ + u_{AB(ij)} + u_{AC(ik)} + u_{BC(jk)} + & \quad \text{másodrendű hatások,} \\ + u_{ABC(ijk)} & \quad \text{harmadrendű hatás.} \end{aligned}$$

A modell feltételei:

$$\sum_{i=1}^{I} u_{A(i)} = \sum_{j=1}^{J} u_{B(j)} = \sum_{k=1}^{K} u_{C(k)} = 0, \qquad (5.85)$$

$$\sum_{i=1}^{I} u_{AB(ij)} = \sum_{j=1}^{J} u_{AB(ij)} = \ldots = \sum_{k=1}^{K} U_{BC(jk)} = 0, \qquad (5.86)$$

$$\sum_{i=1}^{I} u_{ABC(ijk)} = \sum_{j=1}^{J} u_{ABC(ijk)} = \sum_{k=1}^{K} u_{ABC(ijk)} = 0. \qquad (5.87)$$

[9] A cellák számát is C-vel jelöltük az (5.76)-ban. A továbbiakban a szövegben utalunk arra, hogy a C változóról vagy az $I \cdot J \cdot K = C$ cellaszámról van szó.

A zérórendű hatás a várt gyakoriságokból határozható meg:

$$u = \frac{1}{IJK} \sum_i \sum_j \sum_k \ln F_{ijk}. \tag{5.88}$$

Az elsőrendű vagy egyváltozós hatást pl. az A változóra felírva:

$$u_{A(i)} = \frac{1}{JK} \sum_j \sum_k \ln F_{ijk} - u. \tag{5.89}$$

A másodrendű vagy kétváltozós hatás pl. az AB változópárra:

$$u_{AB(ij)} = \frac{1}{K} \sum_k \ln F_{ijk} - \frac{1}{JK} \sum_j \sum_k \ln F_{ijk} - \frac{1}{IK} \sum_i \sum_k \ln F_{ijk} - u. \tag{5.90}$$

A harmadrendű vagy háromváltozós hatás:

$$u_{ABC(ijk)} = \ln F_{ijk} - \frac{1}{K} \sum_k \ln F_{ijk} - \frac{1}{J} \sum_j \ln F_{ijk} - \frac{1}{I} \sum_i \ln F_{ijk} + \tag{5.91}$$

$$+ \frac{1}{JK} \sum_j \sum_k \ln F_{ijk} + \frac{1}{IK} \sum_i \sum_k \ln F_{ijk} + \frac{1}{IJ} \sum_i \sum_j \ln F_{ijk} - u.$$

Telített modell esetén a várható gyakoriságok háromdimenziós táblázatban is egyenlők a megfigyelt gyakoriságokkal.

A függetlenség hipotézisét feltételezve három dimenzióban több modell írható fel.

a) *A feltételes függetlenség modellje*

Tételezzük fel, hogy a B és C változók függetlenek az adott A változó esetén, vagyis B és C változók feltételesen függetlenek. Ekkor a háromdimenziós gyakoriságtábla celláinak valószínűsége (5.56) alapján:

$$p_{ijk} = \frac{P(A = i \cap B = j)P(C = k \cap A = i)}{P(A = i)} = \frac{f_{ij0}}{n} \frac{f_{i0k}}{n} \frac{n}{f_{i00}}. \tag{5.92}$$

A várható gyakoriságok a feltételes függetlenség hipotézisével:

$$F_{ijk} = f_{ij0} f_{i0k} / f_{i00}. \tag{5.93}$$

A feltételes függetlenség hipotézisével felírt loglineáris modell:

$$\ln F_{ijk} = u + u_{A(i)} + u_{B(j)} + u_{C(k)} + u_{AB(ij)} + u_{AC(ik)}. \tag{5.94}$$

Ebben a modellben feltételezzük, hogy $u_{BC(jk)} = 0$ és $u_{ABC(ijk)} = 0$.

A feltételes függetlenség hipotézisének másik két lehetősége:

– A és C független adott B esetén ($u_{AC} = 0$ és $u_{ABC} = 0$).

– A és B független adott C esetén ($u_{AB} = 0$ és $u_{ABC} = 0$).

A feltételes függetlenség hipotézise megegyezik a zéró parciális korreláció fogalmával.

b) *A kölcsönös függetlenség modellje*

A kölcsönös függetlenség hipotézise (5.55) alapján:

$$p_{ijk} = P(A = i)P(B = j)P(C = k) = \frac{f_{i00}}{n} \frac{f_{0j0}}{n} \frac{f_{00k}}{n}, \tag{5.95}$$

és a várható gyakoriság:

$$F_{ijk} = \frac{f_{i00} f_{0j0} f_{00k}}{n^2}. \tag{5.96}$$

A kölcsönös függetlenséget feltételező loglineáris modell:

$$\ln F_{ijk} = u + u_{A(i)} + u_{B(j)} + u_{C(k)}. \tag{5.97}$$

A kölcsönös függetlenség modelljéből a másod- és harmadrendű hatásokat kizártuk.

c) *A többszörös függetlenség modellje*

Ha A és B kapcsolódása a C változó minden kategóriájában azonos, akkor A és B független C-től. A többszörös függetlenség hipotézise (5.57) szerint

$$p_{ijk} = P(A = i \cap B = j)P(C = k) = \frac{f_{ij0}}{n} \frac{f_{00k}}{n}. \tag{5.98}$$

A többszörös függetlenség loglineáris modellje:

$$\ln F_{ijk} = u + u_{A(i)} + u_{B(j)} + u_{C(k)} + u_{AB(ij)}, \tag{5.99}$$

ahol feltételezzük, hogy $u_{AC} = 0$, $u_{BC} = 0$ és $u_{ABC} = 0$.

A többszörös függetlenség hipotézise megegyezik háromváltozós normális eloszlású sokaságban a zéró többszörös korrelációval.

A háromváltozós modellben a többszörös függetlenség másik két lehetősége:

– a (BC) változó független A-tól ($u_{AB} = u_{AC} = u_{ABC} = 0$),

– az (AC) változó független B-től ($u_{AB} = u_{BC} = u_{ABC} = 0$).

d) *A páronkénti asszociáció modellje*

A páronkénti asszociáció hipotézise szerint minden változópár összefügg, de a harmadik változó ezt az összefüggést nem befolyásolja. A telített loglineáris modellből ki kell hagynunk a háromváltozós hatást ($u_{ABC} = 0$). Ha behelyettesítjük a modellbe az u-tagokat, azt kapjuk, hogy:

$$\ln F_{ijk} = u + u_{A(i)} + u_{B(j)} + u_{C(k)} + u_{AB(ij)} + u_{AC(ik)} + u_{BC(jk)}. \tag{5.100}$$

A páronkénti asszociáció modellje abban is különbözik az előző modellektől, hogy a páronkénti asszociáció hipotézisében a p_{ijk} nem fejezhető ki csupán a megfigyelt peremeloszlásokkal, de függvénye azoknak.

A hierarchikus modellek

Az eddig tárgyalt loglineáris modellekben a várható értékeket a változók modellben szereplő legmagasabb rendű hatásainak megfelelő peremeloszlások függvényével fejeztük ki. A függetlenség modelljében pl. az f_{i00}, f_{0j0}, és f_{00k} marginálisok az $u_{A(i)}$, $u_{B(j)}$ és $u_{C(k)}$ paraméterekkel függnek össze. Birch (1963) mutatta meg, hogy a megfigyelt és a várható gyakoriságok megegyeznek akkor, ha a megfelelő u-t tartalmazza a modell. Így pl. ha egy háromdimenziós táblában a nem telített modell tartalmazza az u_{AB} paramétert, akkor $\sum_k F_{ijk} = \sum_k f_{ijk} \quad \forall\, i, j$-re. (Ha a modellben nem szerepel az u_{AB}, akkor nem szükségszerűen teljesül az egyenlőség.) Ha az A változó kategóriái szerint is szummázunk, akkor

$$F_{0j0} = f_{0j0} \quad \forall\, j\text{-re}.$$

Ezért, ha egy modell tartalmazza az u_{AB}-t, akkor tartalmazza az u_A-t és u_B-t is. Vagyis nem illesztjük az $\ln F_{ij} = u + u_{AB(ij)}$ modellt anélkül, hogy ne illesztenénk a másik két paramétert is. Így ha az (ABC) táblára egy $A + (BC)$ modellt illesztünk, ez azt jelenti, hogy a modell az u, u_A, u_B, u_C, u_{BC} paramétereket tartalmazza. De ha csak a (BC) modellt illesztjük, akkor a modell paraméterei u, u_B, u_C, u_{BC}. A fenti sajátosság a loglineáris modellek jellegzetessége: ha a modell egy változó magasabb rendű hatását tartalmazza, akkor tartalmazza annak alacsonyabb rendű hatásait is. Az ilyen modelleket *hierarchikus modelleknek* nevezzük. A háromdimenziós táblánál ebből következően a telített modell az egyetlen olyan hierarchikus modell, amelyik tartalmazza az $u_{ABC} - t$. Az a modell, amelyikből kizártuk az interakciókat, szintén hierarchikus modell. De az a modell, amelyik csak az u_A, u_B és u_{BC} paramétereket tartalmazza, nem hierarchikus modell, mivel nem szerepel benne az u_C.

A null-modell csak az átlagos hatást tartalmazza (u paraméter). Ha az eloszlás szorzatpolinomiális, akkor az u paraméteren kívül a rögzített peremeloszlásoknak megfelelő u paramétereket is tartalmazó modellt nevezzük minimális modellnek.

A háromdimenziós kereszttáblára illeszthető modelleket az 5.11. táblázat foglalja össze.

A modell specifikációja	A modell paraméterei							
	u	u_A	u_B	u_C	u_{AB}	u_{AC}	u_{BC}	u_{ABC}
(ABC)	×	×	×	×	×	×	×	×
$(AB) + (BC) + (AC)$	×	×	×	×	×	×	×	
$(AB) + (BC)$	×	×	×	×	×		×	
$(AB) + (AC)$	×	×	×	×	×	×		
$(BC) + (AC)$	×	×	×	×		×	×	
$(AB) + C$	×	×	×	×	×			
$(AC) + B$	×	×	×	×		×		
$(BC) + A$	×	×	×	×			×	
(AB)	×	×	×		×			
(BC)	×		×	×			×	
(AC)	×	×		×		×		
$A + B + C$	×	×	×	×				
$A + B$	×	×	×					
$A + C$	×	×		×				
$B + C$	×		×	×				
A	×	×						
B	×		×					
C	×			×				
u	×							

5.11. táblázat. A háromdimenziós kereszttáblára illeszthető modellek

Loglineáris modell felírható háromnál több változóra is.

Legyen Z számú változónk, és a változók kategóriáinak a számát jelöljük I, J, K, ..., V-vel. A telített loglineáris modell:

$$\begin{aligned}
\ln F_{ij\ldots v} = u+ \quad & \text{zérórendű hatás,}\\
+u_{A(i)} + u_{B(j)} + \ldots \quad & \text{elsőrendű hatások,}\\
+u_{AB(ij)} + u_{AC(ik)} + \ldots + \quad & \text{másodrendű hatások,}\\
+u_{ABC(ijk)} + \ldots + \quad & \text{harmadrendű hatás,}\\
\vdots + \quad & \\
+u_{ABC\ldots Z(ijk\ldots v)} \quad & Z\text{-edrendű hatás.}
\end{aligned}$$

A modell illesztésénél arra törekszünk, hogy minél kevesebb számú paraméterrel becsüljük a tábla gyakoriságait. A modellek illeszkedésének jóságát magasabb dimenzióban is a χ^2-statisztika (5.82) vagy a G^2 likelihood hányados (5.83) méri.

A szabadságfok a cellák száma és a független paraméterek száma[10] közötti különbséggel egyenlő.

A vizsgálandó modellek száma a tábla méretének növelésével nagyon megnő, pl. a négydimenziós táblánál a hierarchikus modellek száma már 166. Nagyobb méretű tábláknál ezért különösen hasznos a lépésenkénti szelekciós eljárás.

A modellek összehasonlítása

A hierarchikus modellek esetén tapasztalt additív tulajdonság miatt a likelihood hányados statisztika felhasználható a loglineáris modellek összehasonlítására is.

Jelölje $F_c^{(a)}$ az (a) modellben a c-edik cella gyakoriságát, ahol $c = 1,2,\ldots C$, és C most az összes cellák száma. $F_c^{(b)}$ a (b) modell azonos cellában várható gyakorisága. Tételezzük fel, hogy az (a) modell paraméterei a (b) modell paramétereinek részhalmazát alkotják.

Az additív tulajdonság miatt a likelihood hányados lehetőséget ad arra, hogy statisztikusan összehasonlítsunk két modellt, és eldöntsük, hogy egy vagy több paraméter hozzáadása a (b) modellhez szignifikánsan csökkentette-e a G^2 értékét.

$$\begin{aligned}
G^2_{(a)} - G^2_{(b)} &= 2 \cdot \sum_{c=1}^{C} f_c \cdot \ln \frac{f_c}{F_c^{(a)}} - 2 \sum_{c=1}^{C} f_c \cdot \ln \frac{f_c}{F_c^{(b)}}\\
&= 2 \sum_{c=1}^{C} f_c \cdot \ln \frac{F_c^{(b)}}{F_c^{(a)}} \qquad (5.101)\\
&= G^2_{(a-b)}. \qquad (5.102)
\end{aligned}$$

A $G^2_{(a-b)}$ feltételes statisztika is közelítőleg χ^2-eloszlású $(S_b - S_a)$ szabadságfokkal.

A G^2-et értelmezhetjük úgy is, mint a gyakoriságok logaritmusai varianciájának azt a részét, amit a modell nem magyaráz. Így $G^2_{(a)}$ az (a) modell által, $G^2_{(b)}$ a (b) modell által nem magyarázott varianciát méri, $G^2_{(a-b)}$ pedig a (b) modellbe bevont plusz paraméterek által megmagyarázott variancia nagyságát fejezi ki.

[10] Az 5.10. táblázat alapján határozható meg.

Ha a $G^2_{(a)} - G^2_{(b)}$ különbséget az egyszerűbb (a hierarchiában alacsonyabb szinten levő) modell likelihood hányadosával elosztjuk, akkor az így kapott hányados a nem magyarázott variancia relatív csökkenését[11] méri:

$$PRE = \frac{G^2_{(a)} - G^2_{(b)}}{G^2_{(a)}}. \tag{5.103}$$

Ha az (a) a null-modell, akkor a mutatót a többváltozós regressziós modellben számított többszörös determinációs együttható mintájára értelmezzük:

$$R^2 = \frac{G^2_{(0)} - G^2_{(b)}}{G^2_{(0)}} = 1 - \frac{G^2_{(b)}}{G^2_{(0)}}. \tag{5.104}$$

E mutató korrigált változatának számításakor figyelembe vesszük a cellák számát is, és így a különböző méretű kereszttáblákra illesztett modellek is összehasonlíthatóak:

$$R^2_{\text{adj}} = 1 - \frac{G^2_{(b)}/(C_b - S_b)}{G^2_{(a)}/(C_a - S_a)}, \tag{5.105}$$

ahol C az adott tábla összes celláinak száma, S pedig a megfelelő modell szabadságfoka.

A modellek összehasonlításának alternatív tesztje az Akaike által javasolt információs kritérium (IK), amelyet minden modellre külön-külön számítunk ki, és a minimális IK értékű modellt választjuk:

$$IK = G^2 - (C - 2S). \tag{5.106}$$

Az (5.106) szerint G^2 csak akkor csökken, ha az adott méretű táblában a hatások száma kisebb, mint $C/2$. A levonással kompenzáljuk a nagy táblákon sok paraméteres modellel elérhető túl jó illeszkedést.

Lépésenkénti modellszelekció

A kereszttábla dimenziószámának növekedésével rohamosan nő a táblára illeszthető modellek száma. Ezért négy vagy annál magasabb dimenzió esetén nagyon idő- és munkaigényes lenne a null-modell és a telített modell közti összes változat becslése és kiértékelése. E probléma hatékony kezelését segíti a lépésenkénti eljárás, amely kétféleképpen valósítható meg.

Backward–eljárást követünk akkor, ha a telített modellből lépésenként egy-egy tagot elhagyunk.

Forward–szelekciót alkalmazunk akkor, ha a legegyszerűbb modellből kiindulva lépésenként egy-egy tagot hozzáveszünk a modellhez.

A kezdő modell megfelelő kiválasztása jelentősen csökkentheti a lépésszámot, ezért a következő eljárás ajánlható. Előállítjuk pl. a négydimenziós táblában

– az összes elsőrendű hatást tartalmazó modellt: A, B, C, D

– az összes kétváltozós hatást tartalmazó modellt: AB, AC, AD, BC, BD, CD

– az összes háromváltozós hatást tartalmazó modellt: ABC, ABD, ACD, BCD

– a negyedrendű hatást tartalmazó modellt: $ABCD$.

A legegyszerűbb – még illeszkedő – modell rendje adja a rend felső korlátját (q_1), és a legmagasabb rendű – már nem illeszkedő – modell adja a rend alsó korlátját (q_2). [A két rend közötti különbség általában egy vagy kettő.]

[11] Angol megfelelőjének rövidítése: Proportional Reduction of Errors=PRE.

Forward–szelekció esetén a q_2-rendű (nem illeszkedő) modellt bővítjük lépésről lépésre úgy, hogy a likelihood arányváltozása maximális legyen. Addig folytatjuk az eljárást, amíg G^2 szignifikánsan nő.

Backward–eljárást követve a q_1-rendű modellből hagyjuk el azokat a tagokat, amelyek G^2-ben a legkisebb csökkenést okozzák. A határokat addig redukáljuk, amíg még találunk nem szignifikáns paramétert.

A modellek közti választást segítik a marginális és a parciális asszociáció tesztjei, melyek a hatások jelentőségét, relatív nagyságát értékelik.

– A k db változó közötti *marginális asszociáció* tesztje a k dimenziós résztáblában azt a hipotézist teszteli, hogy a k változó interakciója nulla-e.[12] Az eljárás során két olyan modell illeszkedésének (G^2) különbségét hasonlítjuk össze, ahol az egyik modell tartalmazza a k változó interakcióját, a másik pedig legfeljebb $(k-1)$-ed rendű hatásokat illeszt. Ha négydimenziós táblánk van, akkor pl. az (ABC) marginális asszociációt vizsgáljuk. Ha az (AB), (AC) és (BC) hatások jól illeszkedő modellt adnak, akkor az (ABC) marginális asszociáció nem jelentős.

– A *parciális asszociáció* tesztje azt vizsgálja, hogy az összes k-adrendű hatás közül az egyik kiválasztott paraméter szignifikáns-e. Négyváltozós tábla másodrendű hatásai közül tesztelhetjük pl. az (AD) parciális asszociációját úgy, hogy mind a hat kétváltozós tagot tartalmazó modell likelihood hányadosát kivonjuk az (AD)-t nem tartalmazó modell G^2 értékéből. A két G^2 különbsége a parciális asszociáció tesztje. A két teszt felhasználásával a modell paramétereit három részhalmazra bonthatjuk:

1. fontosak, nem hagyhatók el,
2. elhagyhatók,
3. további vizsgálat szükséges.

Ha egy paraméterre mindkét teszt nagy (szignifikáns), akkor ez a hatás nem hagyható el, fontos (1. eset). Ha mindkét teszt értéke kicsi, akkor a kérdéses paraméter kihagyható a modellből (2. eset). Ha a tesztek nem egyeznek, további vizsgálat szükséges ahhoz, hogy eldönthessük a paraméter szerepeltetését.

A loglineáris modell paramétereinek becslése

A paraméterek becslése

A loglineáris modellt az általános lineáris modell formájában írjuk fel.

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{e}, \tag{5.107}$$

ahol: $\mathbf{y}$: a gyakoriságok logaritmusait tartalmazza,
$\mathbf{w}$: az $\mathbf{u}$ paraméterek a kiegészítő feltételekkel,
$\mathbf{X}$: az együtthatómátrix a modellben szereplő változók specifikálására,
$\mathbf{e}$: hibavektor.

Az általános lineáris modell paramétereinek becslésére a klasszikus legkisebb négyzetek módszere nem alkalmazható, mert ez a módszer feltételezi, hogy a függő (y) változó varianciája konstans. A loglineáris modellben az y varianciája változik, ha a cellák gyakoriságai változnak, így a variancia függ a modell paramétereitől.

[12] Ez egyenértékű azzal, hogy a $(k+1)$-edik változó elhagyható. Itt a Simpson-féle paradoxon felmerülésének lehetőségét is figyelembe kell venni.

Az egyik lehetséges megoldás, ha *az általánosított legkisebb négyzetek módszerét* alkalmazzuk, amelynek alapján a paraméterek becslése $\hat{\mathbf{w}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$, ahol $\mathbf{V}$ a hibavektor varianciamátrixa. A paraméterek variancia-kovarianciamátrixa var $(\hat{\mathbf{w}}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$.

Goodman (1972) és sokan mások *a maximum likelihood becslést* tartják előnyösebbnek, mivel a paraméterek becsléseire kisebb varianciát ad, mint a többi becslő eljárás.

Birch (1963) bizonyította, hogy a loglineáris modell paramétereinek maximum likelihood becslése a modellben szereplő legmagasabb rendű hatásoknak megfelelő peremeloszlások függvénye, és a becslés megegyezik a minta különböző valószínűségeloszlásai esetében. Többféle maximum likelihood becslési eljárás létezik, ezek közül kettőre térünk ki röviden, mert ezek szerepelnek a két legelterjedtebb számítógépes programban, a GLIM-ben és az ECTA-ban.

a) *Az iteratív súlyozott legkisebb négyzetek eljárás*

Nelder és Wederburn (1972) olyan eljárást közölt a maximum likelihood becslésre, amely a lineáris modellek széles körében alkalmazható. Eljárásuk ekvivalens a súlyozott legkisebb négyzetek módszerével, ha a függő változót a következőképpen módosítjuk:

$$y'_c = y_c + \frac{f_c - u_c}{u_c}, \quad c = 1,2, \ldots, C \tag{5.108}$$

ahol f_c a megfigyelt gyakoriság,
a súly $w = u_c = e^{y_c}$.

A becslést iteratív eljárással végzik. A részletes leírás Neldertől (1974) származik. Ezt a becslési eljárást alkalmazták a GLIM (General Linear Interactive Modelling) programban. Részletesebben az 5.2.5. alfejezetben tárgyaljuk.

b) *Iteratív skálázó eljárás*

Haberman (1972) dolgozta ki az iteratív skálázó eljárást, ami az ECTA (Everyman's Contingency Table Analysis) programban szerepel. Az iteratív eljárást a páronkénti asszociáció modelljére mutatjuk be.

A loglineáris modell, amit becsülni akarunk (5.100) szerint:

$$y_{ijk} = \ln F_{ijk} = u + u_{A(i)} + u_{B(j)} + u_{C(k)} + u_{AB(ij)} + u_{AC(ik)} + u_{BC(jk)}.$$

A modell a legmagasabb rendű marginálisokhoz $[(AB), (AC), (BC)]$ pontosan illeszkedik, így az y_{ijk} maximum likelihood becslése ki kell hogy elégítse a következő három feltételt:

- $\hat{y}_{ij0} = f_{ij0}$,
- $\hat{y}_{i0k} = f_{i0k}$,
- $\hat{y}_{0jk} = f_{0jk}$.

Az iterációt az $\hat{y}^0_{ijk} = 1$ értékkel indítjuk. Az iteráció egy-egy ciklusában a három peremeloszláshoz illesztjük a gyakoriságokat.

- $\hat{y}^{(1)}_{ijk} = \hat{y}^{(0)}_{ijk} \frac{f_{ij0}}{\hat{y}^{(0)}_{ij0}}$ – ahol az $\hat{y}^{(1)}_{ij0}$ az (AB) peremeloszlásával egyenlő $(\hat{y}^{(1)}_{ij0} = f_{ij0})$;
- $\hat{y}^{(2)}_{ijk} = \hat{y}^{(1)}_{ijk} \frac{f_{i0k}}{\hat{y}^{(1)}_{i0k}}$ – ami az (AC) marginálisokhoz illeszkedik $(\hat{y}^{(2)}_{i0k} = f_{i0k})$;

– $\widehat{y}^{(3)}_{ijk} = \widehat{y}^{(2)}_{ijk} \frac{f_{0jk}}{\widehat{y}^{(2)}_{0jk}}$ – ami a (BC) peremeloszláshoz illeszkedik, és kielégíti a harmadik feltételt ($\widehat{y}^{(3)}_{0jk} = f_{0jk}$).

A második lépésben az (AC) peremeloszláshoz igazodva elrontjuk az első lépés illeszkedését, ugyanígy a következő lépésben az előző feltételek teljesítését rontjuk el. Ezért az eljárást egészen addig folytatjuk, amíg minden cellára a becsült gyakoriságok egymást követő különbsége kisebb lesz, mint egy előre adott kicsi érték, pl. 0,01. Fienberg (1980) bizonyította be, hogy ez az eljárás mindig konvergens, és a konvergencia sebessége nagyon gyors.

A becslési eljárás során a nulla gyakoriságú cellák problémát okoznak, ezekre *nem lehet becslést* adni. Megkülönböztetünk azonban kétféle nulla gyakoriságú cellát. Az egyiket mintabeli nullának, a másikat rögzített vagy strukturális nullának nevezzük. A mintabeli nulla[13] azt jelenti, hogy a változók kategóriáinak adott kombinációjára azért nem találtunk a mintában megfigyeléseket, mert a minta nagysága nem volt elég nagy. A rögzített nullák azt az esetet jelölik, amikor az adott változókombináció egyáltalán nem fordul elő a vizsgált populációban. A többdimenziós táblánk a rögzített nullák miatt hiányos lesz, és ilyenkor a becsléseket is rögzítenünk kell nullához.

5.2.3. A logit-modell

A többdimenziós kereszttáblában a gyakorlatban előfordul, hogy különbséget teszünk a változók között függőségi viszony szerint. Így azt vizsgáljuk, hogy a függőnek tekintett változó kategóriái gyakoriságainak az arányai hogyan függnek a táblázat többi változójától (a magyarázó változóktól). Különösen akkor gyakori ez a kérdésfeltevés, amikor a függő változónak csak két kategóriája van (dichotom). Az egyik kategória valószínűsége p, a másiké pedig $(1 - p)$. Az ilyen táblákat vizsgálhatjuk a varianciaelemzés módszerével vagy többváltozós regresszióval, ahol a függő változó bináris változó, a magyarázó változók pedig dummy változók (a magyarázó változók kategóriáival megegyező bináris változók). A modell paramétereinek a becslésénél a klasszikus legkisebb négyzetek elvét használjuk, ahol feltételezzük, hogy a függő változó varianciája konstans. A kontingenciatáblánál a függő változó egy arányszám (p), amelynek a varianciája $p(1 - p)$. Ezért a variancia- és regresszióelemzés helyett az ilyen táblák elemzésére Cox (1970) és Theil (1970) javasolta a p logit transzformációját, azaz a függő változóra az $\ln[p/(1 - p)]$ logit függvény bevezetését. Szokásos a

$$Z = \ln[p/(1 - p)] \tag{5.109}$$

és az

$$e^Z = \frac{p}{1 - p} \tag{5.110}$$

jelölés is. A logit függvény az esélyek logaritmusa, és az erre felírt modell a loglineáris modell speciális esete.

[13] Goodman (1975) javasolta, hogy adjunk hozzá minden cella gyakoriságához 1/2-t, így a becsléseket már ki tudjuk számítani.

Ha a Z értékét ismerjük vagy becsültük,[14] akkor az (5.110)-ből a p valószínűséget kifejezhetjük:

$$p = \frac{e^Z}{1 + e^Z}. \tag{5.111}$$

Ha az A változó két kategóriájához (5.109) szerint illesztett logit értékeket kivonjuk egymásból, akkor az esélyek hányadosának logaritmusát kapjuk:

$$Z_1 - Z_2 = \ln \frac{p_1}{1 - p_1} - \ln \frac{p_2}{1 - p_2} = \ln \frac{p_1(1 - p_2)}{p_2(1 - p_1)}. \tag{5.112}$$

Az esélyek hányadosát (5.110) alapján közvetlenül felírhatjuk

$$e^{Z_1 - Z_2} = \frac{p_1(1 - p_2)}{p_2(1 - p_1)}. \tag{5.113}$$

Így a logit-modellel az esélyek hányadosait elemezhetjük, míg a loglineáris modellel az esélyeket.

A hatások értelmezése

Tekintsük a háromdimenziós kontingenciatáblát. Legyen az A a függő változó, és jelölje f_{1jk} és f_{2jk} az A változó két kategóriájának megfigyelt gyakoriságait a másik két (B és C) magyarázó változó (jk) cellájában. Ekkor $p_{jk} = f_{1jk}/(f_{1jk} + f_{2jk})$.

A logit-modell:

$$Z_{jk} = \ln \frac{p_{jk}}{1 - p_{jk}} = \ln \frac{f_{1jk}}{f_{2jk}} = \ln f_{1jk} - \ln f_{2jk}. \tag{5.114}$$

Ha külön-külön illesztjük a loglineáris modellt az f_{1jk} és f_{2jk} gyakoriságok logaritmusaihoz, és behelyettesítjük az (5.114) logit-modellbe, azt kapjuk, hogy[15]

$$\begin{aligned} \ln f_{1jk} - \ln f_{2jk} = (u_{A(1)} - u_{A(2)}) + (u_{AB(1j)} - u_{AB(2j)}) + \\ + (u_{AC(jk)} - u_{AC(2k)}) + (u_{ABC(1jk)} - u_{ABC(2jk)}). \end{aligned} \tag{5.115}$$

Felhasználva a loglineáris modell (5.85)–(5.87) ANOVA-típusú feltételeit, azt kapjuk, hogy

$$\begin{aligned} u_{A(1)} - u_{A(2)} &= 2u_{A(1)} = w, \\ u_{AB(1j)} - u_{AB(2j)} &= 2u_{AB(1j)} = w_{B(j)}, \\ u_{AC(1k)} - u_{AC(2k)} &= 2u_{AC(1k)} = w_{C(k)}, \\ u_{ABC(1jk)} - u_{ABC(2jk)} &= 2u_{ABC(1jk)} = w_{BC(jk)}, \end{aligned} \tag{5.116}$$

azaz a logit-modell felírható két loglineáris modell különbségeként:

$$\ln f_{1jk} - \ln f_{2jk} = w + w_{B(j)} + w_{C(k)} + w_{BC(jk)}. \tag{5.117}$$

Az A esélyhányadosa a másik két változó (jk) cellájában:

$$e^{Z_{jk}} = e^w \cdot e^{w_{B(j)}} \cdot e^{w_{C(k)}} \cdot e^{w_{BC(jk)}}. \tag{5.118}$$

[14] Pl. $Z = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$ modellel, amelyet logisztikus regressziónak nevezünk.

[15] Figyeljük meg, hogy a logit-modellből hiányzik az u_{BC} tag (mivel ez nem tartalmazza az A függő változót.

Ha a C változó két kategóriájához illesztett logit értékek különbségébe helyettesítjük be az (5.115) szerinti telített modellt, akkor a w és $w_{B(j)}$ tagok kiejtik egymást:

$$Z_{j1} - Z_{j2} = \ln \frac{p_{j1}}{1 - p_{j1}} - \ln \frac{p_{j2}}{1 - p_{j2}} = w_{C(1)} - w_{C(2)} + w_{BC(j1)} - w_{BC(j2)} \quad (5.119)$$

és az esélyhányadost két arány határozza meg:

$$e^{Z_{j1} - Z_{j2}} = \frac{e^{w_{C(1)}}}{e^{w_{C(2)}}} \cdot \frac{e^{w_{BC(j1)}}}{e^{w_{BC(j2)}}}. \quad (5.120)$$

Az első tényező azt fejezi ki, hogy a C változó 1. kategóriája az A változó 1. kategóriájával vagy az $A(2)$-vel jár-e inkább együtt, míg a második tényező a B változó adott kategóriájától függő interakciót méri.

5.2.4. A loglineáris modell néhány sajátossága

A reziduálisok elemzése

A modellek illeszkedésének vizsgálatánál a statisztikai gyakorlat a reziduálisokat vizsgálja. Haberman (1973) javasolta, hogy a loglineáris modell illesztésénél a megfigyelt (f_c) és a modellel illesztett (F_c) gyakoriságok különbségének standardizált értékét (e_c) elemezzük:

$$e_c = (f_c - F_c)/\sqrt{F_c}. \quad (5.121)$$

Ha ábrázoljuk ezeket a reziduálisokat, akkor könnyen szembetűnnek a rosszul illeszkedő cellák.

Nelder (1974) egy másik reziduális mértéket definiált az illeszkedés jellemzésére:

$$3[f_c^{2/3} - (F_c - 1/6)^{2/3}]/2F_c^{1/6}. \quad (5.122)$$

Sajnos ezek a reziduálisok nem követnek normális eloszlást (mint a regressziós reziduálisok), így a kiértékelésük nehézkesebb.

A kereszttábla particionálása

A loglineáris modell illesztésének sajátossága, hogy egy u-taggal összefüggő összes paramétert vagy beveszünk a modellbe, vagy kihagyunk a modellből. Így például az u_{ABC} szerepeltetése a modellben összesen $(I - 1)(J - 1)(K - 1)$ paraméter bevételét jelenti, és ha az u_{ABC} a teljes táblára illesztett modellnél szignifikáns, akkor elfogadjuk azt a hipotézist, hogy az $u_{ABC(ijk)} \neq 0$ minden i, j, k indexre. Azonban *a priori* feltételezhetjük, hogy ezek közül az $u_{ABC(ijk)}$ tagok közül

a) vannak, amelyek különböznek nullától, de vannak közöttük nullák, illetve

b) vannak, amelyek egyenlők egymással, és vannak, amelyek nullával egyenlők.

Ezeket a hipotéziseket úgy tudjuk tesztelni, hogy a mintát a hipotéziseknek megfelelően csoportokba soroljuk, és a csoportokra különböző modellt illesztünk.

Felhasználhatjuk az **X** együtthatómátrixot is a particionált loglineáris modell illesztésére. Kullback és Fisher (1973) ad erre példát. Egy másik megközelítés lehet, hogy a modellt a táblának csak bizonyos celláihoz illesztjük, és így a táblát hiányosnak tekintjük. Ezt az eljárást mutatja be Bishop, Fienberg és Holland (1975).

A loglineáris modell problémái

A következőkben sorra veszünk néhány főbb problémát.

– A paraméterek szignifikanciapróbája érzékeny a minta nagyságára. Kis mintánál csak a nagyon erős hatások szignifikánsak, és csak kevés cellából álló táblázatot elemezhetünk. Nagy mintánál viszont már az egészen kicsi hatások is szignifikánsak, és a lehetséges nagy számú cella interpretálása meglehetősen nehézkes.

– Ha a táblához illesztett modellt particionálni akarjuk, akkor nehézkes az együtthatómátrix előállítása.

– Ugyanígy nehézkes a nemhierarchikus modellek illesztése a meglévő számítógépes eljárásokkal.

– Nem megoldott a reziduálisok értelmezése, és a modell validitásának vizsgálata.

– A paraméterek értelmezésében problémát okoz, hogy a loglineáris modell túlparametrizált. Egy túlparametrizált modellben az egyedi paramétereket nem lehet addig becsülni, amíg plusz feltételeket nem adnak a modellhez. A feltételek lehetővé teszik ugyan a paraméterek becslését, de a paramétereket csak a feltételek kontextusában lehet értelmezni. A leggyakoribb feltétel, hogy bármelyik változóra (bármelyik indexre futtatva) a hatások összege nulla legyen. Ez megegyezik a szóráselemző modell feltételeivel, ezért nevezzük ezeket ANOVA-típusú feltételeknek.

Kevésbé elterjedt alternatív feltételezés, hogy az összes paramétert, amelyiknek egy vagy több indexe eggyel egyenlő, tekintsük nullának. Ezt a feltételezést alkalmazzuk akkor, amikor a szóráselemző modellt regresszióval becsüljük úgy, hogy a kategorikus változók helyett dichotom változókat vezetünk be. Ezért ezeket a feltételeket regresszió típusú feltételeknek nevezzük. Láttuk, hogy a logit-modell w paraméterei a loglineáris modell megfelelő u paraméterei kétszeresével egyenlők az (5.116) szóráselemzés típusú feltételek esetén.

5.2.5. Az általánosított lineáris modell (GLIM)

A loglineáris modellel kategorikus változók többdimenziós kereszttábláit vizsgáljuk. Nelder és Wedderburn (1972) kidolgozta az általánosított lineáris modellt és az ezzel összefüggő számítógépes programot (GLIM: Generalized Linear Interactive Modelling Program), amely alkalmas kevert mérési skálájú változók lineáris modellezésére is.

A GLIM-modellnek index*speciális esetei* a következő modellek:

(1) Lineáris regresszió mennyiségi magyarázó változókkal és normális eloszlású hibataggal.

(2) Minőségi változók kereszttáblájához szóráselemző modell illesztése normális eloszlású hibataggal.

(3) Loglineáris modell illesztése kereszttáblához.

(4) Logit-modell illesztése.

(5) Probit elemzés.

(6) Szórásnégyzet komponenseinek becslése, ha a négyzetösszegek függetlenek és gamma-eloszlásúak.

A GLIM a fenti modellek becslésén túl lehetőséget ad arra is, hogy a magyarázó változók kevertek legyenek (kvantitatívak és kvalitatívak), ahol egy kvantitatív változó együtthatója egy vagy több minőségi változó kategóriái szerint változhat.

A GLIM-modell elemei az alábbi vektorok:

y: a függő valószínűségi változó megfigyelési értékei (y_i, $i = 1, \ldots, n$),

$\boldsymbol{\mu}$: várható érték = $E(y)$, a függő változó szisztematikus komponense (μ_i, $1, \ldots, n$),

e: hibatag, a függő változó véletlen komponense ($y_i = \mu_i + e_i$),

$\boldsymbol{\eta}$: lineáris becslés,

$\mathbf{x}_1, \ldots, \mathbf{x}_p$: magyarázó változók (lehetnek dummy változók is) (x_{ij}, $i = 1, \ldots, n$; $j = 1, \ldots, p$).

A modell három feltétele:

a) A függő változó valószínűségeloszlása négyféle lehet (normális, binomiális, Poisson- és gamma-), a várható érték: μ.

b) A modell lineáris egyenlete (lineáris struktúra):

$$\eta_i = \sum_j b_j x_{ij}, \tag{5.123}$$

ahol a b_j paraméterek ismeretlenek, az x_{ij} értékek adottak.

Mátrixjelölésekkel:

$$\boldsymbol{\eta} = \mathbf{X}\,\mathbf{b}, \tag{5.124}$$

ahol $\mathbf{X}$ $n \times p$ típusú mátrix.

Ha x_{ij} két értéke (1, 0) a magyarázó változó kérdéses kategóriájának meglétét vagy hiányát jelenti, akkor a b_j az adott kategória hatását fejezi ki; ha pedig x_{ij} egy mennyiségi változó megfigyelt értéke, akkor a b_j a j-edik változó súlya.

c) A függő változó szisztematikus komponense és a lineáris becslés közötti kapcsolatot leíró $\eta = f(\mu)$ függvény a GLIM-modellben hétféle lehet, amint ezt az 5.12. táblázat[16] mutatja.

A tábla belsejében az (1)–(6) jelölések a GLIM speciális eseteire utalnak.

Függvény	Eloszlás			
	Normális	Poisson-	Binomiális-	Gamma-
Azonosság $\eta = \mu$	(1)(2)			(6)
Logaritmus $\eta = \ln \mu$		(3)		
Inverz $\eta = 1/\mu$				
Négyzetgyök $\eta = \sqrt{\mu}$				
Logit $\eta = \ln \frac{p}{1-p}$			(4)	
Probit $\eta = \Phi(p)$			(5)	
Komplementer log log $\eta = \ln(-\ln(1-p))$				

5.12. táblázat. GLIM-modellek[17]

[16] A táblázat forrása: Nelder, J. A.: Applied Statistics. 24, 259–261, 1975.

[17] A Probit függvényben a Φ a normális eloszlás eloszlásfüggvénye.

Példa kevert modellre

Legyenek A és B minőségi változók, és f_{ij} jelölje a kereszttábla megfelelő gyakoriságait. Az A változó minden kategóriájához tartozzon $x_{1i}, x_{2i}, \ldots$ mennyiségi változó. A kevert loglineáris modellt a következőképpen írjuk fel:

$$\ln F_{ij} = u + u_{A(i)} + u_{B(j)} + u_{AB(ij)} + b_{1i}x_{1i} + b_{2i}x_{2i} + \ldots \tag{5.125}$$

Példaként említhetjük azt a modellt, amikor A változó 2 értéke jelzi, hogy a férjezett asszonynak van-e állása, a B változó az iskolai végzettséget méri, x_1 az életkor, x_2 a gyermekek száma, x_3 a férj jövedelme.

A GLIM-modell paramétereinek becslése

A GLIM-modell ismeretlen $\mathbf{b}$ paramétereit ($\boldsymbol{\eta} = \mathbf{X}\,\mathbf{b}$) *maximum likelihood* eljárással becsüljük.

Tételezzük fel először, hogy az $\mathbf{X}$ mátrix nem szinguláris, és így létezik az inverze. A maximum likelihood módszer iteratív eljárása a b paraméterekre a következő becslést adja (az eljárást részletesen lásd Nelder, J. A. és Wedderburn [1972]):

$$\widehat{\mathbf{b}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{z}, \tag{5.126}$$

ahol $\mathbf{V}$ a hibavektor varianciamátrixa (feltételezzük, hogy diagonális), és

$$z_i = \eta_i + (y_i - \eta_i)d\eta/d\mu. \tag{5.127}$$

A becsült paraméterek varianciamátrixa:

$$\mathrm{var}(\widehat{\mathbf{b}}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}. \tag{5.128}$$

Ha az $\mathbf{X}$ mátrix szinguláris, akkor az $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}) = \mathbf{A}$ mátrix általánosított inverzét számítjuk ($\mathbf{A}$ általánosított inverze $\mathbf{A}^-$, ha $\mathbf{A}^-$ kielégíti az $\mathbf{A}\,\mathbf{A}^-\mathbf{A} = \mathbf{A}$ egyenletet). Az általánosított inverzből több is létezhet, így a $\widehat{\mathbf{b}}$ becslésre nem egyértelmű a megoldás.

Másképpen, ha a modell strukturális egyenlete r független feltételt tartalmaz, és a $\mathbf{b}$ paraméterek tere p-dimenziós ($r < p$), akkor végtelen sok $\widehat{\mathbf{b}}$ becslés kielégíti az egyenletet. Ahhoz, hogy egyértelmű legyen a megoldás (a megoldások halmaza egy pontra redukálódjon), plusz feltételeket kell a modellhez hozzátennünk. Meg kell jegyezni, hogy bármely partikuláris megoldás ugyanazt a becslést adja η-ra és μ-re (csak a $\widehat{\mathbf{b}}$-k lesznek különbözők), így a modell illesztett értékeit a pótlólagos feltételek nem befolyásolják. A loglineáris modellnél a pótlólagos feltételek azt írják elő, hogy a hatások változónkénti összege nulla legyen. A GLIM-modellben a $p - r$ pótlólagos feltételként $p - r$ számú paraméternek 0 értéket adnak, és a többi paraméter értékét ennek függvényében határozzák meg. Az így meghatározott paraméterek varianciamátrixa

$$\mathrm{var}(\widehat{\mathbf{b}}) \simeq \mathbf{A}^-. \tag{5.129}$$

Természetesen a rögzített paramétereknek megfelelő sor-, illetve oszlopvektor $\mathbf{0}$ lesz.

Az illeszkedés jóságának mérése

Az a modell, amelyik tartalmazza az összes lehetséges paramétert (telített modell), tökéletesen illeszkedik a megfigyelt adatokhoz. Az a modell, amelyik csak egy paramétert tartalmaz (null-modell), az esetek többségében nagyon rosszul illeszkedik. Ezen a két extrém eseten belül még megkülönböztetünk két szélső modellt. A minimális modell csak azokat a paramétereket tartalmazza, amelyeket a hipotézis szerint minimálisan tartalmaznia kell (pl. amikor rögzítjük valamelyik változó peremeloszlását). A másik szélső

modell, amelyik a hipotézis szerinti legtöbb paramétert tartalmazza, a maximális modell. A minimális és a maximális modellen belül keressük a lehető legkevesebb paramétert tartalmazó, de még adott szignifikanciaszinten illeszkedő modellt.

Jelölje ℓ_v a vizsgált modell likelihood függvényének értékét, ℓ_t pedig a telített modellét.

A vizsgált modell jóságát a telített modelltől való eltérésével mérjük. A statisztika (scaled variance):

$$S(v,t) = -2\ln(\ell_v/\ell_t)$$

közelítőleg χ^2-eloszlású (szf_t–szf_v) szabadságfokkal (ahol szf_t a független paraméterek száma a t modellnél). Az S statisztika normális eloszlású hibatag és $\eta = \mu$ függvény esetén pontosan χ^2-eloszlást követ, egyébként közelítőleg χ^2-eloszlású.

Tételezzük fel, hogy az (a) modell paraméterei két lineárisan független csoportra bonthatók: $b_1, \ldots, b_t$ és $b_{t+1}, \ldots, b_p$ és a (b) modell csak a $b_{t+1}, \ldots, b_p$ paramétereket tartalmazza.

A $b_1 = b_2 = \ldots = b_t = 0$ hipotézist az $S(b,a)$ statisztikával teszteljük. Mivel $E(\chi_t^2) = t$, praktikus szabály, hogy ha $S(b,a)$ közel esik t-hez, akkor a hipotézist elfogadjuk. Ha a $b_1, \ldots b_t$ paraméterek vagy a $b_{t+1}, \ldots, b_p$ paraméterek nemlineárisan függetlenek, akkor a χ^2 szabadságfoka kevesebb lesz, mint t. Általánosságban, ahogyan azt a loglineáris modellnél is láttuk,

$$S(c,a) - S(b,a) = S(c,b)$$

($\text{szf}_b - \text{szf}_c$) szabadságfokú χ^2-eloszlású valószínűségi változó, ahol (c) kisebb modell, mint (b), és (b) kisebb modell, mint (a).

6. fejezet

Útelemzés

A korreláció – általában – nem bizonyítja az okozati kapcsolatot, csak a kapcsolat szorosságáról tájékoztat, de a kapcsolat irányáról nem.

Két változó korrelációja (zéró rendű) önmagában nem bizonyítja azt sem, hogy egyáltalán van-e a két változó között kapcsolat.

Hamis korrelációnak nevezzük két változó korrelációját, ha az egy harmadik változó hatásaként jött létre. Ilyenkor ha a két változó (1 és 2) kapcsolatából kiszűrjük a harmadik (3) változó hatását, vagyis kiszámítjuk a parciális korrelációs együtthatót ($r_{12,3}$), akkor a parciális korreláció megközelítően 0 lesz.

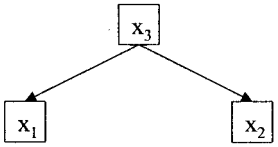
Felmerül a kérdés, hogy a parciális korreláció 0, ez önmagában bizonyítja-e a hamis korrelációt.

Vegyünk három változót, x_1-et, x_2-t, és x_3-at. Az általánosítás megszorítása nélkül tegyük fel, hogy a változók standardizáltak.

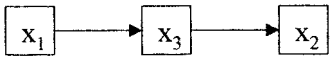
Kiszámítottuk x_1 és x_2 közötti korrelációt r_{12}-t, és azt találtuk, hogy szignifikánsan különbözik 0-tól. Kíváncsiak vagyunk, hogy ez igazi korreláció-e vagy hamis.

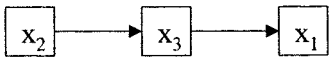
Kiszámítjuk a parciális korrelációt $r_{12,3}$, x_1 és x_2 korrelációját x_3 kiszűrése után, és összehasonlítjuk a zéró rendű korrelációval. Ha a parciális korrelációs együttható közelítően 0 (és r_{12} nem), akkor két esetet különböztetünk meg:

a) x_3 változónak a másik két változóra gyakorolt hatása eredményezte x_1 és x_2 közötti korrelációt, így az hamis korreláció. Az alábbi ábra illusztrálja ezt az esetet.



b) x_3 egy közbenső változó, x_1 (okozati) hatása x_2-re x_3 változón keresztül hat (vagy fordítva, x_2-nek x_1-re gyakorolt hatása x_3-on keresztül hat).





Ezt a korrelációt igaznak kell tekintenünk, holott a parciális korreláció 0 volt. Az okozati irányt persze ebből még nem tudjuk.

Önmagában tehát a parciális korreláció nem dönti el, hogy hamis korrelációt mértünk-e. Azonban a gyakorlatban a „józan ész" dönti el, hogy a két eset közül melyik állhat fenn.

Vizsgáljuk meg először, milyen okozati kapcsolatok lehetnek három változó között.

1. Ha a változók közvetlenül hatással vannak egymásra, a következő modellt építhetjük fel:

$$\begin{aligned} x_1 + a_{12}x_2 + a_{13}x_3 &= u_1 \\ a_{21}x_1 + x_2 + a_{23}x_3 &= u_2 \\ a_{31}x_1 + a_{32}x_2 + x_3 &= u_3 \end{aligned} \tag{6.1}$$

ahol u-k a hibakomponensek (jelezve, hogy valószínűségi változókról van szó, s ugyanakkor regressziós egyenletekről). Az $\mathbf{A} = |a_{ij}|$ mátrixot együtthatómátrixnak nevezik.

Az első egyenletben x_1, a másodikban x_2, a harmadikban x_3 felfogható függő változóként, míg a másik kettő független változóként.

2. Ha feltesszük, hogy nem minden változó hat közvetlenül a többi változóra, az egyes egyenletekből kizárunk független változókat. Ez azt jelenti, hogy egyes független változók együtthatói nullák, azaz A néhány eleme 0.

Ha az $a_{12} = a_{13} = a_{23} = 0$, akkor a (6.1) a következő formára redukálódik:

$$\begin{aligned} x_1 &= u_1 \\ a_{21}x_1 + x_2 &= u_2 \\ a_{31}x_1 + a_{32}x_2 + x_3 &= u_3 \end{aligned} \tag{6.2}$$

Ebben az esetben megkülönböztethetjük az okozatokat. Így mondhatjuk, hogy x_2 okozatilag függ x_1-től, és x_3 az x_1 és x_2 okozati függvénye.

Ha x_3 és x_2 korrelált, $a_{32} \neq 0$, akkor mondhatjuk, hogy a korreláció igazi a (6.2) rendszerben (b) eset).

3. Ha feltesszük, hogy $a_{32} = 0$, akkor

$$\begin{aligned} x_1 &= u_1 \\ a_{21}x_1 + x_2 &= u_2 \\ a_{31}x_1 + x_3 &= u_3 \end{aligned} \tag{6.3}$$

Ebben az esetben x_2, x_3 közötti korrelációt hamis korrelációnak kell tekintenünk (a) eset).

Most megmutatjuk, hogy ha *a priori* feltételezésekkel élünk, akkor eldönthető, hogy a korreláció hamis-e.

Nézzük a (6.1) egyenletrendszer paramétereinek becslését. Mivel x_1, x_2 és x_3 változókra n megfigyelési értékkel rendelkezünk, n véletlen (reziduális) tag (u_i) és 6 változó alkotja az ismeretleneket ($n + 6$), és összesen n egyenletünk van, így a szabadságfok 6. Egyértelmű megoldást csak akkor nyerünk, ha még 6 egyenletet felállítunk (ekkor a szabadságfok 0 lesz). Az *a priori* feltételek ezekkel a plusz egyenletekkel függnek össze. *A priori* feltételeknek két csoportját különböztetjük meg:

1. Egyes változók közvetlenül függnek más változóktól. Ilyen pl. ha egyes változók időben megelőznek más változókat. Ha pl. x_2 időben megelőzi x_1-et, akkor $a_{21} = 0$, azaz x_1 nincs közvetlen hatással x_2-re. Ez a feltétel tehát arra vonatkozik, hogy *a priori* ismert, hogy $a_{ij} = 0$.

2. A véletlen (reziduális) tagok (u_i) páronként korrelálatlanok, függetlenek az összes megelőző x_h ($h < i$)-tól és standardizáltak. Így ha feltesszük, hogy a változók időben megelőzik egymást és a reziduális tagok páronként korrelálatlanok, a korreláció bizonyítja az okozatot.

Feladatunk az okozati modell paramétereinek becslése. Láttuk, hogy a modell regressziós egyenesből áll. *A regressziós probléma ilyen megközelítését nevezzük útelemzésnek* (path analysis).

Az útelemzésnél általában megengedjük a magyarázó változók közötti kapcsolatokat, így a multikollinearitás problémája itt nem lép fel.

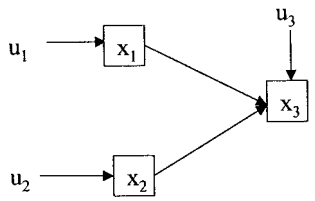
Az útelemzést Scwall Wright (1918, 1921, 1924, 1934, 1954, 1960) fejlesztette ki, majd főleg Hood és Koopman (1953), Tukey (1954) és Blaloch (1964) foglalkozott a továbbfejlesztésével.

Az útelemzés az egyszerű diagramok (gráfok) felhasználásával nem szakemberek számára is könnyen érzékelhetővé teszi a regressziós problémák leírását.

Induljunk ki a (6.2) egyenletrendszer átalakított formájából, amit a 6.1. ábra illusztrál:

$$\begin{aligned} x_1 &= u_1 \\ x_2 &= b_{21}x_1 + b_{2u}u_2 \\ x_3 &= b_{31}x_1 + b_{32}x_2 + b_{3u}u_3 \end{aligned} \tag{6.4}$$

Ezeket az egyenleteket hívják strukturális egyenleteknek, az együtthatókat pedig súlyoknak.



6.1. ábra

A (6.4) egyenletrendszer változói helyébe képzelhetjük a változók megfigyelési értékeit tartalmazó vektorokat.

A (6.4) második egyenletét szorozzuk balról $\mathbf{x}_1'$ sorvektorral.

$$\mathbf{x}_1'\mathbf{x}_2 = b_{21}\mathbf{x}_1'\mathbf{x}_1,$$

ha mindkét oldalt elosztjuk n-nel, kapjuk

$$r_{12} = b_{21}$$

(figyelembe véve, hogy a változók standardizáltak).

Ezzel megkaptuk az első paraméter becslését.

Ha (6.4) második egyenletét balról megszorozzuk $\mathbf{u}_2'$-vel, kapjuk

$$\mathbf{u}_2'\mathbf{x}_2 = b_{21}\mathbf{u}_2'\mathbf{x}_1 + b_{2u}\mathbf{u}_2'\mathbf{u}_2$$

és osztva n-nel

$$r_{2u} = b_{2u},$$

majd ugyancsak a második egyenletet $\mathbf{x}_2'\mathbf{x}_1 + b_{2ur_{2u}}$ ahonnan b_{2u} paraméter becslését kaphatjuk meg:

$$b_2 = \sqrt{1 - b_{21}r_{21}}$$

A (6.4) harmadik egyenletét egyszer $\mathbf{x}_1'$, majd $\mathbf{x}_2'$ vektorral szorozzuk:

$$\mathbf{x}_1'\mathbf{x}_3 = b_{31}\mathbf{x}_1'\mathbf{x}_1 + b_{32}\mathbf{x}_1'\mathbf{x}_2 + \mathbf{x}_1'\mathbf{u}_3$$

és

$$\mathbf{x}_2'\mathbf{x}_3 = b_{31}\mathbf{x}_2'\mathbf{x}_1 + b_{32}\mathbf{x}_2'\mathbf{x}_2 + \mathbf{x}_2'\mathbf{u}_3$$

Ha osztjuk mind a két oldalt a megfigyelések számával (n), és figyelembe vesszük a feltételeket, a következőt kapjuk:

$$r_{13} = b_{31} + b_{32}r_{12} \tag{6.5}$$

és

$$r_{23} = b_{31}r_{12} + b_{32}$$

Ebből a két egyenletből azután meghatározhatjuk b_{31} és b_{32} paraméterek becsléseit.

Ha (6.4) harmadik egyenletét $\mathbf{u}_3'$ és $\mathbf{x}_3'$-mal beszorozzuk és osztunk n-nel, kapjuk:

$$\mathbf{u}_3'\mathbf{x}_3 = b_{3u} = r_{3u}$$

és

$$1 = b_{31}r_{13} + b_{32}r_{23} + b_{3u}r_{3u}$$

ebből

$$b_{3u} = \sqrt{1 - b_{31}r_{13} - b_{32}r_{23}}$$

A (6.5) egyenletrendszert felírhatjuk a következő formában:

$$\mathbf{r}_3 = \mathbf{R}_2\mathbf{b}_3 \quad \text{és ebből} \quad \mathbf{b}_3 = \mathbf{R}_2^{-1}\mathbf{r}_3$$

feltételezve, hogy az inverz létezik.

A paraméterek (súlyok) tehát az egyes változók közvetlen hatásait fejezik ki a függő változóra, miközben a többi változó konstans. Így a paramétereket *parciális együtthatónak* is nevezik.

Ha háromnál több változót építünk be a modellbe, a paraméterek becslése az előzőekkel analóg módon határozható meg.

Az útelemzésnek – a fent ismertetett általános esetén kívül – több változata fejlődött ki.

Így pl. a következő egyenletben:

$$x_4 = b_{41}x_1 + b_{42}x_2 + b_{43}x_3 + b_{4u}u_4$$

x_3 változó helyébe beírhatjuk az x_1 és x_2 változókkal vett becslését, így a *redukált strukturális egyenlethez* jutunk; azt kapjuk, hogy

$$x_4 = b_{41}x_1 + b_{42}x_2 + b_{43}(b_{31}x_1 + b_{32}x_2 + b_{3u}u_3) + b_{4u}u_4 =$$

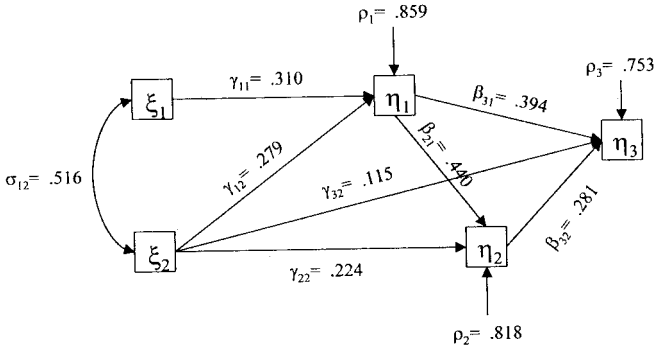
$$= (b_{41} + b_{43}b_{31}) \cdot x_1 + (b_{42} + b_{43}b_{32}) \cdot x_2 + b_{43}b_{3u}u_3 + b_{4u}u_4$$

Az x_1 együtthatója most a közvetlen hatáson kívül (b_{41}) az x_3-on keresztül kifejezett hatását ($b_{43}b_{31}$) is tartalmazza.

6.1. Direkt és indirekt hatások a lineáris strukturális egyenletek modelljeiben

A lineáris strukturális egyenletek modelljei tartalmazzák *egyrészt* a sztochasztikus lineáris egyenleteket, amelyek a változók egymásra gyakorolt hatásait fejezik ki a modellben (szokták ezeket kauzális, okozati kapcsolatoknak is nevezni, azonban a gyakorlatban ezek valóságosan ritkán „okozati" kapcsolatok, sokkal inkább „együttjárások"), *másrészt* a különböző feltételezéseket.

Nézzünk először egy konkrét példát, mégpedig Blau és Duncan (1967) rétegződési modelljét. A modellt a következő diagram ábrázolja:



6.2. ábra. A Blau és Duncan-féle rétegződési modell (1967)

ξ_1 = apa iskolai végzettsége (0–8) Minta elemszáma = 14401
ξ_2 = apa foglalkozása (0–96)
η_1 = vsz iskolai végzettsége (vsz = vizsgált személy)
η_2 = vsz első foglalkozása
η_3 = vsz foglalkozása (1962)

	ξ_1	ξ_2	η_1	η_2	η_3
ξ_1	1,000	0,516	0,453	0,332	0,322
ξ_2		1,000	0,438	0,417	0,405
η_1			1,000	0,538	0,596
η_2				1,000	0,541
η_3					1,000

Az öt státus változó korrelációs mátrixa

A Blau–Ducan-féle rétegződési modell egyenletei:

$$\begin{aligned} \eta_1 &= \gamma_{11}\xi_1 + \gamma_{12}\xi_2 + \rho_1 \\ \eta_2 &= \gamma_{22}\xi_2 + \beta_{21}\eta_1 + \rho_2 \\ \eta &= \gamma_{32}\xi_2 + \beta_{31}\eta_1 + \beta_{32}\eta_2 + \rho_3, \end{aligned} \tag{6.6}$$

vagy mátrix-formában:

$$\boldsymbol{\eta} = \boldsymbol{\beta}\,\boldsymbol{\eta} + \boldsymbol{\Gamma}\,\boldsymbol{\xi} + \boldsymbol{\rho}, \tag{6.7}$$

ahol

$$\boldsymbol{\eta} = \begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} 0 & 0 & 0 \\ \beta_{21} & 0 & 0 \\ \beta_{31} & \beta_{32} & 0 \end{pmatrix}, \quad \boldsymbol{\xi} = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}, \quad \boldsymbol{\Gamma} = \begin{pmatrix} \gamma_{11} & \gamma_{12} \\ 0 & \gamma_{22} \\ 0 & \gamma_{32} \end{pmatrix}.$$

A $\boldsymbol{\rho}$ vektor a sztochasztikus hibakomponenseket (eltéréseket) tartalmazza. Feltételezzük, hogy $E(\boldsymbol{\rho}) = \mathbf{0}$, valamint $\mathrm{cov}(\rho_1, \rho_2) = \mathrm{cov}(\rho_1\rho_3) = \mathrm{cov}(\rho_2\rho_3) = 0$. A $\boldsymbol{\rho}$ kovarianciamátrixa ($\boldsymbol{\psi}$) a feltételek miatt diagonális.

A $\boldsymbol{\beta}$ mátrix az endogén változók ($\boldsymbol{\eta}$) egymásra gyakorolt közvetlen (direkt) hatásait tartalmazza. A modell rekurzív, mivel a $\boldsymbol{\beta}$ mátrix alsó háromszög mátrix (a diagonális elemek nullák és a diagonális elemek feletti elemek is rendre nullák).

A $\boldsymbol{\Gamma}$ mátrix elemei az exogén változóknak ($\boldsymbol{\xi}$) az endogén változókra gyakorolt közvetlen (direkt) hatásait tartalmazza. A γ és β együtthatókat (közvetlen, direkt hatásokat) az ábrán a nyilakon is jelöltük.

Az ábrán azonban az is látszik, hogy az apa iskolai végzettsége (ξ_1) közvetlenül ugyan nem hat a vizsgált személy első foglalkozási státusára (η_2), de közvetetten, a megfigyelt személy iskolai végzettségén (η_1) keresztül hat. Ezt az indirekt hatást a $\beta_{21}\gamma_{11}$ szorzat fejezi ki. Ugyanakkor az apa foglalkozási státusa (ξ_2) hat a vizsgált személy foglalkozási státusára (η_3) részben a személy iskolai végzettségén (η_1) keresztül [(0,279)(0,394)], részben az iskolai végzettség és első foglalkozáson keresztül [(0,279)(0,440)(0,281)], részben az első foglalkozáson (η_2) keresztül [(0,224)(0,281)], vagyis az apa foglalkozási státusa a vsz foglalkozási státusára összességében közvetetten (indirekten)

$$(0{,}279)(0{,}394) + (0{,}279)(0{,}440)(2{,}81) + (0{,}224)(0{,}281) = 0{,}205$$

mértékben hat.

A közvetlen hatása ξ_2-nek η_3-ra 0,115, vagyis az indirekt hatás nagyobb, mint a direkt hatás. A teljes hatás a közvetlen és a közvetett hatások összege: $0,115 + 0,205 = 0,320$.

6.2. A teljes, közvetlen és közvetett hatások

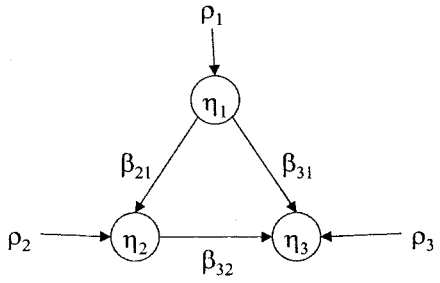
Láttuk az előzőekben, hogy az apa foglalkozási státusa hatással van a fiú (vsz) foglalkozási státusára mind közvetlenül, mind közvetetten.

A közvetett (indirekt) hatáson a legalább egy közbülső változón keresztül kifejtett hatást értjük. Az apa foglalkozási státusának a teljes hatása a két hatás, a közvetlen és közvetett (direkt és indirekt) hatás összege.

Jelöljük általában $\boldsymbol{\eta}$ endogén változók hatását az endogén változókra ($\boldsymbol{\eta}$) $\mathbf{T}_{\eta\eta}$-vel. A $\boldsymbol{\eta}$ teljes hatása $\boldsymbol{\eta}$-re:

$$\mathbf{T}_{\eta\eta} = \sum_{k=1}^{\infty} \mathbf{B}^k. \tag{6.8}$$

A $\mathbf{T}_{\eta\eta}$ akkor definiált, ha a végtelen szumma konvergált egy véges elemű mátrixhoz. Ennek illusztrálásához nézzük a következő példát:



6.3. ábra. Egy egyszerű rekurzív modell három endogén változóra

Ekkor a közvetlen hatásokat tartalmazó $\mathbf{B}$ mátrix a következő.

$$\mathbf{B} = \begin{pmatrix} 0 & 0 & 0 \\ \beta_{21} & 0 & 0 \\ \beta_{31} & \beta_{32} & 0 \end{pmatrix}.$$

A teljes hatást a definíció szerint a következőképpen írhatjuk:

$$\begin{aligned} \mathbf{T}_{\eta\eta} &= \mathbf{B} + \mathbf{B}^2 + \mathbf{B}^3 + \dots \\ &= \begin{pmatrix} 0 & 0 & 0 \\ \beta_{21} & 0 & 0 \\ \beta_{31} & \beta_{32} & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \beta_{21}\beta_{32} & 0 & 0 \end{pmatrix} + \mathbf{0} + \dots \end{aligned} \tag{6.9}$$

Láthatjuk, hogy $\mathbf{B}^k \quad k \geq 3$-ra nullmátrix, így az összeg konvergens, és a teljes hatás meghatározható. Az összeg első tagja ($\mathbf{B}$) a közvetlen hatásokat tartalmazza, a magasabb

hatványú tagok (2 vagy több) a közvetett hatásokat tartalmazzák a kapcsolódások számának (hosszának) megfelelően. A példában a második tag fejezi ki a közvetett hatást, ami 2 hosszúságú kapcsolódás, η_1-nek η_3- ra gyakorolt hatását mutatja, amit az η_2-n keresztül fejt ki. A $\mathbf{B}^3$ és a magasabb hatványú tagok rendre nullák, ami azt mutatja, hogy a három vagy több közvetítőn keresztül, háromnál hosszabb kapcsolódások rendre nullák. A példában a teljes hatások mátrixa:

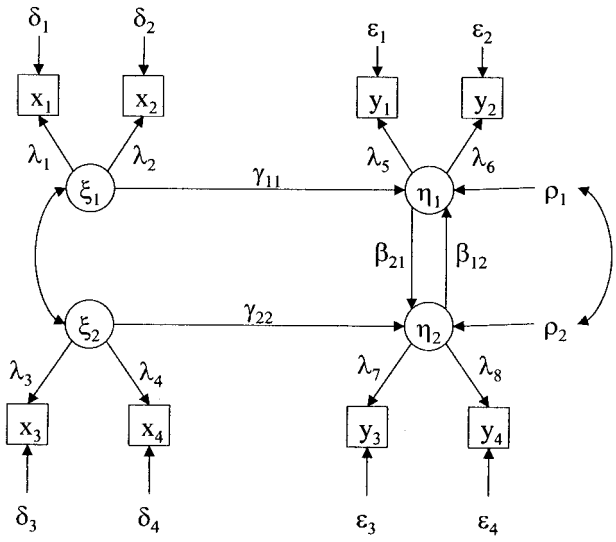
$$\mathbf{T}_{\eta\eta} = \begin{pmatrix} 0 & 0 & 0 \\ \beta_{21} & 0 & 0 \\ \beta_{31} + \beta_{21}\beta_{32} & \beta_{32} & 0 \end{pmatrix}.$$

Általában, az indirekt hatásokat a teljes és a direkt hatások különbségeként határozzuk meg.

A rekurzív modelleknél, ahol a $\mathbf{B}$ mátrix alsó háromszög mátrix, $\mathbf{B}^k = \mathbf{0}$ ha $k \geq m$, ahol m a $\boldsymbol{\eta}$, endogén változók száma.

A teljes hatások mátrixa ekkor $T = \sum_{k=1}^{m-1} \mathbf{B}^k$. Ebből következik, hogy a közvetett hatások mátrixa $\mathbf{I}_{\eta\eta} = \mathbf{T}_\eta - \mathbf{B} = \sum_{k=2}^{m-1} \mathbf{B}^k$.

A nemrekurzív modelleknél a helyzet bonyolultabb. Nézzük először a következő példát:



6.4. ábra. Nemrekurzív latens változós modell és a mérési modell

A példában $\mathbf{B} = \begin{pmatrix} 0 & \beta_{12} \\ \beta_{21} & 0 \end{pmatrix}$, ahol β_{21} η_1 hatását fejezi ki η_2-re, β_{12} pedig η_2 hatását η_1-e. A közvetett hatások a $\mathbf{B}$ mátrix hatványai:

$$\mathbf{B}^2 = \begin{pmatrix} \beta_{21}\beta_{12} & 0 \\ 0 & \beta_{21}\beta_{12} \end{pmatrix}, \qquad \mathbf{B}^3 = \begin{pmatrix} 0 & \beta_{21}\beta_{12}^2 \\ \beta_{21}^2\beta_{12} & 0 \end{pmatrix},$$

$$= \mathbf{B}^4 = \begin{pmatrix} \beta_{21}^2\beta_{12}^2 & 0 \\ 0 & \beta_{21}^2\beta_{12}^2 \end{pmatrix}.$$

A rekurzív modellel ellentétben a $\mathbf{B}^n$ nem szükségképpen nulla, ha $k \geq m$. A $\mathbf{B}^2$ és $\mathbf{B}^4$ illusztrálja, hogy a η_1-nek és η_2-nek is van közvetett hatása saját magára. Ebből látszik, hogy $\mathbf{T}_{\eta\eta}$ elemei végtelen összegek.

Általánosságban, hogy a $\mathbf{T}_{\eta\eta}$ teljes hatásokat meg tudjuk határozni, a $\mathbf{B}^n$-nek nullához kell konvergálnia $k \longrightarrow \infty$ esetén. Ez akkor és csak akkor teljesül, ha a $\mathbf{B}$ legnagyobb sajátértékének abszolút értéke kisebb mint 1 (lásd Bentler és Freeman 1983, 144). Határozzuk meg $\mathbf{T}_{\eta\eta}$-t a definíciós egyenletből

$$\mathbf{T}_{\eta\eta} = \mathbf{B} + \mathbf{B}^2 + \ldots + \mathbf{B}^k, \tag{6.10}$$

mivel

$$\mathbf{B}^{k+1} \longrightarrow \mathbf{0} \quad \text{ha} \quad k \longrightarrow \infty$$

Adjunk az egyenlet mindkét oldalához $\mathbf{I}$-t ($\mathbf{I} = \mathbf{B}^0$)

$$\mathbf{I} + \mathbf{T}_{\eta\eta} = \mathbf{I} + \mathbf{B} + \mathbf{B}^2 + \ldots + \mathbf{B}^k.$$

Szorozzuk meg az egyenlet mindkét oldalát

$$\begin{aligned}(\mathbf{I} - \mathbf{B})(\mathbf{I} + \mathbf{T}_{\eta\eta}) &= (\mathbf{I} - \mathbf{B})(\mathbf{I} + \mathbf{B} + \mathbf{B}^2 + \ldots + \mathbf{B}^k) \\ &= \mathbf{I} - \mathbf{B}^{k+1}.\end{aligned}$$

Mivel $\mathbf{B}^{k+1} \longrightarrow \mathbf{0}$ ha $k \longrightarrow \infty$

$$\mathbf{T}_{\eta\eta} = (\mathbf{I} - \mathbf{B})^{-1} - \mathbf{I}. \tag{6.11}$$

Az indirekt hatásokat megkaphatjuk, ha $\mathbf{T}_{\eta\eta}$-ből levonjuk $\mathbf{B}$-t

$$\mathbf{I}_{\eta\eta} = (\mathbf{I} - \mathbf{B})^{-1} - \mathbf{I} - \mathbf{B}. \tag{6.12}$$

Idáig csak az endogén latens változók egymásra gyakorolt hatásaival foglalkoztunk. Nézzük ezután, hogyan kaphatjuk meg a latens exogén változók közvetett hatásait a latens endogén változókra.

Induljunk ki a lineáris strukturális egyenletek alapképletéből.

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\,\boldsymbol{\xi} + \boldsymbol{\rho}. \tag{6.13}$$

Helyettesítsük az egyenlet jobb oldalát η helyébe:

$$\begin{aligned}\boldsymbol{\eta} &= \mathbf{B}[\mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\,\boldsymbol{\xi} + \boldsymbol{\rho}]\boldsymbol{\eta} + \boldsymbol{\Gamma}\,\boldsymbol{\xi} + \boldsymbol{\rho} \\ &= \mathbf{B}^2\boldsymbol{\eta} + (\mathbf{I} + \mathbf{B})(\boldsymbol{\Gamma}\,\boldsymbol{\xi} + \boldsymbol{\rho}) \\ &= \mathbf{B}^2[\mathbf{B}\,\boldsymbol{\eta} + \boldsymbol{\Gamma}\,\boldsymbol{\xi} + \boldsymbol{\rho}] + (\mathbf{I} + \mathbf{B})(\boldsymbol{\Gamma}\,\boldsymbol{\xi} + \boldsymbol{\rho}) \\ &= \mathbf{B}^3\boldsymbol{\eta} + (\mathbf{I} + \mathbf{B} + \mathbf{B}^2)(\boldsymbol{\Gamma}\,\boldsymbol{\xi} + \boldsymbol{\rho}) \\ &\vdots \\ &= \mathbf{B}^k\boldsymbol{\eta} + (\mathbf{I} + \mathbf{B} + \mathbf{B}^2 + \ldots + \mathbf{B}^{k-1})(\boldsymbol{\Gamma}\,\boldsymbol{\xi} + \boldsymbol{\rho}).\end{aligned} \tag{6.14}$$

Az endogén változók ($\boldsymbol{\xi}$) teljes hatását az utolsó egyenlet jobb oldalán a $\boldsymbol{\xi}$ együttható-mátrixa tartalmazza: $(\mathbf{I} + \mathbf{B} + \mathbf{B}^2 + \ldots + \mathbf{B}^{k-1})\boldsymbol{\Gamma}$. A szorzat első tényezője $(\mathbf{I} + \mathbf{B} + \mathbf{B}^2 + \ldots + \mathbf{B}^{k-1})$ konvergál a $(\mathbf{I} - \mathbf{B})^{-1}$ inverzmátrixhoz, ahogyan ezt korábban már bemutattuk. (Ennek feltétele, hogy a $\mathbf{B}$ legnagyobb sajátértékének abszolút értéke kisebb legyen 1-nél).

A teljes hatások mátrixa így

$$\mathbf{T}_{\eta\xi} = (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma}. \tag{6.15}$$

Mivel $\boldsymbol{\xi}$ közvetlen hatásait $\boldsymbol{\eta}$-re a $\boldsymbol{\Gamma}$ mátrix tartalmazza, $\boldsymbol{\xi}$ közvetett hatása $\boldsymbol{\eta}$-ra:

$$\mathbf{I}_{\eta\xi} = (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma} - \boldsymbol{\Gamma} = \left[(\mathbf{I} - \mathbf{B})^{-1} - \mathbf{I}\right]\boldsymbol{\Gamma}. \tag{6.16}$$

Ez az egyenlet azt mutatja, hogy $\boldsymbol{\xi}$ exogén változók indirekt hatásai a latens endogén változókra ($\boldsymbol{\eta}$) a latens endogén változók egymásra gyakorolt teljes hatásainak és a $\boldsymbol{\xi}$-nek $\boldsymbol{\eta}$-ra gyakorolt direkt hatásainak a szorzata.

A latens exogén változóknak a megfigyelt endogén változókra ($\mathbf{y}$) gyakorolt hatásait a fentiekhez hasonlóan kaphatjuk meg.

$$\begin{aligned}\mathbf{y} &= \boldsymbol{\Lambda}_y\boldsymbol{\eta} + \boldsymbol{\varepsilon}\\ &= \boldsymbol{\Lambda}_y[\mathbf{B}^k\boldsymbol{\eta} + (\mathbf{I} + \mathbf{B} + \mathbf{B}^2 + \ldots + \mathbf{B}^{k-1})(\boldsymbol{\Gamma}\,\boldsymbol{\xi} + \boldsymbol{\rho}) + \boldsymbol{\varepsilon}] \qquad (6.17)\\ &= \boldsymbol{\Lambda}_y\mathbf{B}^k\boldsymbol{\eta} + \boldsymbol{\Lambda}_y(\mathbf{I} + \mathbf{B} + \mathbf{B}^2 + \ldots + \mathbf{B}^{k-1})\boldsymbol{\Gamma}\,\boldsymbol{\xi}\\ &\quad + \boldsymbol{\Lambda}_y(\mathbf{I} + \mathbf{B} + \mathbf{B}^2 + \ldots + \mathbf{B}^{k-1})\boldsymbol{\rho} + \boldsymbol{\varepsilon}.\end{aligned}$$

Feltételezzük, hogy a $\mathbf{B}$ konvergens, a teljes hatások mátrixa:

$$\mathbf{T}_{y\xi} = \boldsymbol{\Lambda}_y(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma}, \qquad (6.18)$$

ami egyben $\boldsymbol{\xi}$-nek $\mathbf{y}$-ra gyakorolt közvetett hatásait is tartalmazza, mivel $\boldsymbol{\xi}$-nek nincs közvetlen kapcsolata $\mathbf{y}$-nal.

Hasonló logikával kaphatjuk meg $\boldsymbol{\eta}$-nak $\mathbf{y}$-ra gyakorolt hatásait is:

$$\begin{aligned}\mathbf{T}_{y\eta} &= \boldsymbol{\Lambda}_y(\mathbf{I} + \mathbf{B} + \mathbf{B}^2 + \ldots + \mathbf{B}^k) \qquad (6.19)\\ &= \boldsymbol{\Lambda}_y(\mathbf{I} - \mathbf{B})^{-1},\end{aligned}$$

és az indirekt hatások:

$$\mathbf{I}_{y\eta} = \boldsymbol{\Lambda}_y(\mathbf{I} - \mathbf{B})^{-1} - \boldsymbol{\Lambda}_y = \boldsymbol{\Lambda}_y[(\mathbf{I} - \mathbf{B})^{-1} - \mathbf{I}]. \qquad (6.20)$$

Ahogyan ezt az előzőekben megállapítottuk, ahhoz, hogy a teljes és az indirekt hatásokat meg tudjuk határozni, szükséges feltételeznünk, hogy a $\mathbf{B}$ mátrix sajátértékeinek abszolút értékei kisebbek legyenek 1-nél. A sajátértékeket azonban nem ismerjük mindig. Két esetben azonban egyszerű a helyzetünk. A rekurzív modellek esetében a $\mathbf{B}$ mátrix alsó háromszögmátrix, és így $\mathbf{B}^k$ egyenlő a nullmátrixszal, ha $k \geq m$. A másik esetben, ha feltételezzük, hogy a $\mathbf{B}$ elemei pozitívak és az összegük minden oszlopban kisebb 1-nél, akkor a sajátértékek abszolút értékei is kisebbek 1-nél (lásd Goldberger 1958, 237–38). Ez utóbbi feltétel azonban csak szükséges, de nem elégséges feltétel. Az egyszerű számítási eljárás miatt viszont hasznos a gyakorlatban.

A következő táblázatban összefoglaljuk a fenti eredményeket. Ezeket alkalmazhatjuk akkor is, ha a modell nem tartalmaz latens változókat, pl. ha $\mathbf{x} \equiv \boldsymbol{\xi}$ és $\mathbf{y} \equiv \boldsymbol{\eta}$.

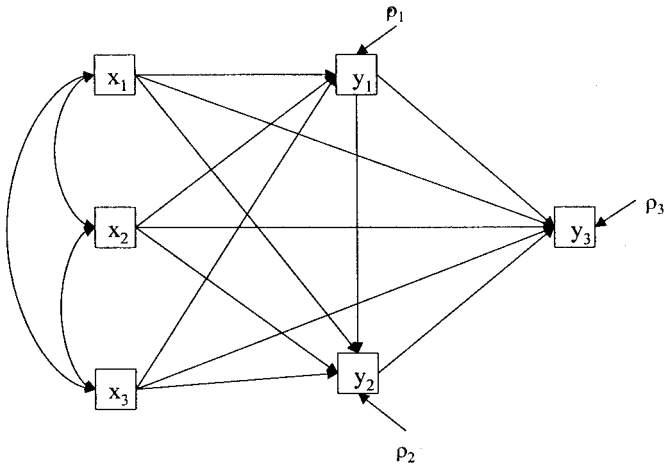
Független változó	Függő változó		
	$\boldsymbol{\eta}$	$\mathbf{y}$	$\mathbf{x}$
$\boldsymbol{\xi}$ hatásai:			
Direkt hatás	$\boldsymbol{\Gamma}$	$\mathbf{0}$	$\boldsymbol{\Lambda}_x$
Indirekt hatás	$(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma} - \boldsymbol{\Gamma}$	$\boldsymbol{\Lambda}_y(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma}$	$\mathbf{0}$
Teljes hatás	$(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma}$	$\boldsymbol{\Lambda}_y(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma}$	$\boldsymbol{\Lambda}_x$
$\mathbf{x}$ hatásai:			
Direkt hatás	$\mathbf{B}$	$\boldsymbol{\Lambda}_y$	$\mathbf{0}$
Indirekt hatás	$(\mathbf{I} - \mathbf{B})^{-1} - \mathbf{I} - \mathbf{B}$	$\boldsymbol{\Lambda}_y(\mathbf{I} - \mathbf{B})^{-1} - \boldsymbol{\Lambda}_y$	$\mathbf{0}$
Teljes hatás	$(\mathbf{I} - \mathbf{B})^{-1} - \mathbf{I}$	$\boldsymbol{\Lambda}_y(\mathbf{I} - \mathbf{B})^{-1}$	$\mathbf{0}$

6.1. táblázat. A strukturális egyenletek változásainak egymásra gyakorolt közvetlen, közvetett és teljes hatásai.

6.3. Specifikus hatások

Az indirekt hatások magukba foglalják az összes utat, amely valamely változóból kiindulva a másik, függő változóba elvezet. Az egyes, közbülső változók hatása így ismeretlen marad. Sokszor érdekes lehet pontosan ismerni egy-egy változó-csoport közvetítő hatását is. Az ilyen hatásokat specifikus indirekt hatásoknak nevezzük. A közvetkezőkben azt mutatjuk meg, hogyan számíthatjuk ezeket a specifikus hatásokat.

Induljunk ki a következő modellből:



6.5. ábra. Rekurzív modell a specifikus hatások bemutatására

A fenti ábrán látható modell paraméter-mátrixai:

$$\mathbf{B} = \begin{pmatrix} 0 & 0 & 0 \\ \beta_{21} & 0 & 0 \\ \beta_{31} & \beta_{32} & 0 \end{pmatrix} \quad \mathbf{\Gamma} = \begin{pmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} \\ \gamma_{21} & \gamma_{22} & \gamma_{23} \\ \gamma_{31} & \gamma_{32} & \gamma_{33} \end{pmatrix}.$$

Tételezzük fel, hogy az **x** változók specifikus hatását akarjuk becsülni **y** változókra, amelyet **x** az **y** endogén változón keresztül kifejt. Ha a modellből elhagyjuk azokat a kapcsolódásokat (nyilakat), amelyek y_1-gyel függnek össze (elhagyjuk azokat a nyilakat, amelyek y_1-be vezetnek, illetve y_1-ből vezetnek tovább), és az így módosított modellben számítjuk az indirekt hatásokat, akkor az már nem tartalmazhatja az y_1-gyel összefüggő hatásokat.

Ha a módosított modellből számított indirekt hatásokat kivonjuk az eredeti modell indirekt hatásaiból, akkor pontosan azokat a specifikus hatásokat kapjuk, amelyek az y_1-en keresztül adódtak az indirekt hatásokhoz. Általánosan a specifikus indirekt hatásokat a következő képpen számolhatjuk: (1) meghatározzuk, milyen változtatások szükségesek a paramétermátrixokon, (2) módosítjuk a paramétermátrixokat, (3) ha **B** változott, akkor ellenőrizzük, hogy **B** legnagyobb sajátértékeinek abszolút értéke kisebb-e 1-nél (lehetséges, hogy az eredeti **B** kielégítette ezt a feltételt, de a módosított **B** esetében nem teljesül (lásd Fisher [1970]; Sobel [1986]), (4) kiszámítjuk a direkt, indirekt és teljes hatásokat a módosított paramétermátrixok alapján, (5) kivonjuk az így kapott paramétermátrixokat az eredetiből, és az indirekt hatások új mátrixában találhatjuk a specifikus

hatásokat. Ezután esetleg más módosítások alapján újabb specifikus indirekt hatásokat keresünk.

Az előbbi példában a módosított paramétermátrixok az y_1 hatásának kiszűrésével a következők:

$$\mathbf{B}_{\ell_1} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & \beta_{32} & 0 \end{pmatrix} \qquad \mathbf{\Gamma}_{\ell_2} = \begin{pmatrix} 0 & 0 & 0 \\ \gamma_{31} & \gamma_{32} & \gamma_{33} \end{pmatrix}$$

ahol

$$\ell_1 = \{(., 1) = 0\}$$

$$\ell_2 = \{(1, .) = 0\}$$

Az ℓ_1 index jelöli azt a feltételt amit a módosítással vezettünk be, itt a $\mathbf{B}$ első oszlopvektorának elemeit – amelyek y_1 hatásait tartalmazzák a többi endogén változóra – rendre egyenlővé tettük 0-val $[(., 1) = 0]$. Hasonlóan ℓ_2 azt jelöli, hogy a $\mathbf{\Gamma}$ mátrix első sorának elemei nullák $[(1, .) = 0]$, mivel elhagytuk az y_1-be vezető nyilakat.

A módosított $\mathbf{B}_{(\ell_1)}$ mátrixnál ellenőrizzük a szükséges stabilitási feltételt, ami itt teljesül, mivel $\mathbf{B}_{(\ell_1)}$ alsó háromszögmátrix.

Az eredeti modellben az indirekt hatások $\mathbf{I}_{yx}$ mátrixát a következőképpen számoljuk:

$$\mathbf{I}_{yx} = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{\Gamma} - \mathbf{\Gamma} = \tag{6.21}$$

$$= \begin{pmatrix} 0 & 0 & 0 \\ \beta_{21}\gamma_{11} & \beta_{21}\gamma_{12} & \beta_{21}\gamma_{13} \\ [(\beta_{31} + \beta_{21}\beta_{32})\gamma_{11} & [(\beta_{31} + \beta_{21}\beta_{32})\gamma_{12} & [(\beta_{31} + \beta_{21}\beta_{32})\gamma_{13} \\ +\beta_{32}\gamma_{21}] & +\beta_{32}\gamma_{22}] & +\beta_{32}\gamma_{33}] \end{pmatrix}$$

Az indirekt hatások a $\mathbf{B}_{(\ell_1)}$ és $\mathbf{\Gamma}_{(\ell_2)}$ mátrixok alapján:

$$\left(\mathbf{I} - \mathbf{B}_{(\ell_1)}\right)^{-1} \mathbf{\Gamma}_{(\ell_2)} - \mathbf{\Gamma}_{(\ell_2)} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \beta_{32}\gamma_{21} & \beta_{32}\gamma_{22} & \beta_{32}\gamma_{23} \end{pmatrix} \tag{6.22}$$

A fenti közvetett hatások az y_1 kihagyásával az $\mathbf{x}$ változóknak az y_2 változón keresztül gyakorolt hatásait tartalmazza.

Ez utóbbi mátrixot, ha kivonjuk az eredeti modellből számított indirekt hatások mátrixából, megkapjuk y_1 specifikus hatásait:

$$\begin{pmatrix} 0 & 0 & 0 \\ \beta_{21}\gamma_{11} & \beta_{21}\gamma_{12} & \beta_{21}\gamma_{13} \\ (\beta_{31} + \beta_{21}\beta_{32})\,\gamma_{11} & (\beta_{31} + \beta_{21}\beta_{32})\,\gamma_{12} & (\beta_{31} + \beta_{21}\beta_{32})\,\gamma_{13} \end{pmatrix} \tag{6.23}$$

Általánosságban egy vagy több változón keresztül gyakorolt specifikus hatásokat a következőképpen számolhatjuk. Először kiszámítjuk az eredeti modell indirekt hatásait. Másodszor, módosítjuk a $\mathbf{B}$ és $\mathbf{\Gamma}$ mátrixokat úgy, hogy nullával tesszük egyenlővé azokat az együtthatókat, amelyek azokhoz a nyilakhoz tartoznak, amelyek abba a változó(k)ba tartanak vagy abból mutatnak tovább, amely(ek)nek a hatásait vizsgáljuk. Harmadszor, feltételezve, hogy az új $\mathbf{B}$ mátrix kielégíti a stabilitási feltételezést, kiszámítjuk az indirekt hatásokat az új, módosított mátrixszal. Végül az új indirekt hatásokat tartalmazó mátrixot (amely nem tartalmazza már a kérdéses változók közvetítő hatásait) kivonjuk az eredeti indirekt hatásokat tartalmazó mátrixból, így megkapjuk azokat a specifikus hatásokat, amelyek a vizsgált változókon keresztül hatottak.

Specifikus hatásokat számíthatunk úgy is, hogy nem egy változón (vagy változókon) keresztül gyakorolt hatásokat vizsgáljuk, hanem egy (vagy több) úton (nyílon) keresztül

kifejtett hatásokat. Tekintsük például a nemrekurzív modellekre bemutatott ábrát (6.4. ábra). Becsülni szeretnénk a $\boldsymbol{\xi}$ változók teljes hatását az endogén változókra $\boldsymbol{\eta}$, amit az η_2-ból η_1-be vezető úton kifejtenek.

Most is hasonlóan járunk el, mint az előzőekben. Először módosítjuk az együtthatómátrixokat úgy, hogy azok ne tartalmazzák a vizsgált hatást. Kiszámítjuk a módosított mátrixokkal a teljes hatások mátrixát, azután az eredeti mátrixból kivonva a kapott eredményeket, jutunk azokhoz a teljes specifikus hatásokhoz, amelyek η_2-ból η_1-en keresztül vezető úthoz kötődnek.

A példában a β_{12} együtthatót ha nullával tesszük egyenlővé, kiszűrjük azokat a hatásokat, amelyek η_2-ből η_1-be, és azon keresztül vezetnek. A módosított $\mathbf{B}$ mátrixa:

$$\mathbf{B}_{(\ell)} = \begin{pmatrix} 0 & 0 \\ \beta_{21} & 0 \end{pmatrix}, \tag{6.24}$$

ahol $\ell = \{\beta_{12} = 0\}$.

A módosított modellben $\boldsymbol{\xi}$ hatásait $\boldsymbol{\eta}$-ra nem változtattuk, így $\boldsymbol{\Gamma}$ mátrix nem változik.

A $\mathbf{B}_{(\ell)}$ együtthatómátrix alsó háromszög mátrix, így teljesíti a stabilitási feltételezést. A következő lépésben kiszámítjuk a teljes hatások együtthatóit, amelyek nem tartalmazzák a $\eta_2 \longrightarrow \eta_1$ utat ($\beta_{12} = 0$):

$$\mathbf{T}_{\eta\xi(\ell)} = \left(\mathbf{I} - \mathbf{B}_{(\ell)}\right)^{-1} \boldsymbol{\Gamma} = \begin{pmatrix} \gamma_{11} & 0 \\ \beta_{21}\gamma_{11} & \gamma_{22} \end{pmatrix}. \tag{6.25}$$

Az eredeti teljes hatások mátrixa:

$$\mathbf{T}_{\eta\xi} = (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma} = (1 - \beta_{21}\beta_{12})^{-1} \begin{pmatrix} \gamma_{11} & \beta_{12}\gamma_{22} \\ \beta_{21}\gamma_{11} & \gamma_{22} \end{pmatrix} \tag{6.26}$$

A módosított mátrixokkal számított teljes hatások mátrixát kivonjuk az eredeti $\mathbf{T}_{\eta\xi}$ mátrixból, megkapjuk a teljes hatásoknak azt a részét, ami az η_2-nek η_1-re gyakorolt hatásával függ össze.

A specifikus teljes hatás:

$$(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma} - \left(\mathbf{I} - \mathbf{B}_{(\ell)}\right)^{-1} \boldsymbol{\Gamma} = \tag{6.27}$$

$$= (1 - \beta_{21}\beta_{12})^{-1} \begin{pmatrix} \beta_{21}\beta_{12}\gamma_{11} & \beta_{12}\gamma_{22} \\ \beta_{21}^2\beta_{12}\gamma_{11} & \beta_{21}\beta_{12}\gamma_{22} \end{pmatrix}.$$

Például a x_1 teljes specifikus hatása η_1 endogén változóra, amely az η_2-nek η_1-re gyakorolt hatásának az eredménye:

$$(1 - \beta_{21}\beta_{22})^{-1}(\beta_{21}\beta_{12}\gamma_{11}).$$

A közvetlen hatásokat kifejező együtthatókkal együtt azok varianciáit és a becsléseit is közvetlenül megkapjuk a becslési eljárás végén. A közvetett és teljes hatások azonban a közvetlen hatások függvényei, így ezek varianciáinak a becslésére nem alkalmazhatjuk a szokásos formulákat.

Folmer (1981) és Sobel (1982, 1986) javasolta a „delta-módszert" (Bishop, Fienberg és Holland [1975]) alkalmazni a teljes, indirekt és a különböző specifikus hatások aszimptotikus varianciáinak a becslésére. A módszer a paraméterek aszimptotikus eloszlásának vizsgálatán és a paraméterek függvényének lineáris approximációján alapul.

Tekintsük először egy paraméter $\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_n, \ldots$ becsléseit, és az ezekhez tartozó $F_1, F_2, \ldots, F_n, \ldots$ eloszlásfüggvényeket.

Ha az F_n eloszlásfüggvény konvergál az F eloszlásfüggvényhez $n \longrightarrow \infty$ esetén, akkor az F a $\widehat{\theta}_n$ konzisztens becslése az elméleti θ értéknek, akkor

$$\lim_{n \longrightarrow \infty} P[(\widehat{\theta}_n - \theta) < \delta] = 1, \tag{6.28}$$

bármilyen $\delta > 0$ értékre, (vagy másképpen írva $\lim_{n \longrightarrow \infty} \theta_n = \theta$), akkor az $F(\ )$ degenerált eloszlás, mivel egy konstanshoz kovergál (feltéve, hogy θ konstans). Ennek ellenére hasznos a $\widehat{\theta}_n$ aszimptotikus eloszlásának tanulmányozása, ha nem ismerjük a minta eloszlását, vagy az nagyon bonyolult.

Tételezzük fel, hogy $\widehat{\theta}_n$ az elméleti θ paraméternek egy becslése, és hogy $p \lim \widehat{\theta}_n = \theta$. Mivel $\widehat{\theta}_n$ konvergál egy konstans θ értékhez, $\widehat{\theta}_n$ eloszlása degenerált. Azonban, ha a $(\widehat{\theta}_n - \theta)$ különbséget megszorozzuk $\sqrt{n}$-nel, a $\sqrt{n}\ (\widehat{\theta}_n - \theta)$ eloszlását tanulmányozhatjuk, és annak alapján következtetéseket vonhatunk le $\widehat{\theta}_n$ viselkedéséről nagy minták esetén.

A központi határeloszlás tétele értelmében a $\sqrt{n}(\widehat{\theta}_n - \theta)$ eloszlása közelíti a normális eloszlást:

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{D} N(0, V),$$

ahol N a normális eloszlást jelöli, 0 várható értékkel és V varianciával, $\xrightarrow{D}$ jelöli azt, hogy aszimptotikus eloszlásról van szó.

Ebben az esetben $\widehat{\theta}_n$ aszimptotikusan normális eloszlású θ várható értékkel és $n^{-1}V$ varianciával:

$$\widehat{\theta}_n \sim AN(\theta, V/n), \tag{6.29}$$

ahol AN az aszimptotikusan normális eloszlást jelöli, $\mathrm{AVAR}(\widehat{\theta}_n) = V/n$ a $\widehat{\theta}_n$ aszimptotikus varianciája.

Illusztrációként nézzük a mintaátlagot, $\overline{X}_n$-t (az n index jelöli a minta elemszámát).

Ha az x valószínűségi változó elméleti várható értéke μ, varianciája σ^2, akkor az előzőek szerint:

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{D} N(0, \sigma^2).$$

Ebben az esetben $\widehat{\theta}_n = \overline{X}_n$, $\theta = \mu$ és $V = \sigma^2$ és $\overline{X} \approx AN(\mu, \sigma^2/n)$, $\mathrm{AVAR}(\overline{X}) = \sigma^2/n$ és ennek becslése $avar(\overline{X}) = s^2/n$.

Az előzőeket könnyen általánosíthatjuk több paraméteres modellre is.

Tekintsük a $\boldsymbol{\theta}$ paraméter vektor-változót, amely a $\mathbf{B}$, $\boldsymbol{\Gamma}$ és $\boldsymbol{\Lambda}_y$ ismeretlen paramétereit tartalmazza. A $\boldsymbol{\theta}$ vektor elemeinek száma legyen s. A $\boldsymbol{\theta}$ becslése legyen $\widehat{\boldsymbol{\theta}}_n$, ahol n a minta elemszáma. Válasszunk olyan becslést, hogy $\widehat{\boldsymbol{\theta}}_n$ eloszlása aszimptotikusan normális legyen $\boldsymbol{\theta}$ várható értékkel és $n^{-1}\mathbf{V}$ aszimptotikus kovarianciamátrixszal ($\mathrm{ACOV}(\widehat{\boldsymbol{\theta}}_n) = n^{-1}\mathbf{V}$), ahol $\mathbf{V}$ kovarianciamátrixa a $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$ határeloszlásnak. A maximum likelihood és az általánosított legkisebb négyzetek módszere kielégíti ezeket a feltételezéseket.

Ezután az általános bevezetés után nézzük a delta-módszert a teljes, közvetett és specifikus hatások aszimptotikus varianciáinak a becslésére. Definiáljuk az r–elemű $\mathbf{f}(\boldsymbol{\theta})$ vektort, ami a $\boldsymbol{\theta}$ paraméterek függvénye. Tételezzük fel, hogy $\mathbf{f}(\boldsymbol{\theta})$ differenciálható függvény. Esetünkben $\mathbf{f}(\boldsymbol{\theta})$ a közvetett (vagy teljes, vagy specifikus) hatásokat tartalmazza, és ezek a közvetlen hatások függvényei.

A delta-módszer szerint ha $\mathbf{f}$ differenciálható, akkor

$$\sqrt{n}(\mathbf{f}(\boldsymbol{\theta}_n) - \mathbf{f}(\boldsymbol{\theta})) \xrightarrow{D} N(\mathbf{0}, (\theta\mathbf{f}/\theta\boldsymbol{\theta})'\mathbf{V}(\boldsymbol{\theta}_n)(\theta\mathbf{f}/\theta\boldsymbol{\theta})).$$

Egy paraméter és annak kétszer deriválható függvénye (f) esetén a delta-módszer a következő: a Taylor kifejtése az f függvénynek:

$$f(\widehat{\theta}_n) - f(\theta) = f'(\theta)(\widehat{\theta}_n - \theta) + f''(\theta^*)(\widehat{\theta}_n - \theta)^2/2, \tag{6.30}$$

ahol θ^* a $(\widehat{\theta}_n - \theta)$ intervallumba eső érték. Nagy n esetén θ^* közel esik θ-hez, és a jobb oldalon az első tag dominálja a második tagot, amit elhagyva:

$$f(\widehat{\theta}_n) - f(\theta) \approx f'(\theta)(\widehat{\theta}_n - \theta),$$

vagyis az $f(\widehat{\theta}_n) - f(\theta)$ lineáris függvénye $\widehat{\theta}_n$-nek. Mivel θ_n közelítőleg normális eloszlású nagy minta esetén (feltételezésünk szerint), a lineáris függvénye is közelítőleg normális eloszlású lesz. Így $f(\widehat{\theta}_n)$ aszimptotikusan normális eloszlású, várható értéke

$$E(f(\widehat{\theta}_n)) \approx f(\theta),$$

varianciája:

$$\mathrm{AVAR}(f(\widehat{\theta}_n)) \approx n^{-1}[f'(\theta)]^2 V.$$

Többváltozós esetben, tekintsük a paraméterek $\boldsymbol{\theta}$ vektorát és annak egy differenciálható függvényét f-et. Ha

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{D} N(\mathbf{0},\ \mathbf{V})$$

akkor

$$\sqrt{n}(f(\widehat{\boldsymbol{\theta}}_n) - f(\boldsymbol{\theta})) \xrightarrow{D} N(\mathbf{0}(\theta f/\theta\boldsymbol{\theta})'\mathbf{V}(\theta f/\theta\boldsymbol{\theta})).$$

Az $f(\widehat{\boldsymbol{\theta}}_n)$ Taylor-kifejtése:

$$f(\widehat{\boldsymbol{\theta}}) - f\boldsymbol{\theta}) = (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})'(\theta f/\theta\boldsymbol{\theta}) + R_n,$$

ahol R_n közelít 0-hoz ha $n \longrightarrow \infty$, így

$$f(\widehat{\boldsymbol{\theta}}) - f(\boldsymbol{\theta}) \approx (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})'(\theta f/\theta\boldsymbol{\theta}).$$

Ez a $(\widehat{\boldsymbol{\theta}}_n)$-nek lineáris függvénye, így aszimptotikusan normális eloszlású, $E(f(\widehat{\theta}_n)) \approx f(\boldsymbol{\theta})$, és aszimptotikus varianciája:

$$\mathrm{AVAR}(f(\widehat{\boldsymbol{\theta}}_n)) \approx n^{-1}[(\theta f/\theta\boldsymbol{\theta})'\mathbf{V}(\theta f/\theta\boldsymbol{\theta})].$$

Ennek általánosításával, ha $\mathbf{f}$ r-elemű vektorfüggvénye a paramétereknek ($\boldsymbol{\theta}$), akkor az $\mathbf{f}(\widehat{\boldsymbol{\theta}}_n)$ vektor várható értéke:

$$E(\mathbf{f}(\widehat{\boldsymbol{\theta}}_n)) \approx \mathbf{f}(\boldsymbol{\theta}),$$

és aszimptotikus kovarianciamátrixa:

$$\mathrm{ACOV}(\mathbf{f}(\widehat{\boldsymbol{\theta}}_n)) = (\theta\mathbf{f}/\theta\boldsymbol{\theta})'\mathrm{ACOV}(\widehat{\theta}_n)(\theta\mathbf{f}/\theta\boldsymbol{\theta}).$$

Az $(\theta\mathbf{f}/\theta\boldsymbol{\theta})$ első sorvektora: $[\theta f_1/\theta\theta_1, \theta f_2/\theta\theta_1, \ldots, \theta f_r/\theta\theta_1]$,
második sorvektora: $[\theta f_1/\theta\theta_2,\ \theta f_2/\theta\theta_2, \ldots, \theta f_r/\theta\theta_2]$ stb. Az $(\theta\mathbf{f}/\theta\boldsymbol{\theta})$ $(s \times r)$ típusú mátrix.

Ha az elemszám elég nagy $\boldsymbol{\theta}$ helyett annak $\widehat{\boldsymbol{\theta}}_n$ becslését alkalmazva, az $\mathbf{f}(\widehat{\boldsymbol{\theta}}_n)$ aszimptotikus kovarianciamátrixa:

$$\left(\frac{\theta\mathbf{f}}{\theta\widehat{\boldsymbol{\theta}}_n}\right)' \mathrm{ACOV}(\widehat{\boldsymbol{\theta}}_n) \left(\frac{\theta\mathbf{f}}{\theta\boldsymbol{\theta}_n}\right).$$

Az eljárást egy egyszerű kauzális modellen mutatjuk be:

$$\eta_1 = \gamma_{11}\xi_1 + \rho_1$$

$$\eta_2 = \beta_{21}\eta_1 + \rho_2.$$

Feltételezzük, hogy ρ_1 és ρ_2 független egymástól és ξ_1-től, $E(\rho_i) = 0$, és a η_1, η_2 és ξ_1 az átlaguktól való eltéréseket tartalmazzák. A $\boldsymbol{\theta}$ paramétervektor a γ_1 és β_{21}, együtthatókat tartalmazza. Vizsgáljuk a ξ_1 indirekt hatását η_2-re: $\gamma_{11}\beta_{21}$, így az $\mathbf{f}(\boldsymbol{\theta})$ függvénynek egy eleme van $\gamma_{11}\beta_{21}$. Az $f_1(\boldsymbol{\theta})$ parciális deriváltja: $[\beta_{21}, \gamma_{11}]$, az $\widehat{\theta}_n$ asszimptotikus kovarianciamátrixa:

$$\text{ACOV}(\widehat{\theta}_n) = \begin{pmatrix} n^{-1}V_{11} & 0 \\ 0 & n^{-1}V_{22} \end{pmatrix}.$$

A mátrix diagonális elemei tartalmazzák a $\widehat{\gamma}_{11}$ és $\widehat{\beta}_{21}$ asszimptotikus varianciáit. A diagonálison kívüli elemek nullák, mivel a két együttható korrelálatlan egymással.

A delta-módszer eredménye szerint az indirekt $\widehat{\gamma}_{11}\widehat{\beta}_{21}$ hatás aszimptotikus varianciája:

$$n^{-1}[\beta_{21}^2 V_{11} + \gamma_{11}^2 V_{22}].$$

Ha a β_{21} vagy a γ_{11} értéke nulla, a delta-módszer nem alkalmazható. Egyébként a minta becsléseit behelyettesítve a fenti kifejezésbe, megkapjuk $\widehat{\gamma}_{11}\widehat{\beta}_{21}$ közvetett hatás aszimptotikus varianciájának egy becslését.

Láttuk, hogy az indirekt, teljes és specifikus hatások számítása és ezek varianciáinak becslése meglehetősen számításigényes.

A bootstrap eljárást alkalmazhatjuk a nem közvetlen hatások kiértékelésére is. Miután a modell több egyenletet tartalmaz, minden egyenletre ugyanazt a bootstrap mintát kell alkalmazni (összesen B-t) a becslések elvégzéséhez. Ezt az eljárást alkalmazva a közvetlen és a közvetett hatásokra, ezek varianciáit a bootstrap eloszlás alapján számíthatjuk.

Általában a bootstrap és a delta-módszer hasonló eredményeket ad, azonban a bootstrap standard hibái általában nagyobbak.

A bootstrap a delta-módszerrel ellentétben (ahol feltételezzük a normális eloszlást), képes a közvetett hatások eloszlásának aszimmetriáját megmutatni. Robert Stine és Kenneth Boolen (1988) közöl példát arra az esetre, amikor a közvetett hatás bootstrap eloszlása aszimmetrikus, a delta-módszer becslése pedig normális.

Robert Stine azonos modellt illesztett nagyobb mintára (n = 172), és akkor a fenti különbség meglehetősen kicsivé zsugorodott.

Jegyzet

Bootstrap-módszerek

Bootstrap-módszernek azokat az eljárásokat nevezzük, amelyek során statisztikák tulajdonságait becsüljük véletlen minták alapján, amelyeket egy megfigyelt mintából veszünk. A Bootstrap egy sajátos megközelítési módszer, amely során becsüljük a varianciákat, konfidenciaintervallumokat és a statisztikák más tulajdonságait. A Bootstrap egy módszer a statisztikák kiértékelésére. Azon a paradigmán alapszik, amin a klasszikus statisztikai becslési eljárás, amely során az alapsokaságból vett minta alapján következtetünk a sokaság eloszlására vagy az eloszlás valamely paraméterére. A Bootstrap-eloszlások varianciák, konfidenciaintervallumok azonban nem a sokaságból vett mintából, hanem a mintából vett mintából származnak. A Bootstrap-minta sajátos minta; minta, amit a megfigyelt adatokból (mintából) *visszatevéssel* veszünk. A Bootstrap-minta visszatevéses eljárással történő mintavétel a mintából. A Bootstrap standard hibák és

intervallumok rendszerint jobbak, mint azok, amelyek nem ellenőrzött feltevéseken alapulnak. A Bootstrap rugalmassága, hogy olyan statisztikák standard hibáit is megadja, amelyek számítása egyébként nagyon nehéz. Bootstrap a bonyolult matematikát helyettesíti a számítások számának növelésével. Ahelyett, hogy pl. a regressziós együtthatóknak egy vagy két becslését adná, a Bootstrap esetleg több százat számít. Ezeket a számításokat elvégezhetjük olyan programokkal, amelyek lehetővé teszik a makróprogramozást, mint pl. a SAS, azonban Bootstrap-becsléseket közvetlenül is kaphatunk az AXIS (An Experimental Interface for Statistics) program segítségével, amit Robert A. Stine (1991) dolgozott ki.

Tételezzük fel, hogy $\{X_1, X_2, \ldots, X_n\}$ egy véletlen minta, ahol a minta elemszáma n, ismeretlen eloszlással, az eloszlásfüggvényt jelölje F, ennek paraméterét pedig θ. A θ becslését jelölje $\widehat{\theta}$, és szeretnénk tudni ennek elméleti (populációbeli) varianciáját: VAR $(\widehat{\theta})$, és mintabeli eloszlását. Ha a populációból sokszor tudnánk mintát venni, akkor $\widehat{\theta}$ mintabeli ingadozását könnyen ki tudnánk számolni. Ez azonban nagyon ritkán lehetséges. A Bootstrap-megközelítés a mintán alapul. Jelölje az X valószínűségi változó $\{X = X_1, X_2, \ldots, X_n\}$ megfigyelt mintabeli értékeit $\{x_1, x_2, \ldots, x_n\}$, a minta empirikus eloszlását F_n. Az F_n empirikus eloszlásfüggvénye:

$$F_n(x) = \#(x_i \leq x)/n,$$

ahol $\#(x_i \leq x)$ azon megfigyelések száma, amelyekre az egyenlőtlenség teljesül.

Az empirikus eloszlásfüggvényről (F_n) feltételezzük, hogy jól közelíti a populáció eloszlását, F-et. Ha ez igaz, akkor nem kell a populációból sokszor mintát venni, hanem elég, ha a mintából veszünk elég sokszor mintát, és ennek alapján becsüljük a paramétert, és annak mintabeli varianciáját.

Jelölje $x^{*(b)} = \left(x_1^{*(b)}, x_2^{*(b)}, \ldots, x_n^{*(b)}\right)$ a Bootstrap-mintát, amelynek elemeit a $\{x_1, x_2, \ldots, x_n\}$ megfigyelt mintából véletlenszerűen, visszatevéssel választottuk. A Bootstrap-minta elemszáma n, megegyezik az eredeti minta elemszámával. A $*$ jelöli, hogy bootstrap mintáról van szó, a (b) pedig a Bootstrap-mintákat különbözteti meg egymástól ($b = 1,2, \ldots, B$). Mindegyik Bootstrap-minta becslést ad θ-ra, jelölje ezt $\widehat{\theta}^*$. A B számú Bootstrap-minta alapján meghatározhatjuk $\widehat{\theta}^*$ Bootstrap-eloszlását, valamint a VAR $(\widehat{\theta}^*)$ Bootstrap-becslését:

$$\text{var}(\widehat{\theta}^*) = \sum_{b=1}^{B} \left(\widehat{\theta}^{*(b)} - \overline{\widehat{\theta}}^*\right)^2 /(B-1),$$

ahol $\overline{\widehat{\theta}}^* = \sum_{b=1}^{B} \widehat{\theta}^{*(b)}/B$, a $\widehat{\theta}$ becslés átlaga a Bootstrap-mintákban (B a Bootstrap-minták száma). Hasonlóan, a $[\text{var}(\widehat{\theta}^*)]^{1/2}$ a $\widehat{\theta}$ becslés Bootstrap standard hibáját adja. Az $\widehat{\theta}^*$ Bootstrap-mintabeli eloszlása a $\widehat{\theta}$ mintabeli eloszlásának egy becslése. Efron (1987) a Bootstrap-minták számát (B) 100 körülinek javasolta a standard hibák becslése esetén.

A Bootstrap-becslést nemparametrikusnak nevezhetjük, mivel nem kell feltételeznünk az F eloszlás típusáról semmit. Csak azt kell feltételeznünk, hogy az empirikus eloszlásfüggvény F_n jól közelíti F-et, és hogy a $\widehat{\theta}^*$ Bootstrap-mintaeloszlása hasonlít a $\widehat{\theta}$ mintaeloszlásához.

7. fejezet

Diszkriminanciaelemzés

A diszkriminanciaelemzés célja egy adott osztályozásról eldönteni, hogy mely változók különítik el leginkább a csoportokat.

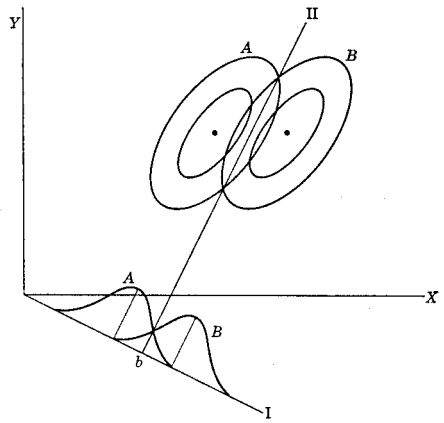
Feltételezzük, hogy a megfigyelési egységek m számú változó alapján g csoportba vannak besorolva, a csoportok elemei többváltozós normális eloszlásúak azonos kovarianciamátrixszal. Ebben az esetben a diszkriminanciafüggvény lineáris, a diszkriminanciafüggvények száma (a diszkriminancia-tér dimenziójára) felső korlát adható:

$$\min\{m, g-1\}$$

A diszkriminanciafüggvény segítségével vizsgálhatjuk a megfigyelt változók szerepét a csoportok különbözőségében, valamint egy újabb objektum csoportba sorolását.

7.1. A diszkriminanciaelemzés grafikus-modellje

Tekintsük a módszer geometriai interpretációját a legegyszerűbb esetben, amikor két csoportunk és két változónak van.



7.1. ábra. Grafikus-modell

A megfigyelt (mérési) változók olyan lineáris függvényét keressük, amely a csoportokat úgy vetíti le egy latens egyenesre, hogy a csoportok eloszlásai a legkisebb mértékben fedjék egymást.

Az elkülönülés annál jobb, minél kisebb az átfedés, ezt alapvetően két tényező befolyásolja

a) a csoportátlagok egymáshoz viszonyított elhelyezkedése és

b) az átlagvektorok körüli szóródás.

7.2. A diszkriminanciafüggvény meghatározása

A diszkriminanciafüggvény meghatározásának egyik eljárása a többváltozós szóráselemzésen alapul. A diszkriminanciafüggvényt a csoportok közötti és a csoporton belüli eltérések négyzetösszegei hányadosának maximalizálásával határozhatjuk meg. Tartalmazza az $\mathbf{X}$ mátrix a megfigyelt változóknak a teljes minta átlagától való eltérését.

$$\mathbf{T} = \mathbf{K} + \mathbf{B} \qquad (\text{ahol } \mathbf{T} = \mathbf{X}'\mathbf{X})$$

A diszkriminanciafüggvény a megfigyelt változók lineáris kombinációjaként állítható elő:

$$\mathbf{y} = \mathbf{X}\,\mathbf{c}$$

($\mathbf{c}'\,\mathbf{c} = 1$ normalizálási kikötéssel).

Az $\mathbf{y}$ tehát nem megfigyelt értékeket tartalmaz. Az $\mathbf{y}'\,\mathbf{y}$ négyzetösszeg két részre bontható:

$$\mathbf{y}'\,\mathbf{y} = \mathbf{c}'\,\mathbf{X}'\,\mathbf{X}\,\mathbf{c} = \mathbf{c}'\,\mathbf{T}\,\mathbf{c} = \mathbf{c}'\,(\mathbf{K} + \mathbf{B})\mathbf{c} = \mathbf{c}'\,\mathbf{K}\,\mathbf{c} + \mathbf{c}'\,\mathbf{B}\,\mathbf{c},$$

ahol az első tag a csoportok közötti eltéréseket, a második tag a csoporton belüli eltéréseket tartalmazza.

Az adott $\mathbf{K}$ és $\mathbf{B}$ mátrixok mellett keressük azt a diszkriminanciafüggvényt, amely az első tag arányát maximalizálja a másodikhoz képest. Feladatunk tehát megkeresni azokat a $\mathbf{c}$ együtthatókat, amelyek esetén a

$$\lambda = \frac{\mathbf{c}'\,\mathbf{K}\,\mathbf{c}}{\mathbf{c}'\,\mathbf{B}\,\mathbf{c}}$$

maximális értéket vesz fel. A λ maximalizálásánál kiindulhatunk a fenti kifejezés logaritmusából is:

$$\ln\lambda = \ln(\mathbf{c}'\,\mathbf{K}\,\mathbf{c}) - \ln(\mathbf{c}'\,\mathbf{B}\,\mathbf{c}).$$

A deriválás az alábbi eredményt adja:

$$\frac{\partial \ln\lambda}{\partial \mathbf{c}'} = \frac{2\mathbf{K}\,\mathbf{c}}{\mathbf{c}'\,\mathbf{K}\,\mathbf{c}} - \frac{2\mathbf{B}\,\mathbf{c}}{\mathbf{c}'\,\mathbf{B}\,\mathbf{c}} = \mathbf{0}.$$

Ebből a következő alapegyenlethez jutunk:

$$\mathbf{K}\,\mathbf{c} - \lambda\mathbf{B}\,\mathbf{c} = \mathbf{0}.$$

Feltéve, hogy a $\mathbf{B}^{-1}$ inverz létezik

$$(\mathbf{B}^{-1}\,\mathbf{K} - \lambda\,\mathbf{I})\mathbf{c} = \mathbf{0}.$$

Így a jól ismert sajátérték- és sajátvektor-problémához jutunk, vagyis $\mathbf{B}^{-1}\mathbf{K}$ mátrixnak λ a sajátértéke és $\mathbf{c}$ a sajátvektora. Ha $g-1$ kisebb mint m, akkor $\mathbf{K}$ rangja $g-1$, és a szorzatmátrix rangja nem lehet ennél nagyobb, mivel $\mathbf{B}$ rangja m, a szorzatmátrixnak $g-1$ különböző sajátértéke lesz.

A j-edik diszkriminanciafüggvény varianciája

$$\sigma_j = \mathbf{c}_j'\left(\frac{1}{n-1}\mathbf{T}\right)\mathbf{c}_j.$$

A diszkriminanciafüggvényt a szórásával standardizálva a diszkriminanciafaktorhoz jutunk

$$\mathbf{f}_j = \frac{1}{\sqrt{\sigma_j}}\mathbf{y}_j = \sigma_j^{-\frac{1}{2}}\left(\mathbf{X}\,\mathbf{c}_j\right).$$

Ha áttérünk standardizált változókra ($\mathbf{z}$):

$$\mathbf{X} = \mathbf{D}_{\text{diag}}^{\frac{1}{2}}\mathbf{Z},$$

ahol $\mathbf{D}_{\text{diag}}^{\frac{1}{2}}$ az $\frac{1}{n-1}\mathbf{T}$ mátrix diagonális elemeiből képzett mátrix (vagyis diagonálelemei az egyes változók szórásait adják).

Így

$$\mathbf{f}_j = \sigma_j^{-\frac{1}{2}}\mathbf{D}_{\text{diag}}^{\frac{1}{2}}\mathbf{Z}\,\mathbf{c}_j = \mathbf{Z}\left(\sigma_j^{-\frac{1}{2}}\mathbf{D}_{\text{diag}}^{\frac{1}{2}}\mathbf{c}_j\right) = \mathbf{Z}\,\mathbf{b}_j,$$

ahol $\mathbf{b}_j = \mathbf{D}_{\text{diag}}^{\frac{1}{2}}\mathbf{c}_j\sigma_j^{-\frac{1}{2}} = \mathbf{D}_{\text{diag}}^{\frac{1}{2}}\mathbf{a}_j$.

A $\mathbf{b}$ a standardizált diszkriminanciafüggvény együtthatóit tartalmazza. Az $\mathbf{s}_j = \mathbf{R}\,\mathbf{b}_j$ szorzat a j-edik diszkriminanciafaktor korrelációját adja az eredeti változókkal, ezzel faktorstruktúrához jutottunk.

Általában r ($r \leq m$) különböző sajátértéket kapunk a

$$\left(\mathbf{B}^{-1}\mathbf{K} - \mathbf{L}\,\mathbf{I}\right)\mathbf{C} = \mathbf{0}$$

egyenlet megoldásával. Ezek alapján állítjuk elő a diszkriminanciafüggvényeket úgy, hogy azok egymástól függetlenek legyenek. A diszkriminanciafaktor értelmezéséhez kiszámíthatjuk a faktorelemzésnél megismert faktorsúlymátrixot. Az $\mathbf{F} = \mathbf{Z}\,\mathbf{B}$ a diszkriminanciafaktorokat tartalmazza, ahol $\mathbf{B}$ mátrixot a diszkriminanciafaktorok együtthatómátrixának nevezzük ($m \times r$).

$$\mathbf{B} = \mathbf{D}_{\text{diag}}^{\frac{1}{2}}\mathbf{C}\left\langle\sigma^{-\frac{1}{2}}\right\rangle = \mathbf{D}_{\text{diag}}^{\frac{1}{2}}\mathbf{A}.$$

A diszkrimináns faktorstruktúra

$$\mathbf{S} = \mathbf{R}\,\mathbf{B}.$$

Az egyes diszkriminanciafüggvények ereje a sajátértékek segítségével mérhető: Például y_j esetén

$$\frac{\lambda_j}{\sum_{k=1}^{m}\lambda_k}.$$

A diszkrimináns hatás a Wilks-féle Λ-segítségével fejezhető ki, ezt két determináns hányadosaként kapjuk:

$$\Lambda = \frac{|\mathbf{B}|}{|\mathbf{T}|}.$$

Ugyanerre az eredményre jutunk, ha a $\mathbf{B}^{-1}\mathbf{K}$ sajátértékeiből számoljuk Λ értékét

$$\Lambda = \prod_{j=1}^{r}\frac{1}{1+\lambda_j}.$$

Ha az első k diszkriminanciafüggvényt szignifikánsnak találtuk, annak eldöntéséhez, hogy a maradék $(r-k)$ diszkriminanciafüggvények hatása szignifikánsnak tekinthető-e, χ^2-próba alkalmazható:

$$\chi^2 = -\left(n - \frac{m+g}{2} - 1\right)\ln\Lambda',$$

a szabadságfok $(m-k)(g-k-1)$,

és $\quad \Lambda' = \prod_{j=k+1}^{r}\frac{1}{1+\lambda_j}.$

A diszkriminanciafüggvények alterében a csoportok átlagai

$$m_{df_k} = \mathbf{A}'(\mathbf{m}_k - \mathbf{m}),$$

és a csoportok szórásai:

$$\mathbf{D}_{df_k} = \mathbf{A}'\,\mathbf{D}_k\mathbf{A}.$$

7.3. A diszkriminanciafüggvény alkalmazása objektumok csoportokba sorolására

Ebben a fejezetben többféle módszer segítségével vizsgáljuk, hogyan oldható meg újabb objektumok csoportba sorolása a diszkriminanciafüggvény alapján.

7.3.1. Általánosított távolság modell

Vizsgáljuk a diszkriminanciafüggvény alkalmazását objektumok csoportba sorolására. Tekintsünk g egymást át nem fedő csoportot, és vizsgáljuk, hogy egy adott objektum a g csoport melyikébe sorolható, melyikbe tartozik leginkább.

A g csoport átlagait jelöljék az

$$\overline{\mathbf{x}}_1, \overline{\mathbf{x}}_2, \ldots, \overline{\mathbf{x}}_g$$

vektorok, a csoportok kovarianciamátrixait pedig

$$\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_g.$$

A vizsgált objektum mérési adatait az $\mathbf{x}$ vektor tartalmazza. Az $\mathbf{x}$ objektumot a j-edik csoporthoz soroljuk, ha a csoportok átlagvektoraitól mért általánosított távolsága a j-edik csoportra a legkisebb.

Az általánosított Mahalanobis-féle távolság:

$$d(\mathbf{x}, \overline{\mathbf{x}}_j) = [(\mathbf{x} - \overline{\mathbf{x}}_j)'\mathbf{S}_j^{-1}(\mathbf{x} - \overline{\mathbf{x}}_j)]^{1/2}.$$

Az általánosított távolság segítségével így az objektumok besorolására egy determinisztikus szabály adható.

Ha feltételezzük, hogy a csoportok azonos szórású normális eloszlásúak, akkor az általánosított távolság felhasználásával valószínűségi szabályt is adhatunk. A nullhipotézis:

$$H_0\colon E(\overline{\mathbf{x}}_j) = \mathbf{x}$$

és az alternatív hipotézis

$$H_1\colon E(\overline{\mathbf{x}}_j) \neq \mathbf{x}.$$

A próbafüggvény

$$F_j = (n_j - m)n_j/[m(n_j - 1)]d^2(\mathbf{x}, \overline{\mathbf{x}}_j)$$

$(m, n_j - m)$ szabadságfokú F-eloszlást követ.

Ha a csoportok elemszámait a teljes mintához viszonyítjuk, így a csoportok nagyságait a $P(1)$, $P(2)$, ..., $P(g)$ valószínűségek ($\sum P(i) = 1$) fejezik ki, és annak valószí-

nűsége, hogy a véletlenszerűen választott **x** objektum a j-edik csoportba tartozik:

$$P(\mathbf{x}|j) = \int f(F|H_0, v_1, v_2)dF \qquad (F = F_j, \ldots, \infty).$$

A hibás klasszifikáció valószínűsége a következő valószínűség komplementere:

$$P(j|\mathbf{x}) = \frac{P(j)P(\mathbf{x}|j)}{\sum_{i=1}^{k} P(i)P(\mathbf{x}|i)},$$

ahol a $P(j|\mathbf{x})$ *a posteriori* valószínűség, annak valószínűsége, hogy a j-edik csoportot jelöltük ki, feltéve, hogy **x** következett be.

PÉLDA

Legyen két csoportunk, 8, illetve 9 foglalkozással. A négy változó megfigyelt értékeit tartalmazza a következő táblázat:

1. csoport $n_1 = 8$
(1. klaszter)

Presztízs	x_1	45	40	16	28	39	38	25	31
Kereset	x_2	3268	4801	4514	4785	4513	3449	4338	3768
Isk. végz.	x_3	9,8	9,4	9,5	9,6	8,5	9,0	9,4	9,0
Munkakör	x_4	4,0	3,1	4,7	3,8	3,4	4,1	4,2	3,9

2. csoport $n_2 = 9$
(3. klaszter)

Presztizs	x_1	32	48	63	42	39	57	39	54	49
Kereset	x_2	6397	4687	3905	4649	3040	3790	5864	2995	3985
Isk. végz.	x_3	8,9	11,0	11,4	9,5	12,0	14,5	11,7	12,0	12,5
Munkakör.	x_4	3,5	4,2	3,7	4,6	6,0	3,5	4,1	5,0	5,2

Az átlagvektorok és kovarianciamátrixok:

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} 32{,}75 \\ 4179{,}5 \\ 9{,}27 \\ 3{,}9 \end{bmatrix}, \quad \mathbf{S}_1 = \begin{bmatrix} 90{,}79 & -2485{,}4 & -0{,}68 & -3{,}13 \\ -2485{,}4 & 362131{,}7 & -7{,}63 & -92{,}37 \\ -0{,}68 & -7{,}63 & 0{,}17 & 0{,}07 \\ -3{,}13 & -92{,}37 & 0{,}07 & 0{,}24 \end{bmatrix},$$

$$\bar{\mathbf{x}}_2 = \begin{bmatrix} 47 \\ 4368 \\ 11{,}5 \\ 4{,}42 \end{bmatrix}, \quad \mathbf{S}_2 = \begin{bmatrix} 98{,}5 & -7053{,}75 & 9{,}6 & -1{,}56 \\ -7053{,}75 & 1359929{,}25 & -1139{,}86 & -586{,}65 \\ 9{,}6 & 1139{,}86 & 0{,}156 & 0{,}73 \\ -1{,}56 & -586{,}65 & 0{,}156 & 0{,}73 \end{bmatrix}.$$

Vizsgáljuk meg, hogy az $\mathbf{x} = [48,\ 3554,\ 7,\ 2]'$ adatokkal jelzett egyed melyik csoporthoz tartozik.

Az általánosított távolság:

$$d(\mathbf{x}, \bar{\mathbf{x}}_1) = 8{,}414,$$

$$d(\mathbf{x}, \bar{\mathbf{x}}_2) = 96{,}09.$$

A megfelelő F értékek:

$$F_1 = [(8-4)8/(4(7))]8{,}414^2 = 80{,}91,$$

$$F_2 = [(9-4)9/(4(8))]96{,}09^2 = 12984{,}3$$

Mivel $d(\mathbf{x}, \overline{\mathbf{x}}_1) < d(\mathbf{x}, \overline{\mathbf{x}}_2)$, a vizsgált egyedet az 1. csoporthoz soroljuk.

Ez az állítás 5%-os szignifikanciaszinten nem erősíthető meg, mivel

$$(F_1 = 80{,}91) > (F_{0{,}05,\ 4,\ 4} = 6{,}39).$$

7.3.2. Diszkriminancia-modell

Alkalmazzuk az x objektum csoportba sorolásához a diszkriminanciafüggvényt.

Legyen adott a két (A és B) csoport átlagvektora és kovarianciamátrixa

Legyen:

$$\overline{\mathbf{x}}_A = \begin{bmatrix} \overline{x}_{1A} \\ \vdots \\ \overline{x}_{mA} \end{bmatrix},$$

$$\mathbf{S}_A = \begin{bmatrix} S_{11A} & S_{12A} & \dots & S_{1mA} \\ S_{21A} & S_{22A} & \dots & S_{2mA} \\ \vdots & \vdots & & \vdots \\ S_{m1A} & S_{m2A} & \dots & S_{mmA} \end{bmatrix}.$$

Hasonlóan írható fel $\overline{\mathbf{x}}_B$ és $\mathbf{S}_B$ is.

Feltételezzük, hogy a két csoport többváltozós normális eloszlású azonos kovarianciamátrixszal rendelkezik.

A diszkriminanciafüggvény együtthatóit a következőképpen számítjuk:

$$\mathbf{a} = \mathbf{S}_{AB}^{-1}(\overline{\mathbf{x}}_A - \overline{\mathbf{x}}_B),$$

ahol $\mathbf{S}_{AB}^{-1}$ a két csoport közötti kovarianciamátrix inverze, amelynek általános eleme:

$$S_{hi} = [(n_A - 1)S_{hiA} + (n_B - 1)S_{hiB}]/[n_A + n_B - 2],$$

ahol S_{hiA} és S_{hiB} a h-adik és az i-edik változó közötti kovariancia az A és B csoportban.

Egy adott $\mathbf{x}$ objektum diszkriminanciaértékét a B csoporthoz viszonyítva a következőképpen számítjuk:

$$d(\mathbf{x}, \overline{\mathbf{x}}_B) = \sum_i a_i(x_i - \overline{x}_{iB}) \qquad (i = 1, \dots, m).$$

Az $\mathbf{x}$ objektumot abba a csoportba soroljuk, amelyre a $d(\mathbf{x}, \overline{\mathbf{x}}_B)$ a legkisebb.

Ha a hibás klasszifikáció „költsége” különbözik a két csoportnál, a következő döntési szabályt alkalmazzuk: ha

$$Q_A > \ln\{P(B)L(A|B)/[P(A)L(B|A)]\},$$

az $\mathbf{x}$ objektumot az A csoporthoz soroljuk.

A kifejezések tartalma:

$$Q_A = [\mathbf{x} - 0{,}5(\overline{\mathbf{x}}_A + \overline{\mathbf{x}}_B)]'\mathbf{a},$$

$$P(A) = n_A/(n_A + n_B),$$

$$P(B) = n_B/(n_A + n_B).$$

Az $L(A|B)$ és $L(B|A)$ a „költség”-függvények. Ha ezek egyenlőek és $n_A = n_B$, az $\mathbf{x}$-et az A csoporthoz soroljuk, ha Q_A nagyobb, mint nulla.

PÉLDA

Legyen adott a két csoport közötti kovarianciamátrix és az átlagvektorok:

$$\mathbf{S}_{12} = \begin{bmatrix} 94{,}902 & -4921{,}853 & 4{,}803 & -2{,}293 \\ -4921{,}853 & 894290{,}393 & -611{,}486 & -355{,}986 \\ 4{,}803 & -611{,}486 & 1{,}517 & 0{,}116 \\ -2{,}293 & -355{,}986 & 0{,}116 & 0{,}501 \end{bmatrix};$$

$$\overline{\mathbf{x}}_1 = \begin{bmatrix} 32{,}75 \\ 4179{,}5 \\ 9{,}27 \\ 3{,}9 \end{bmatrix}; \qquad \overline{\mathbf{x}}_2 = \begin{bmatrix} 47{,}0 \\ 4368{,}0 \\ 11{,}5 \\ 4{,}4 \end{bmatrix}.$$

A diszkriminanciafüggvény együtthatói:

$$a = \begin{bmatrix} -1{,}283 \\ -0{,}016 \\ -2{,}521 \\ -17{,}712 \end{bmatrix}.$$

A diszkriminanciaértékek a két csoporthoz viszonyítva:

$$d(\mathbf{x}, \overline{\mathbf{x}}_1) = -1{,}283(48 - 32{,}75) - 0{,}016(3554 - 4179{,}5) - 2{,}521(7 - 9{,}27) - 17{,}712(2 - 3{,}9) = 29{,}866,$$

$$d(\mathbf{x}, \overline{\mathbf{x}}_2) = 65{,}987.$$

A fenti értékek szerint az $\mathbf{x} = [48,\ 3554,\ 7,\ 2]$ adatokkal jellemzett egyed az 1-es csoporthoz sorolható.

A Q értéke függ a hibás klasszifikáció költségétől.

Például ha

$$L(1|2) = L(2|1) = 1,$$

$$P(1) = 8/17 = 0{,}47 \quad \text{és}$$

$$P(2) = 0{,}53,$$

akkor

$$\ln[P(1)L(2|1)]/[P(2)L(1|2)] = \ln[0{,}47/0{,}53] = \ln[0{,}88] = -0{,}12$$

és

$$Q_1 = \left[\begin{bmatrix} 48 \\ 3554 \\ 7 \\ 2 \end{bmatrix} - 0{,}5 \left\{ \begin{bmatrix} 32{,}75 \\ 4179{,}5 \\ 9{,}27 \\ 3{,}9 \end{bmatrix} + \begin{bmatrix} 47 \\ 4368 \\ 11{,}5 \\ 4{,}42 \end{bmatrix} \right\} \right]' \begin{bmatrix} -1{,}283 \\ -0{,}016 \\ -2{,}521 \\ -17{,}712 \end{bmatrix} = 47{,}926.$$

Mivel $Q_1 > -0{,}12$, a vizsgált objektumot az 1. csoporthoz soroljuk.

7.3.3. I-divergencia

Tegyük fel, hogy g csoportunk van, és a csoportokat minőségi változókkal jellemezzük. A j-edik csoportot a valószínűségeloszlással írjuk le:

$$P_j = [p_{1j}, p_{2j}, \ldots, p_{sj}],$$

ahol s a kvalitatív változó kategóriáinak a száma,

$$p_{ij} = f_{ij}/f_{0j} \qquad \text{minden} \qquad i = 1,2, \ldots, s\text{-re}.$$

Azt vizsgáljuk, hogy az

$$F = [f_1, f_2, \ldots, f_s]$$

eloszlással adott objektum melyik csoportba sorolható. Az F divergenciája a j-edik csoporttól

$$2I(F; F_j) = 2 \sum_{i=1}^{s} f_i \ln[(f_i/f_0)/p_{ij}].$$

A kérdéses objektumot ahhoz a csoporthoz soroljuk, amelyikre minimális divergenciát ad. Ezt a besorolást akkor fogadjuk el, ha a következő egyenlőtlenség teljesül:

$$2I(F; F_j) < \chi^2_{\alpha,s-1},$$

ahol $\chi^2_{\alpha,s-1}$ a khi négyzet eloszlás táblázatából vehető $s - 1$ szabadságfokkal.

PÉLDA

Tegyük fel, hogy három csoportot a következő valószínűségeloszlással jellemzünk:

$$P_1 = [0,23 \quad 0,15 \quad 0,40 \quad 0,22],$$

$$P_2 = [0,54 \quad 0,30 \quad 0,14 \quad 0,02],$$

$$P_3 = [0,12 \quad 0,23 \quad 0,19 \quad 0,46].$$

Megvizsgáljuk, hogy az

$$F = [20 \quad 16 \quad 24 \quad 15]$$

gyakorisággal jellemzett objektum melyik csoporthoz sorolható.
A számítások:

$$2I(F; F_1) = 2[20 \ln [20/75)/0,23] + 16 \ln [(16/75)/0,15] +$$
$$+ 24 \ln [(24/75)/0,40] + 15 \ln [(15/75)/0,22]] = 3,6131,$$

$$2I(F; F_2) = 69,6258,$$

$$2I(F; F_3) = 29,5682.$$

Az F-eloszlással jellemzett objektumot az első csoportba soroljuk. A besorolást elfogadjuk, mivel

$$(2I(F; F_1)) < (\chi^2_{0,05,\ 3} = 7,815).$$

7.3.4. Bayes-féle elemzés

Tekintsünk egy objektumot, amelyet p dichotom változó eloszlásával jellemzünk:

$$F = \begin{bmatrix} f_1 & n - f_1 \\ f_2 & n - f_2 \\ \vdots & \vdots \\ f_p & n - f_p \end{bmatrix}.$$

Az f_i azon esetek száma, amelyek rendelkeznek az i-edik változó tulajdonságával, az $n - f_i$ azon esetek száma, amelyek nem rendelkeznek.

Feltételezzük, hogy g csoportunk van. A vizsgált objektumot ahhoz a csoporthoz soroljuk, amelyhez a legnagyobb *a priori* valószínűséggel tartozik.

Az *a priori* valószínűség:

$$P(j|F) = P(j)P(F|j) / \sum_{i=1}^{g} [P(i)P(F|i)].$$

Ez annak a valószínűsége, hogy a j-edik csoportot választjuk, amikor az F-eloszlást figyeltük meg.

Az *a priori* valószínűséget minden csoportra kiszámítjuk, és az F-eloszlással jellemzett objektumot ahhoz a csoporthoz soroljuk, amelyre a legnagyobb az *a priori* valószínűség.

$$P(j) = n_j / n$$

és

$$P(F|j) = \prod_{i=1}^{p} P(F_i|j),$$

ahol

$$P(F_i|j) = p_{ij}^{f_i}(1 - p_{ij})^{n-f_i} n!/[f_i!(n - f_i)!].$$

A besorolást elfogadjuk, ha a következő feltétel teljesül:

$$P(j|F) \geq \alpha.$$

PÉLDA

A táblázatban négy dichotom változó előfordulásának a valószínűsége található három csoportra.

Csoportok	Változók				
j	1	2	3	4	n_j
1	0,6	0,5	0,3	0,1	200
2	0,2	0,7	0,7	0,8	150
3	0,9	0,2	0,1	0,6	100

Tételezzük fel, hogy egy objektumnak négy dichotom változóra vontakozóan rendelkezünk az alábbi gyakoriság-eloszlásával:

$$F = \begin{bmatrix} f_1 & n - f_1 \\ f_2 & n - f_2 \\ f_3 & n - f_3 \\ f_4 & n - f_4 \end{bmatrix} = \begin{bmatrix} 7 & 33 \\ 28 & 12 \\ 30 & 10 \\ 39 & 1 \end{bmatrix}.$$

Az $f_1 = 7$ azt jelenti, hogy az első változó 7 esetre jellemző és 33-ra nem.

A számítás a 2. csoportra:

$$F_1 = [f_1 \quad n - f_1] = [7 \quad 33],$$

$$P_{12} = [p_{12} \quad 1 - p_{12}] = [0{,}2 \quad 0{,}8],$$

$$P(F_1|2) = 0{,}2^7(0{,}8)^{33}40!/[7! \quad 33!] = 0{,}15125,$$

$$P(F_2|2) = 0{,}13657,$$

$$P(F_3|2) = 0{,}11282,$$

$$P(F_4|2) = 0{,}00133,$$

$$P(F|2) = 0{,}15125(0{,}13657)(0{,}11282)(0{,}00133) = 3{,}09948 \cdot 10^{-6}.$$

Ez utóbbi annak a valószínűségét adja, hogy F a második csoporthoz tartozik.
A fenti számításokhoz hasonlóan:

$$P(F|1) = 3{,}347278(10^{-56}),$$

$$P(F|3) = 0{,}68231(10^{-67}).$$

Az első csoport *a posteriori* valószínűsége:

$$P(1|F) = 0{,}4444(3{,}47278(10^{-56}))/[0{,}4444(3{,}47278(10^{-56}))+$$
$$+ 0{,}3333(3{,}09948(10^{-6})) + 0{,}2222(9{,}68231(10^{-67}))] =$$
$$= 1{,}49392 \cdot 10^{-50},$$

ahol $P(j) = 200/450 = 0{,}4444$.
Az *a posteriori* valószínűségek a második és harmadik csoportra:

$$P(2|F) = 0{,}999,$$

$$P(3|F) = 2{,}08257(10^{-61}).$$

Az eredmények azt mutatják, hogy a kérdéses objektum a 2. csoporthoz tartozik.

7.4. Kanonikus diszkriminancia-faktorelemzés

Bartell (1938) érdekes összefüggéseket mutatott meg a diszkriminanciaelemzés és a kanonikus korrelációelemzés között.

Ha az egyes csoportokba tartozást változókkal fejezzük ki úgy, hogy az i-dik változó 1 értéket vesz fel, ha az eset az i-edik csoportba esik, és 0-t különben. Ezen bináris változók és a diszkriminancia-függvény közötti maximális korrelációt hívjuk kanonikus korrelációnak, amit a következőképpen számíthatunk ki:

$$r_{cj} = \frac{\lambda_j}{1 + \lambda_j}$$

Feltételeztük, hogy az $\mathbf{x}$ változót a teljes minta átlagából való eltéréseivel mérjük.

Ennek alapján, ha x_{ijk} jelöli a k-adik megfigyelést az i-edik csoportban a j-edik változó szerint:

$$x_{ijk} = m_{ij} + u_{ijk},$$

ahol m_{ij} a j-edik változónak az átlaga az i-edik csoportban, u_{ijk} pedig a hibakomponens.

A teljes **X** mátrix felbontható eszerint két részre:

X = **M**+**U** és ez a kanonikus faktorelemzés kiinduló egyenletével egyenlő, ahol **M** sorai csoporton belül azonosak. A teljes eltérések négyzetösszeg-mátrixa felbontható ennek alpján:

$$\mathbf{T} = \mathbf{X}'\mathbf{X} = (\mathbf{M} + \mathbf{U})'(\mathbf{M} + \mathbf{U}) = \mathbf{M}'\mathbf{M} + \mathbf{U}'\mathbf{U} = \mathbf{K} + \mathbf{B},$$

feltételezve, hogy $\mathbf{M}'\mathbf{U} = \mathbf{0}$ vagyis a csoportos átlagok korrelálatlanok a hibakomponenssel.

A kanonikus feladat szerint **X** és **M** olyan lineáris kombinációit (**v** és **w**) keressük, amely maximális korrelációt biztosít.

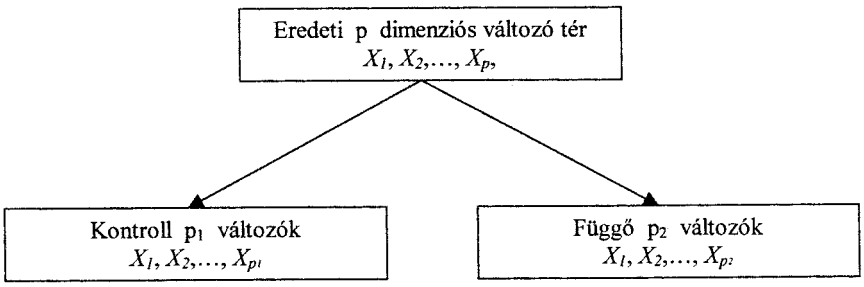
Megmutatható, hogy megoldva ezt a feladatot a diszkriminanciaelemzés $\mathbf{K}\,\mathbf{c} - \lambda\mathbf{B}\,\mathbf{c} = \mathbf{0}$ alapegyenletéhez jutunk, ami viszont egyenlő a kanonikus faktorelemzés $\mathbf{R}^*\,\mathbf{c} = \alpha\mathbf{U}^2\mathbf{c}$ egyenletével.

Ez indokolja, hogy ezt a módszert kanonikus diszkriminancia-faktorelemzésnek nevezzük.

7.5. Többszörös kovarianciaelemzés

Induljunk ki a többváltozós szóráselemzés alaphelyzetéből. Az egyik oldalon adott a legalább intervallum szintű skálákon mért n dimenziós vektorváltozó, a másik oldalon pedig egy szempont, egy nominális szintű, kategóriaképző ismérv.

Mindenekelőtt vegyük tüzetes vizsgálat alá a vektorváltozó n komponensét, melyeknek n dimenziós állapotterében megfigyelési egységeink elhelyezkednek. Osszuk fel a p változót két csoportra:



ahol $p = p_1 + p_2$

A két csoportra osztást:

1. a szaktudomány szempontjai,
2. esetleg meglévő előző vizsgálatok eredményei,
3. hipotézisek felhasználásával

végezhetjük (esetleg több változatban is elkészíthetjük őket). Ezután a második p_2 elemű csoport minden változóját az első csoport p_1 változójának felhasználásával, lineáris regressziós függvény segítségével előállítjuk:

$$\widehat{X}_{i_r} = \widehat{X}_{i_r}(X_{j_1}, X_{j_2}, \ldots, X_{j_{p_1}})$$

ezután elvégezzük a „kontrollváltozók” hatásának kiszűrését:

$$\widehat{X}^*_{i_r} = X_{i_r} - \widehat{X}_{i_r}(X_{j_1}, X_{j_2}, \ldots, X_{j_{p_1}})$$

megkapva ezúton az $\widehat{X}^*_{i_r}$ „reziduálisokat”.

A vázolt vizsgálatok után a p_2 db csillagos változóra:

$$X^*_{i_1}; X^*_{i_2}; \ldots; X^*_{i_{p_2}}$$

többváltozós szóráselemzést végzünk.

Az eljárás lényegét abban rögzíthetjük, hogy a leírt módon megvizsgálható, a változók közötti belső kapcsolatok nem módosítják-e nem kívánatos mértékben a szóráselemzés eredményét.

Ha azt szeretnénk tudni, hogy valamely kritérium szerinti csoportok különböznek-e bizonyos változókban, akkor a többváltozós szóráselemzés módszerét használhatjuk. Abban az esetben, amikor egy változóhalmaz hatását keressük a csoportok közötti eltérés magyarázásában egy kontroll-változóhalmaz információjának, megkülönböztető hatásának kiszűrése után, a többváltozós kovarianciaelemzés módszerét használhatjuk.

A kontrollváltozók (p_1) regressziós becslését kiszűrve a függő változóhalmazból (p_2) a függő változó reziduumának a megkülönböztető hatását vizsgáljuk a csoportokban. A reziduális mátrixot a többszörös parciális korrelációelemzés segítségével számítottuk ki:

$$\widetilde{\mathbf{R}}_{22} = \mathbf{R}_{22} - \mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{R}_{12} = \mathbf{R}_{22} - \widehat{\mathbf{R}}_{22},$$

ahol $\widetilde{\mathbf{R}}_{22}$ a reziduálisok kovarianciamátrixa. A többváltozós szóráselemzésnél használt jelöléssel $\mathbf{R}_{2,1}$-gyel is jelölhetjük.

A többváltozós szóráselemzés fogalmaira átültetve a reziduális mátrix gondolatát, két mátrixhoz is juthatunk.

Egyrészt a teljes minta eltérés-négyzetének reziduális mátrixához. A

$$T = \left[\begin{array}{c|c} \mathbf{T}_{11} & \mathbf{T}_{12} \\ \hline \mathbf{T}_{21} & \mathbf{T}_{22} \end{array}\right]$$ particionált mátrixból kiindulva,

a teljes reziduális mátrix:

$$\mathbf{T}_{2,1} = \mathbf{T}_{22} - \mathbf{T}_{21}\mathbf{T}_{11}^{-1}\mathbf{T}_{12},$$

a csoportokon belüli eltérések négyzetösszegeinek reziduális mátrixa:

$$\mathbf{B} = \left[\begin{array}{c|c} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \hline \mathbf{B}_{21} & \mathbf{B}_{22} \end{array}\right]$$ particionált formából kiindulva

$$\mathbf{B}_{2,1} = \mathbf{B}_{22} - \mathbf{B}_{21}\mathbf{B}_{11}^{-1}\mathbf{B}_{12}.$$

A Wilks-lambdakritérium:

$$\Lambda = \frac{|\mathbf{B}_{2,1}|}{|\mathbf{T}_{2,1}|}.$$

Az F-próba megegyezik a többváltozós szóráselemzésnél leírtakéval.

Az egyváltozós F-hányadosok

$$F_j = \frac{\left(t_{2,1\ jj} - b_{2,1\ jj}\right)/f_1}{b_{2,1\ jj}/f_2},$$

ahol $f_1 = g - 1$ és $f_2 = n - g - p_1$ a két szabadságfok.

7.6. Faktoriális diszkriminanciaelemzés

A többváltozós szóráselemzés két vagy több csoportba osztott mintában elemzi a csoportosító ismérv és egy vagy több függő változó kapcsolatát. Így választ ad arra a kérdésre, hogy a függő változók terében a csoportok különbözőnek tekinthetők-e, és mely változók különbözősége mutat szignifikáns eltérést. A diszkriminanciaelemzés ezen túl azt is megmutatja, hogy a csoportok lehetséges elkülönítésében az egyes változók milyen szerepet játszanak. A többváltozós szóráselemzést kiterjeszthetjük két csoportképző ismérv és a függő változók vizsgálatára. A többváltozós szóráselemzés ilyen irányú továbbfejlesztését nevezzük *faktoriális diszkriminanciaelemzésnek*.

A teljes mintát tehát két ismérv szerint csoportosítjuk:

	k
j	$\mathbf{x}_{jki}$

$$j = 1,2,\ldots,r$$
$$k = 1,2,\ldots,c$$
$$i = 1,2,\ldots,n$$

Minden j, k cellában (csoportban) rendelkezésünkre áll a p számú függő változó megfigyelése n esetre. A megfigyelések száma (n) tehát minden cellában azonos. Vizsgáljuk a sorok és az oszlopok között lévő különbségeket, vagyis a sorban található csoportosítási ismérv hatását (sorhatás) és az oszlopváltozó hatását (oszlophatás), valamint a két csoportosító ismérv interakcióját a függő változók halmazára, és vizsgáljuk a függő változóknak a sorok és az oszlopok különbözőségében játszott szerepét.

7.6.1. A módszer leírása

A sorhatás, az oszlophatás és az interakciók meghatározásához a minta teljes eltéréseit írjuk fel additív formában. A teljes minta átlagtól való eltéréseinek keresztszorzat- és négyzetösszeg-mátrixa:

$$\mathbf{T} = \sum_{j=1}^{r}\sum_{k=1}^{c}\sum_{i=1}^{n}(\mathbf{x}_{jki} - \mathbf{m})(\mathbf{x}_{jki} - \mathbf{m})',$$

ahol $\mathbf{m}$ a függő változók átlagvektora a teljes mintában.

A cellákon belüli eltérések keresztszorzat- és négyzetösszeg-mátrixa (hibamátrix):

$$\mathbf{B} = \sum_{j=1}^{r}\sum_{k=1}^{c}\sum_{i=1}^{n}(\mathbf{x}_{jki} - \mathbf{m}_{jk})(\mathbf{x}_{jki} - \mathbf{m}_{jk})',$$

ahol $\mathbf{m}_{jk}$ a j,k cella átlagvektora.

A sorok átlagvektorának a teljes átlagvektortól való eltérésének szorzatösszege, a *sorhatás*:

$$\mathbf{K}_r = nc \sum_{j=1}^{r} (\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})',$$

ahol $\mathbf{m}_j$ a j-edik sor átlagvektora.

Az oszlopok átlagai és a teljes átlag (vektor) eltérései keresztszorzat- és négyzetösszeg-mátrixa (*oszlophatás*):

$$\mathbf{K}_c = nr \sum_{k=1}^{c} (\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})'$$

A sor- és oszlopváltozó egymásra gyakorolt hatása (*interakció*):

$$\mathbf{K}_i = \mathbf{T} - (\mathbf{K}_r + \mathbf{K}_c + \mathbf{B})$$

vagy közvetlen számolással:

$$\mathbf{K}_i = n \sum_{j=1}^{r} \sum_{k=1}^{c} (\mathbf{m}_{jk} - \mathbf{m}_j - \mathbf{m}_k + \mathbf{m})(\mathbf{m}_{jk} - \mathbf{m}_j - \mathbf{m}_k + \mathbf{m})',$$

A teljes eltérés keresztszorzat- és négyzetösszeg-mátrixának teljes felbontása, amely a módszer alapegyenlete:

$$\mathbf{T} = \mathbf{K}_r + \mathbf{K}_c + \mathbf{K}_i + \mathbf{B}$$

A felbontáshoz tartozó mátrixok szabadságfokai:

$$\begin{aligned} &\mathbf{T} : rcn - 1 \\ &\mathbf{K}_r : r - 1 \\ &\mathbf{K}_c : c - 1 \\ &\mathbf{K}_i : (r-1)(c-1) \\ &\mathbf{B} : rc(n-1) \end{aligned}$$

A felbontásból látható, hogy a teljes eltérések négy részre bonthatók fel: a sorok eltérései, az oszlopok eltérései, a sor és oszlop változó interakciója, és végül a cellákon belüli eltérések, amit hibatagnak nevezünk. Ennek megfelelően háromféle hatás szignifikanciáját kell ellenőrizni. A többváltozós szóráselemzésnél definiált Wilks-féle lambdát használhatjuk a hipotézisvizsgálat elvégzéséhez.

A sorhatás hipotéziséhez:

$$\Lambda_r = \frac{|\mathbf{B}|}{|\mathbf{K}_r + \mathbf{B}|},$$

az oszlophatás hipotéziséhez:

$$\Lambda_c = \frac{|\mathbf{B}|}{|\mathbf{K}_c + \mathbf{B}|},$$

az interakció hipotéziséhez:

$$\Lambda_i = \frac{|\mathbf{B}|}{|\mathbf{K}_i + \mathbf{B}|}.$$

A Λ-k alapján a megfelelő F értéket a következőképpen számíthatjuk:

$$F_{N_2}^{N_1} = \left(\frac{1 - \Lambda^{1/s}}{\Lambda^{1/s}} \right) \frac{N_2}{N_1},$$

ahol

$$s = \sqrt{\frac{p^2(rc-1)^2 - 4}{p^2 + (rc-1)^2 - 5}},$$

$$N_1 = p(rc-1)$$

$$N_2 = s\left((n-1) - \frac{p+rc}{2}\right) - \frac{p(rc-1)-2}{2}$$

A változókra külön is számolható F hányados a $\mathbf{K}_r$, $\mathbf{K}_c$, $\mathbf{K}_i$ és $\mathbf{B}$ mátrixok elemeiből. Pl. a sorhatás hipotézisének ellenőrzése egy változóra:

$$F^{r-1}_{(r-1)(c-1)} = \frac{(k_{jj}/(r-1))}{(b_{jj}/(r-1)(c-1))}$$

A $\mathbf{K}_r$, $\mathbf{K}_c$, $\mathbf{K}_i$ mátrixok felhasználásával a változóknak a sorok és oszlopok elkülönülésében játszott szerepére a külön-külön végzett diszkriminanciaelemzésekkel kaphatunk választ. Így pl. a változók szerepét, fontosságát a sorok különbségeiben a $(\mathbf{B}^{-1}\mathbf{K}_r)$ mátrix sajátvektora alapján határozhatjuk meg.

7.6.2. Példa a faktoriális diszkriminanciaelemzésre

A NIM Távlati Tervezési és Beruházási Főosztálya az MKKE Matematikai és Számítástudományi Intézetével együttműködve kidolgozta a várható beruházási költség előrejelzésének eljárását. Az e kutatásban felhasznált alapadatokra támaszkodva lehetőség nyílt további vizsgálatok elvégzésére is, ezek egyikét foglaljuk össze a következőkben. A rendelkezésünkre álló, hitellel finanszírozott beruházásokat két szempont szerint csoportosítottuk. Az egyik szempont szerint ágazatonként, a másik szerint területi bontásban választottuk szét az adatokat, majd faktoriális diszkriminanciaelemzés segítségével vizsgáltuk, hogy e két ismérv szerinti csoportosításnak milyen hatása van a beruházások hatékonyságára, valamint a beruházásra fordított összegre. A hatékonyságot a 100 Ft eszközre jutó nyereséggel fejezzük ki. Számításainkban 5 ágazatot és 3 regionális területet vettünk fel, az egyes cellákban 19 adat szerepelt. Ez összesen 285 adatot jelent. (A módszer megkívánja, hogy mindegyik csoportban ugyanannyi adat szerepeljen.) A vizsgálatba vont ágazatok és régiók a következők:

ágazatok: nehézipar és gépipar; építő- és építőanyag-ipar; könnyűipar és egyéb ipar; mezőgazdaság; termelő szolgáltatás;

régiók: központi körzet; Kelet-Magyarország; Nyugat-Magyarország.

Számítási eredmények. Két komponenst tartalmazó függő vektorváltozót vettünk figyelembe. E komponensek: a beruházási költség (y_1) és a hatékonyság (y_2):

	y_1	y_2
teljes átlag	54,414	21,668
standard eltérés	102,263	14,973

Sorhatás. A sorhatás szerepe most az ágazati hatás nyomon követése. Először a várható értékek eltérését vizsgáljuk meg F-próbával. A szabadságfokok: 4 és 270. Mindkét függő változót (y_1-et és y_2-t) külön-külön tekintjük, y_1 esetében

a belső szóráshoz tartozó négyzetösszeg 34 556,566,

a külső szóráshoz tartozó négyzetösszeg 9 485,589,

a kapott empirikus érték: $F_{\text{emp}} = 3{,}64$, az elméleti érték: $F_{\text{elm}} = 3{,}4$; y_2 esetében

a belső szóráshoz tartozó négyzetösszeg 1 553,467,

a külső szóráshoz tartozó négyzetösszeg 255,936,

az empirikus érték: $F_{emp} = 7{,}54$, az elméleti érték: $F_{elm} = 3{,}4$.

Ezután a sorhatás hipotézisének ellenőrzéséhez állítsuk elő a

$$\Lambda_r = \frac{|\mathbf{B}|}{|\mathbf{K}_r + \mathbf{B}|} \qquad \text{kifejezést,}$$

ahol $\mathbf{K}_r$ és $\mathbf{B}$ a sorok eltéréseinek, illetve a cellákon belüli eltéréseknek a mátrixai. Esetünkben $\Lambda_r = 0{,}8402596$, így $F_r = 6{,}3417321$ az $F_{elm} = 3{,}4$ értékével szemben. Amikor csak a hatékonyságot vettük figyelembe függő változóként, és így vizsgáltuk meg e sorhatást, F_r értéke 7,823 volt.

A diszkriminanciafüggvényhez tartozó együtthatómátrix és a kommunalitások:

	y_1	y_2	kommunalitás
1.	0,5375	–0,8601	1,0287
2.	0,7966	0,5786	0,9693

Oszlophatás. Az oszlophatás elemzése a területi hatás fellépését vizsgálja. Itt is kiszámítottuk először a két függő változóhoz tartozó szórásokat, és F-próbával különkülön megmértük a várható értékek eltéréseit. A szabadságfok: 2 és 270.

y_1 esetében

a belső szóráshoz tartozó négyzetösszeg 60 390,862,

a külső szóráshoz tartozó négyzetösszeg 9 485,589,

a kapott $F_{emp} = 6{,}87$ elég nagy az $F_{elm} = 4{,}75$ értékhez képest;

y_2 esetében

a belső szóráshoz tartozó négyzetösszeg 408,742,

a külső szóráshoz tartozó négyzetösszeg 205,936,

a kapott $F_{emp} = 1{,}98$ lényegesen kisebb, mint az $F_{elm} = 4{,}75$.

Az oszlophatás hipotézisének ellenőrzéséhez kiszámítjuk a Wilks-féle Λ-t:

$$\Lambda_c = 0{,}93826520 \quad \text{és} \quad F_c = 9{,}277338.$$

Amikor függő változóként csak a hatékonyság szerepelt, $F = 2{,}07$ volt az $F_{elm} = 4{,}75$ értékkel szemben.

A diszkriminanciafüggvényhez tartozó mátrix és a kommunalitások

	y_1	y_2	kommunalitás
1.	–0,9362	–0,2996	0,9663
2.	–0,2738	0,9753	1,0262

Interakció. A sor- és oszlophatások, azaz az ágazati, illetve a területi szempontok egymás irányába megnyilvánuló befolyását mutatja meg. A várható értékek eltérésére vonatkozó F-próbánál most a szabadságfokok: 8 és 270.

y_1 esetében

a belső szóráshoz tartozó négyzetösszeg 18 737,752,

a külső szóráshoz tartozó négyzetösszeg 9 485,589,

az empirikus érték: $F_{emp} = 1{,}98$, az elméleti érték $F_{elm} = 2{,}6$;

y_2 esetében

a belső szóráshoz tartozó négyzetösszeg 129,177,

a külső szóráshoz tartozó négyzetösszeg 205,936,

az empirikus érték: $F_{emp} = 0{,}63$.

Az előbbiekhez hasonlóan egyetlen függő változóval (hatékonyság) számoltunk, ekkor $F_{\text{emp}} = 0{,}6412$ volt. Az interakció hipotéziséhez most a

$$\Lambda_i = \frac{|\mathbf{B}|}{|\mathbf{K}_i + \mathbf{B}_i|}$$

értéket határozzuk meg, ahol $\mathbf{K}_i$ az interakcióhoz tartozó eltérések mátrixa; $\Lambda_i = 0{,}9279081$ és $F_i = 1{,}3103725$.

Az elméleti érték 2,6 volt. Amikor csak a hatékonyság szerepelt függő változóként, $F_{\text{emp}} = 0{,}6412$ volt.

A diszkriminanciafüggvényhez tartozó együtthatómátrix és a kommunalitások:

	y_1	y_2	kommunalitás
1.	0,9994	0,0340	0,9999
2.	−0,0457	0,9933	0,9888

Végeredményként megállapíthatjuk, hogy a hatékonyság inkább a területi felosztást befolyásolja, míg a beruházási összeg az ágazatok elkülönülésében játszik szerepet.

8. fejezet

Kanonikus korrelációelemzés

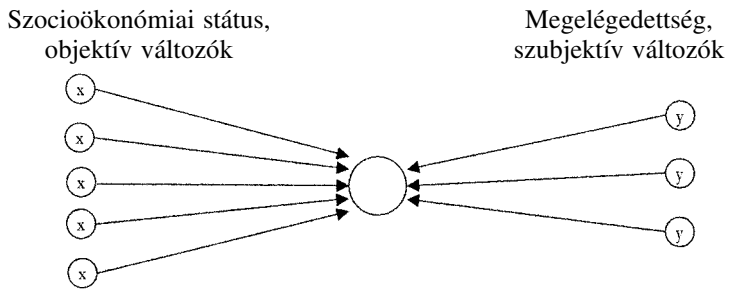
Az adatelemzéseknél gyakorta kerülünk olyan helyzetbe, amikor a változóhalmazt természetes módon két részre kell bontanunk, és a két változóhalmaz kapcsolatát kell vizsgálnunk. (Ilyen eset az, amikor egy fogalom különböző szféráit jellemezzük változókkal, és a köztük lévő viszonyt akarjuk elemezni.) Például a gazdasági hatékonyságot vizsgálva az élőmunka-hatékonyságot és az eszközhatékonyságot mérjük egy-egy változóhalmazzal, és a két terület változóstruktúráinak kapcsolódásait akarjuk elemezni. De ilyen probléma az is, amikor egy modell két elemét akarjuk összekapcsolni, és mindkét elem önmagában is sokváltozós rendszer.

Sokszor teszünk különbséget változók között aszerint, hogy függő vagy független, magyarázott vagy magyarázó, becsült vagy becslő változóról van szó. Például a megelégedettség szubjektív változói és a szocioökonómiai státus objektív változói közötti kapcsolatot jellemezhetjük kanonikus korrelációs modellel (8.1. ábra):

A kanonikus korrelációelemzést tekinthetjük a többszörös korreláció általánosításának. A többszörös korreláció esetében m darab x_i független (becslő vagy magyarázó) változó kapcsolatát vizsgáljuk egyetlen y függő (becsült vagy magyarázott) változóval.

A többszörös korrelációs modellben keressük az x_i változók azon függvényét (leggyakrabban lineáris függvényét), amely maximálisan korrelál y-nal.

A kanonikus korreláció esetén több x_i $(i = 1, \ldots, m_1)$ magyarázó változó sztochasztikus kapcsolatát több y_j $(j = 1, \ldots, m_2)$ függő változóval vizsgáljuk. A kanonikus



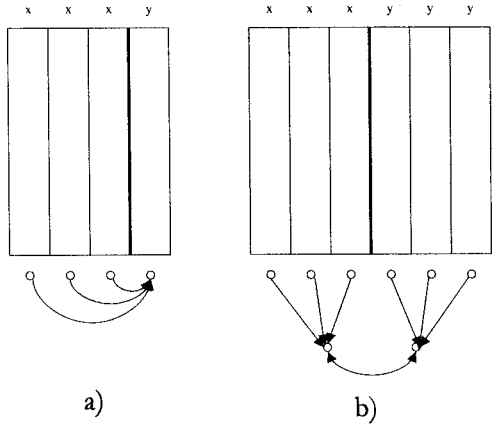
8.1. ábra. A kanonikus korreláció fogalmának szemléltetése

korrelációs modellben az x_i és y_j változók olyan lineáris függvényeit keressük, amelyek közötti korreláció maximális.

A 8.2. ábra ezt a két alapesetet világítja meg.

a) a többváltozós korreláció, többváltozós regresszió esete,

b) a kanonikus korrelációelemzés esete.



8.2. ábra. A többszörös és a kanonikus korreláció összehasonlítása

A kanonikus korreláció más megközelítésben tekinthető tulajdonképpen kettős faktorelemzésnek. Keressük a két változóhalmaz azon faktorait, amelyek közötti korreláció maximális.

8.1. A módszer leírása

Fogalmazzuk meg a problémát matematikai formulákkal! Legyen $\mathbf{X}$ az x_i $(i = 1,2,\ldots,m_1)$ változók n-szeri megfigyelésének $(n \times m_1)$-es mátrixa, és $\mathbf{Y}$ $(n \times m_2)$-es mátrix tartalmazza az m_2 db y változó n-szeri megfigyelését. Az általánosítás megszorítása nélkül feltehetjük, hogy $m_2 < m_1$. Hasznos egyszerűsítés kínálkozik a változók standardizálásával, vagyis ha változóinkat 0 várható értékű és 1 szórású változókká

transzformáljuk. Ebben az esetben ugyanis a kovarianciamátrix megegyezik a korrelációs mátrixszal. A továbbiakban ezért feltesszük, hogy x_i és y_i változók standardizáltak (az egyszerűség kedvéért ezt külön nem jelöljük). A kanonikus korrelációelemzés során az $\mathbf{x} = \left[x_1, x_2, \ldots, x_{m_1}\right]'$ és az $\mathbf{y} = \left[y_1, y_2, \ldots, y_{m_2}\right]'$ vektorváltozók sztochasztikus kapcsolatát vizsgáljuk az $\mathbf{x}$ és $\mathbf{y}$ vektorváltozók komponenseinek lineáris függvényein keresztül.

Legyenek $\mathbf{v}$ vektor elemei az x_i változók lineáris függvényei (lineáris kombinációi), tehát

$$\mathbf{v} = \mathbf{X}\,\mathbf{c},$$

ahol a $\mathbf{c}$ vektor $(m_1 \times 1)$-es méretű elemeit feltételi súlyoknak nevezzük. Hasonlóan definiáljuk az y_i változók lineáris kombinációját:

$$\mathbf{w} = \mathbf{Y}\,\mathbf{d},$$

ahol a $\mathbf{d}$ egy $(m_2 \times 1)$-es vektor, és az ún. következménysúlyokat tartalmazza.

A kanonikus korrelációelemzés feladata megtalálni azokat a $\mathbf{c}$ és $\mathbf{d}$ súlyokat, amelyek mellett a $\mathbf{v}$ és $\mathbf{w}$ ún. kanonikus változók közötti korreláció maximális. A $\mathbf{v}$ és $\mathbf{w}$ közötti korrelációt nevezzük *kanonikus korrelációs együtthatónak*. A kanonikus korrelációelemzés a két változóhalmazból állít elő olyan nem megfigyelt változópárokat, amelyek maximálisan korrelálnak. Eszerint a kanonikus korrelációelemzés két változóhalmaz szimultán főkomponens-elemzésének egy fajtája.

Tételezzük fel, hogy x_i és y_i változók standardizáltak. Ebben az esetben a korrelációmátrixokat a következőképpen kaphatjuk meg:

a) Az x_i valószínűségi változók között

$$\mathbf{R}_{xx} = \frac{1}{n}\mathbf{X}'\mathbf{X} \qquad (m_1 \times m_1),$$

b) az y_i valószínűségi változók között

$$\mathbf{R}_{yy} = \frac{1}{n}\mathbf{Y}'\mathbf{Y} \qquad (m_2 \times m_2),$$

c) és az x_i és y_i változók között

$$\mathbf{R}_{xy} = \frac{1}{n}\mathbf{X}'\mathbf{Y} \qquad (m_1 \times m_2).$$

A fenti mátrixokat tekinthetjük egy általános korrelációmátrix, $\mathbf{R} = \frac{1}{n}(\mathbf{X}\,\mathbf{Y})'(\mathbf{X}\,\mathbf{Y})$ partícióinak. A továbbiakban feltételezzük, hogy a $\mathbf{v}$ és a $\mathbf{w}$ kanonikus változók szintén standardizáltak, vagyis

$$\frac{1}{n}\mathbf{v}'\mathbf{v} = \frac{1}{n}\mathbf{c}'\mathbf{X}'\mathbf{X}\,\mathbf{c} = \mathbf{c}'\mathbf{R}_{xx}\mathbf{c} = 1, \tag{8.1}$$

$$\frac{1}{n}\mathbf{w}'\mathbf{w} = \frac{1}{n}\mathbf{d}'\mathbf{Y}'\mathbf{Y}\,\mathbf{d} = \mathbf{d}'\mathbf{R}_{yy}\mathbf{d} = 1.$$

Feltételezve, hogy a kanonikus változók standardizáltak $\mathbf{v}$ és $\mathbf{w}$ közötti korreláció

$$\frac{1}{n}\mathbf{v}'\mathbf{w} = \frac{1}{n}\mathbf{c}'\mathbf{X}'\mathbf{Y}\,\mathbf{d} = \lambda. \tag{8.2}$$

Célunk olyan $\mathbf{c}$ és $\mathbf{d}$ súlyokat találni, hogy λ maximális érték legyen, figyelembe véve a (8.1) szerinti feltételeket. Tehát egy feltételes szélsőérték-feladatot kell megoldanunk, ahol a feltételek egyenlőség formájában adottak.

A Lagrange-féle multiplikátor-módszer kínál megoldást. A Lagrange-függvény:

$$L = \mathbf{c}'\mathbf{R}_{xy}\mathbf{d} - \frac{1}{2}\mu[\mathbf{c}'\mathbf{R}_{xx}\mathbf{c} - 1] - \frac{1}{2}\rho[\mathbf{d}'\mathbf{R}_{yy}\mathbf{d} - 1]. \tag{8.3}$$

Az L függvény maximuma ott kereshető, ahol a parciális deriváltak egyenlők nullával. A μ és a ρ multiplikátor előtti $\frac{1}{2}$ szorzó egyszerűsíti a deriválás után kapott egyenleteket. A parciális deriváltak:

$$\frac{\partial L}{\partial \mathbf{c}'} = \mathbf{R}_{xy}\mathbf{d} - \mu\mathbf{R}_{xx}\mathbf{c} = \mathbf{0}, \tag{8.4}$$
$$\frac{\partial L}{\partial \mathbf{d}'} = \mathbf{R}_{yx}\mathbf{c} - \rho\mathbf{R}_{yy}\mathbf{d} = \mathbf{0},$$

ebből

$$\mathbf{R}_{xy}\mathbf{d} = \mu\mathbf{R}_{xx}\mathbf{c} \quad \text{és} \quad \mathbf{R}_{yx}\mathbf{c} = \rho\mathbf{R}_{yy}\mathbf{d}. \tag{8.5}$$

Ha a (8.5) egyenletek közül az elsőt szorozzuk balról $\mathbf{c}$-vel, a másodikat balról $\mathbf{d}'$-vel, akkor láthatjuk, hogy

$$\mathbf{c}'\mathbf{R}_{xy}\mathbf{d} = \mu\mathbf{c}'\mathbf{R}_{xx}\mathbf{c} = \mu, \tag{8.6}$$
$$\mathbf{d}'\mathbf{R}_{yx}\mathbf{c} = \rho\mathbf{d}'\mathbf{R}_{yy}\mathbf{d} = \rho,$$

A (8.2) szerint $\mathbf{c}'\mathbf{R}_{xy}\mathbf{d} = \mathbf{d}'\mathbf{R}_{yx}\mathbf{c} = \lambda$, amiből következik, hogy $\mu = \rho = \lambda$. Így a (8.4) helyett a következőket írhatjuk:

$$-\lambda\mathbf{R}_{xx}\mathbf{c} + \mathbf{R}_{xy}\mathbf{d} = \mathbf{0}, \tag{8.7}$$
$$\mathbf{R}_{yx}\mathbf{c} - \lambda\mathbf{R}_{yy}\mathbf{d} = \mathbf{0}.$$

Ezzel definiáltunk egy $m_1 + m_2 = m$ homogén egyenletből álló rendszert m ismeretlennel ($\mathbf{c}$ és $\mathbf{d}$) és egy ismeretlen együtthatóval (λ). A (8.7) egyenletrendszert a következőképpen írhatjuk:

$$\begin{bmatrix} -\lambda\mathbf{R}_{xx} & \mathbf{R}_{yx} \\ \mathbf{R}_{yx} & -\lambda\mathbf{R}_{yy} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} = \mathbf{0}.$$

Az egyenletrendszernek akkor és csak akkor van a triviálistól ($\mathbf{c} \neq \mathbf{0}$ és $\mathbf{d} \neq \mathbf{0}$) különböző megoldása, ha a particionált mátrix szinguláris, vagyis ha a determinánsa egyenlő nullával, tehát

$$\begin{vmatrix} -\lambda\mathbf{R}_{xx} & \mathbf{R}_{xy} \\ \mathbf{R}_{yx} & -\lambda\mathbf{R}_{yy} \end{vmatrix} = 0. \tag{8.8}$$

Használjuk fel a determinánsokra vonatkozó megfelelő előírásokat: ha egy oszlopot vagy egy sort szorzunk (vagy osztunk) egy konstanssal, a determináns értéke is szorzódik (vagy osztódik) a konstanssal. Ha a (8.8) determináns első m_1 sorát megszorozzuk ($-\lambda$)-val, és az utolsó m_2 oszlopát elosztjuk ($-\lambda$)-val, a determináns értéke 0 marad. Így kapjuk azt, hogy

$$\begin{vmatrix} \lambda^2\mathbf{R}_{xx} & \mathbf{R}_{xy} \\ \mathbf{R}_{yx} & \mathbf{R}_{yy} \end{vmatrix} = 0. \tag{8.9}$$

A determináns kifejtésével λ^2-nek egy m_1-ed fokú polinomjához jutunk, így általában m_1 különböző megoldást kapunk.

E megoldások pozitív gyökeit nevezzük kanonikus korrelációs együtthatóknak.

Először a legnagyobb érték érdekel bennünket, ez adja a maximális korrelációt: jelöljük λ_1-gyel. Ha ezt a λ_1 becslést beírjuk a (8.7) egyenletbe, s az így adódó homogén egyenleteket megoldjuk, megkapjuk a $\mathbf{c}$ és a $\mathbf{d}$ ismeretlenek becsléseit, amelyeket szintén indexszel látunk el ($\mathbf{c}_1$ és $\mathbf{d}_1$).

A (8.7) egyenleteket röviden a következő alakban írhatjuk:

$$\mathbf{c}_1 = \frac{1}{\lambda_1}\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}\mathbf{d}_1 \quad \text{és} \quad \mathbf{d}_1 = \frac{1}{\lambda_1}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\mathbf{c}_1.$$

Ezek a megoldások azonban nem biztos, hogy kielégítik a (8.1) szerinti feltételeket, mivel egy önkényesen megválasztott skalár (λ_1) is befolyásolja őket. Jelöljük a (8.7) egyenletek megoldásait $\tilde{\mathbf{c}}_1$-gyel és $\tilde{\mathbf{d}}_1$-gyel. Ha elvégezzük a következő korrekciót:

$$\mathbf{c}_1 = \frac{\tilde{\mathbf{c}}_1}{(\tilde{\mathbf{c}}_1\mathbf{R}_{xx}\tilde{\mathbf{c}}_1)^{\frac{1}{2}}}$$

és

$$\mathbf{d}_1 = \frac{\tilde{\mathbf{d}}_1}{(\tilde{\mathbf{d}}_1\mathbf{R}_{yy}\tilde{\mathbf{d}}_1)^{\frac{1}{2}}},$$

akkor az így kapott $\mathbf{c}_1$ és $\mathbf{d}_1$ már biztosan kielégíti a (8.1) szerinti feltételt. Vegyük figyelembe, hogy a továbbiakban a transzformáció után kapott $\mathbf{c}_1$ és $\mathbf{d}_1$ értékekkel dolgozunk.

Ezután a maradék gyökök közül választjuk a legnagyobbat, λ_2-t, a második kanonikus korrelációt, és kiszámítjuk a hozzá tartozó $\mathbf{c}_2$ és $\mathbf{d}_2$ értékeket, és így, általánosságban a λ_i-hez tartozó $\mathbf{c}_i$-t és $\mathbf{d}_i$-t.

Jelöljük a $\mathbf{c}_i$ vektorokból álló mátrixot $\mathbf{C}$-vel, és $\mathbf{D}$ tartalmazza a $\mathbf{d}_i$ vektorokat. A kanonikus korrelációs együtthatókat (λ_i) helyezzük el a $\mathbf{L}$ diagonális mátrix diagonálelemeibe! A nem megfigyelt $\mathbf{v}$ és $\mathbf{w}$ kanonikus változókat pedig a $\mathbf{V}$ és $\mathbf{W}$ mátrixok tartalmazzák.

Általánosságban a feltételi- és következménysúlyokat a következő egyenletekből számíthatjuk ki (a (8.7) átalakított formája):

$$\mathbf{C} = \mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}\mathbf{D}\,\mathbf{L}^{-1}$$

és

$$\mathbf{D} = \mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\mathbf{C}\,\mathbf{L}^{-1}.$$

Ha a változók száma nagy, az előző eljárás nehézkessé válik. Különösen a (8.9) determináns kifejtése és megoldása okoz gondot. Ezt a problémát megkerülve nézzünk egy másik eljárást! Induljunk ki a (8.5) egyenletekből! Szorozzuk be a második egyenlet mindkét oldalát balról $\lambda^{-1}\mathbf{R}_{yy}^{-1}$-zel (feltéve, hogy az inverz létezik):

$$\mathbf{d} = \lambda^{-1}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\mathbf{c}. \tag{8.10}$$

Ezt az elsőbe behelyettesítve:

$$\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\mathbf{c} = \lambda^2\mathbf{R}_{xx}\mathbf{c}.$$

Balról mindkét oldalt szorozva $\mathbf{R}_{xx}^{-1}$-zel (feltéve, ha az inverz létezik):

$$\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\mathbf{c} = \lambda^2\mathbf{c}. \tag{8.11}$$

Ez a következő alakban írható fel:

$$\left(\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx} - \lambda^2\mathbf{E}\right)\mathbf{c} = \mathbf{0}.$$

Ebből láthatjuk, hogy λ^2 nem más, mint az $\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}$ mátrix sajátértéke, $\mathbf{c}$ pedig a hozzátartozó sajátvektor.

A maximális kanonikus korrelációt a legnagyobb sajátérték adja. Az ehhez tartozó $\mathbf{c}_1$ ismeretében (8.10)-ből kiszámíthatjuk $\mathbf{d}_1$-et. Itt ugyanazzal a problémával találkozunk, amivel az előző megoldásnál. Ahhoz, hogy a (8.1) szerinti feltételt ki tudjuk elégíteni, normalizálni kell a $\mathbf{c}_1$ és a $\mathbf{d}_1$ vektorokat (hogy $\mathbf{c}_1'\mathbf{R}_{xx}\mathbf{c} = 1$ legyen). Természetesen, ha először $\mathbf{c}$-t fejezzük ki (8.5)-ből, nem $\mathbf{d}$-t, az eredmény ugyanaz lesz. A sajátérték és a sajátvektor számolásánál problémát okozhat, hogy az $\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}$ mátrix nem szimmetrikus. A legtöbb sajátértéket meghatározó algoritmus feltételezi, hogy a mátrix szimmetrikus, ezért meghatározzuk (8.11) egy szimmetrikus alternatíváját. Definiáljuk a következő kisegítő vektort:

$$\mathbf{q} = \mathbf{R}_{xx}^{\frac{1}{2}}\mathbf{c}, \quad \text{amelyből} \quad \mathbf{c} = \mathbf{R}_{xx}^{\frac{1}{2}}\mathbf{q}. \tag{8.12}$$

Ha ezt beírjuk a (8.11) egyenletbe, és balról megszorozzuk az egyenlet mindkét oldalát $\mathbf{R}_{xx}^{\frac{1}{2}}$-vel, olyan egyenlethez jutunk, amelyben a bal oldali mátrix már bizonyíthatóan szimmetrikus:

$$\mathbf{R}_{xx}^{\frac{1}{2}}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\mathbf{R}_{xx}^{\frac{1}{2}}\mathbf{q} = \lambda^2\mathbf{q}. \tag{8.13}$$

A (8.13)-ban található mátrix $(m_1 \times m_1)$-es, m_1 különböző sajátértéke és m_1 különböző sajátvektora van. Tartalmazza a $\mathbf{Q}$ mátrix ezeket a sajátvektorokat! A $\mathbf{Q}'\mathbf{Q}$ szorzatmátrix diagonális lesz, és ha a sajátvektorok egységnyi hosszúságúak, akkor $\mathbf{Q}'\mathbf{Q} = \mathbf{E}$.

A (8.12) alapján $\mathbf{Q}'\mathbf{Q} = \mathbf{C}'\mathbf{R}_{xx}\mathbf{C} = \mathbf{E}$, amely éppen a (8.1) szerinti feltétel kielégítését jelenti. Mivel a $\mathbf{V} = \mathbf{X}\mathbf{C}$, a kanonikus változók szórásnégyzete:

$$\frac{\mathbf{V}'\mathbf{V}}{n} = \mathbf{C}'\mathbf{R}_{xx}\mathbf{C} = \mathbf{E}. \tag{8.14}$$

Ez azt a követelményt foglalja magában, hogy a $\mathbf{V}$ oszlopvektorai ortonormáltak, vagyis a lineáris kombinációval képzett $\mathbf{v}_i$ kanonikus változók páronként korrelálatlanok, valamint standardizáltak. Ugyanez belátható $\mathbf{W}$-re is (azaz az y változókból képzett $\mathbf{w}_i$ kanonikus változók egymással korrelálatlanok és standardizáltak). A $\mathbf{V}$ és $\mathbf{W}$ közötti korrelációról

$$\frac{\mathbf{V}'\mathbf{W}}{m} = \mathbf{C}'\mathbf{R}_{xy}\mathbf{D}$$

pedig be lehet látni, hogy diagonális mátrix: ($\mathbf{L}$). A (8.5) egyenletből

$$\mathbf{R}_{xy}\mathbf{D} = \mathbf{R}_{xx}\mathbf{C}\,\mathbf{L},$$

amit balról beszorzunk $\mathbf{C}'$-vel:

$$\mathbf{C}'\mathbf{R}_{xy}\mathbf{D} = \mathbf{C}'\mathbf{R}_{xx}\mathbf{C}\,\mathbf{L} = \mathbf{L}. \tag{8.15}$$

Ez azt jelenti, hogy minden $\mathbf{v}_i$-hez tartozik egy $\mathbf{w}_j$, amellyel maximálisan korrelál, míg a többi $\mathbf{w}_j$-vel korrelálatlan,

$$\mathbf{v}_i\mathbf{w}_j = \begin{cases} 0, & \text{ha} \quad i \neq j, \\ \lambda_i, & \text{ha} \quad i = j. \end{cases}$$

A kanonikus korrelációs együttható nagyságának vizsgálata mellett fontos lehet a kanonikus változók (faktorok) értelmezése is. A $\mathbf{c}$ és a $\mathbf{d}$ együtthatók ismeretében tudjuk, hogy az eredeti változók milyen súlyú lineáris kombinációi állítják elő a kanonikus faktorokat. A kanonikus faktoroknak könnyen értelmezhető módjához jutunk, ha kiszámítjuk

a faktorelemzésnél jól bevált faktorsúlyok mátrixához hasonló, kanonikus faktorsúlyok mátrixát. A két változóhalmaz között maximális korrelációt adó első kanonikus faktorpár és az azokat előállító változók közötti korrelációkat tartalmazó kanonikus faktorstruktúrát a következőképpen számíthatjuk ki.

Az első bal oldali kanonikus faktor ($\mathbf{v} = \mathbf{X}\,\mathbf{c}$) és a bal oldali változók $\mathbf{x}$ közötti korrelációk (felhasználva, hogy mind $\mathbf{x}$, mind $\mathbf{v}$ standardizált)

$$\mathbf{s}_1 = \frac{1}{n}\mathbf{X}'\mathbf{v} = \frac{1}{n}\mathbf{X}'\mathbf{X}\,\mathbf{c} = \mathbf{R}_{xx}\mathbf{c}.$$

Ehhez hasonlóan az első jobb oldali kanonikus faktor struktúrája

$$\mathbf{s}_2 = \frac{1}{n}\mathbf{Y}'\mathbf{w} = \frac{1}{n}\mathbf{Y}'\mathbf{Y}\,\mathbf{d} = \mathbf{R}_{yy}\mathbf{d}.$$

A többi kanonikus faktor esetén ugyanígy megkaphatjuk, hogy az adott kanonikus faktor előállításában melyek a legjelentősebb változók. A kanonikus faktorsúlyokat felhasználva kiszámíthatjuk, hogy a kanonikus faktorok a változók varianciájának milyen arányát magyarázzák. Így az első bal oldali kanonikus faktor ($\mathbf{v}$) a bal oldali változók varianciájának $\frac{\mathbf{s}_1'\mathbf{s}_1}{m_1}$ arányát reprodukálja. Ha ezt az arányt megszorozzuk az első (a megfelelő) kanonikus korreláció négyzetével, akkor a bal oldali változók varianciájának a jobb oldali változók első (megfelelő) kanonikus faktora által magyarázott arányát kapjuk: $r_x = \frac{\mathbf{s}_1'\mathbf{s}_1}{m_1}\lambda_1^2$.

D. Stewart és W. Love ezt nevezte a bal oldali változóhalmaz *redundanciájának* az adott jobb oldali változóhalmaz kanonikus faktora esetén.

Az első jobb oldali kanonikus faktor a jobb oldali változóhalmaz varianciájának $\left(\frac{\mathbf{s}_2'\mathbf{s}_2}{m_2}\right)$ 100%-át magyarázza. A jobb oldali teljes variancia $r_y = \frac{\mathbf{s}_2'\mathbf{s}_2}{m_2}\lambda_1^2$ arányát magyarázza a bal oldali változóhalmaz első kanonikus faktora, és az lesz a jobb oldali változóhalmaz redundanciája az első bal oldali kanonikus faktorra.

A két arány, r_x és r_y természetesen nem kell hogy megegyezzen. Például, ha a bal oldali első kanonikus faktor az első főkomponenshez hasonló, a jobb oldali kanonikus faktor pedig a jobb oldali változóhalmaz egy kis varianciájú (sajátértékű) főkomponenséhez hasonlít, akkor a bal oldali változóhalmazhoz tartozó redundancia nagyobb lesz ($r_x > r_y$).

Mivel több kanonikus faktorpár is számítható, egy bal oldali változóhalmaz teljes redundanciáját, adott jobb oldali változóhalmaz esetén az egy-egy kanonikus faktorpárra számított redundanciák összege adja:

$$r_{d_1} = \sum_{i=1}^{m_1} r_{x_i}.$$

Ugyanígy a jobb oldali változóhalmaz teljes redundanciája a kanonikus modell szerint:

$$r_{d_2} = \sum_{i=1}^{m_2} r_{y_i}.$$

A kanonikus korrelációs együtthatók szignifikancia-próbáját végezzük el Bartlett-féle χ^2-próba alapján. E próba elvégzéséhez fel kell tennünk, hogy $\mathbf{x}$ és $\mathbf{y}$ többvál-

tozós normális eloszlású valószínűségi változók. A próbához definiáljuk még a Wilks-féle lambdát:

$$\Lambda_1 = \prod_{i=1}^{m_2}(1 - \lambda_i^2). \tag{8.16}$$

Tekintsük a következő változót (amely Λ függvénye):

$$\chi^2 = -[n - 1 - 0{,}5(m_1 + m_2 + 1)] \ln \Lambda_1, \tag{8.17}$$

amely közelítően χ^2 eloszlású, $(m_1 \times m_2)$ szabadságfokkal.

A nullhipotézisünk az, hogy $\mathbf{x}$ vektorváltozó korrelálatlan $\mathbf{y}$ vektorváltozóval. A próba a szokásos módon végezhető el. Ha elvetjük a hipotézist, akkor az első (maximális) kanonikus korrelációt elhagyjuk Λ_1-ből, és a maradék $(m_2 - 1)$ kanonikus korrelációs együttható szignifikanciáját vizsgáljuk. Az új Λ_2 és χ^2 a következő:

$$\Lambda_2 = \prod_{i=2}^{m_2}(1 - \lambda_i^2)$$

és

$$\chi^2 = -[n - 1 - 10{,}5(m_1 + m_2 + 1)] \ln \Lambda_2,$$

$(m_1 - 1)(m_2 - 1)$ szabadságfokkal. Általában, ha Λ-ból $(r - 1)$ kanonikus korrelációs együtthatót hagyunk el

$$\Lambda_r = \prod_{i=r}^{m_2}\left(1 - \lambda_i^2\right) \tag{8.18}$$

és

$$\chi^2 = -[n - 1 - 0{,}5(m_1 + m_2 + 1)] \ln \Lambda_r, \tag{8.19}$$

$(m_1 - r + 1)(m_2 - r + 1)$ szabadságfokkal.

Azok a kanonikus korrelációs együtthatók fognak szignifikánsan különbözni 0-tól, amelyek esetében még elvetettük a nullhipotézist.

A függő változók regressziós becslése a kanonikus változók segítségével

Két változóhalmaz közötti sztochasztikus kapcsolatot vizsgáltuk a kanonikus változók segítségével. A kapcsolat szorosságán kívül érdekelhet bennünket az is, hogy a két halmaz közül a függőnek tekintett változóhalmaz hogyan becsülhető a magyarázó változók segítségével. Ezt a regressziós problémát a kanonikus változók felhasználásával oldjuk meg.

Tegyük fel, hogy az $\mathbf{x}$ változók függetlenek, azaz $\mathbf{R}_{xx} = \mathbf{I}$, valamint, hogy az összes független változó hatást gyakorol az összes függő változóra (y). Ebből következik, hogy a függő változók között érvényesülnek kölcsönhatások (interdependencia).

A kanonikus korrelációs együtthatók és a kanonikus változók becsléseit az előzőek alapján kaphatjuk meg. Az egyszerűség kedvéért a becsléseket csak a regressziós egyenlet függő változója esetén jelöljük külön, más esetben nem teszünk a jelölésben különbséget a becsült és az elméleti érték között, mivel azok mindig értelemszerűen következnek. A kanonikus változók az előzőek alapján:

$\mathbf{V} = \mathbf{X}\,\mathbf{C}$ és $\mathbf{W} = \mathbf{Y}\,\mathbf{D}$, valamint a variancia-kovarianciamátrixok:

$$\frac{\mathbf{V}'\mathbf{V}}{n} = \mathbf{I}, \quad \frac{\mathbf{W}'\mathbf{W}}{n} = \mathbf{I} \quad \text{és} \quad \mathbf{V}'\mathbf{W} = \mathbf{L}. \tag{8.20}$$

Először a „kanonikus függő” változókat becsüljük a „kanonikus független” változók segítésével. $\mathbf{W} = \mathbf{V}\,\mathbf{B}$, ahol $\mathbf{B}$ becslését a legkisebb négyzetek módszerével kaphatjuk meg:

$$\widehat{\mathbf{B}} = (\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}'\mathbf{W} = \mathbf{L}, \tag{8.21}$$
$$\widehat{\mathbf{W}} = \mathbf{V}\,\mathbf{L}.$$

Ezután $\mathbf{V}$ kanonikus változókkal becsüljük az $\mathbf{Y}$ megfigyelt függő változókat:

$$\widehat{\mathbf{Y}} = \mathbf{V}\,\mathbf{A}.$$

Az $\mathbf{A}$ becslését a legkisebb négyzetek módszerével határozzuk meg, és felhasználjuk a (8.20)-ban megadott feltételeket:

$$\widehat{\mathbf{Y}} = \mathbf{X}\,\mathbf{C}(\mathbf{C}'\mathbf{X}'\mathbf{X}\,\mathbf{C})^{-1}\mathbf{C}'\mathbf{X}'\,\mathbf{Y},$$
$$\widehat{\mathbf{Y}} = \mathbf{X}\,\mathbf{C}(n\,\mathbf{E})^{-1}\mathbf{C}'\mathbf{R}_{xy}n, \tag{8.22}$$
$$\widehat{\mathbf{Y}} = \mathbf{X}\,\mathbf{C}\,\mathbf{C}'\mathbf{R}_{xy}.$$

A (8.7) egyenletekből

$$\mathbf{R}_{yx}\mathbf{C} = \mathbf{R}_{yy}\mathbf{D}\,\mathbf{L}.$$

Ha transzponáljuk, behelyettesíthetünk vele a (8.22)-be:

$$\mathbf{C}'\mathbf{R}'_{yx} = \mathbf{L}'\mathbf{D}'\mathbf{R}_{yy}.$$

Így a kanonikus korrelációs együttható, a kanonikus változók, a kanonikus súlyok segítségével megadtuk az y függő változók regressziós becslését. Megmutatható, hogy ez megegyezik a legkisebb négyzetek módszerével közvetlenül kapható becsléssel.

8.2. Példa a kanonikus korrelációelemzésre

A kanonikus korreláció alkalmazását bemutató példa adatai a KSH Megyei évkönyvek 1975-ös köteteiből származnak, és 82 város infrastrukturális fejlettségét, foglalkozási, kereseti adatait, egészségügyét és iskolarendszerét jellemző változóit tartalmazza. Az alkalmazott változók a következők:

– 1000 fős népességre jutó vándorlási különbözet,
– a szocialista iparban foglalkoztatottak aránya a lakónépességben,
– a munkás- és alkalmazott nők aránya a szocialista szektorban,
– a munkások aránya a szocialista szektorban,
– a havi átlagbér, Ft,
– a mezőgazdasági termelőszövetkezeti tagok (nyugdíjasok nélkül) aránya a népességben,
– a mezőgazdasági termelőszövetkezeti nő tagok (nyugdíjasok nélkül) aránya a népességben,
– a mezőgazdasági termelőszövetkezeti tagok keresete, Ft,
– a kiépített utak (1000 m^2) és a tanácsi belterületi utak aránya,
– az egy lakásra jutó lakónépesség, fő,
– az 1000 lakosra jutó épített lakások száma nyaraló nélkül,
– a vízhálózatba bekapcsolt lakások aránya,
– a zárt közcsatorna-hálózatba bekapcsolt lakások aránya,

– a rendszeres szemétgyűjtésbe bekapcsolt lakások aránya,
– a 1000 lakosra jutó orvosok száma, fő,
– az 1000 lakosra jutó kórházi ágyak száma,
– a megfelelő szaktanár által leadott órák aránya a felső tagozatban,
– az osztott iskolák aránya,
– az egy osztályteremre jutó tanulók száma, fő,
– az 1000 lakosra jutó középfokú tanintézetek száma,
– az 1000 lakosra jutó felsőfokú tanintézetek száma.

A közölt számítások az MTA Szociológiai Intézetében folytatott, *A településrendszer változás-típusai* című kutatási téma keretében készültek. A kutatási kérdés az volt, hogy a város infrastrukturális fejlettségi szintjét milyen mértékben határozza meg, illetve milyen szorosan kapcsolódik hozzá a városi társadalmat jellemző foglalkozási szerkezet, a kereseti viszonyok, az egészségügy és az iskolarendszer szférája. További kérdés, hogy az infrastruktúra változóival szembeállított, a helyi társadalmat jellemző változók milyen struktúrája kapcsolódik legjobban az infrastruktúrához.

A települések infrastruktúráját a következő mutatók mérték:

– a kiépített utak és a tanácsi belterületi utak aránya,

– az egy lakásra jutó lakónépesség,

– az 1000 lakosra jutó épített lakások száma nyaraló nélkül,

– a vízhálózatba bekapcsolt lakások aránya,

– a zárt közcsatorna-hálózatba bekapcsolt lakások aránya,

– a rendszeres szemétgyűjtésbe bekapcsolt lakások aránya.

A számítások eredményeit a 8.1–8.3. táblázat tartalmazza.

A legnagyobb kanonikus korreláció 0,86-os értéke szoros kapcsolatot jelez a két vizsgált változóhalmaz között. A két változóhalmaz első kanonikus faktorstruktúrája azt mutatja, hogy a közinfrastruktúra mutatóihoz és a laksűrűséghez legszorosabban a foglalkozási szerkezet (ipari foglalkoztatottak és a mezőgazdasági termelőszövetkezeti tagok), a kereset és az iskolarendszer szférájának két mutatója kapcsolódik.

A bal oldali változókból kiszűrt teljes variancia: 0,489, a jobb oldali változókból kiszűrt teljes variancia: 1,000.

A bal oldali változók teljes redundanciája, adottnak véve a jobb oldaliakat: 0,204; a jobb oldali változók teljes redundanciája, adottnak véve a bal oldaliakat: 0,522.

A Wilks-féle Λ az összes változóra: 0,0601661.

A χ^2 az összes változóra: 196,7451897.

Szabadságfok: 90.

Az első jobb oldali kanonikus faktor a 6 infrastruktúra-mutató varianciájának 50%-át magyarázza. Ez az érték egészen közel áll a feltétel nélkül számítható főkomponens összvarianciából való részesedéséhez, amely (a külön elvégzett főkomponens-elemzés alapján) 54%. A jobb oldali változóhalmaz teljes redundanciája is magasnak mondható: 0,52. A bal oldali változóhalmaz kanonikus faktorai lényegesen eltérnek a főkomponensektől, a teljes redundancia mutatója is lényegesen alacsonyabb (0,2), ami azt jelzi, hogy a városi társadalmi szerkezetet és az urbanizációt leíró szférák ugyan jól meghatározzák a városok infrasturktúráját leíró 6 változó megfigyelési értékeit, de fordítva az állítás már nem tartható, lényegesen gyengébb a kapcsolódás.

Változó	Bal oldali kanonikus súlyok faktoronként			Jobb oldali kanonikus súlyok faktoronként		
	F_1	F_2	F_3	F_1	F_2	F_3
1000 fős népességre jutó vándorlási különbözet	0,425	0,225	−0,824			
A szocialista iparban foglalkoztatottak aránya a lakónépességben	0,303	0,493	0,274			
A munkás- és alkalmazott nők aránya a szocialista iparban	0,005	−0,254	0,137			
A munkások aránya a szocialista iparban	0,283	−0,038	0,079			
A havi átlagbér, Ft	0,311	−0,360	0,223			
A mezőgazdasági tsz-tagok (nyugdíjasok nélkül) aránya a népességben	0,272	0,638	−0,079			
A mezőgazdasági tsz nő tagok (nyugdíjasok nélkül) aránya a termelőszövetkezetben	−0,172	−0,425	0,195			
A mezőgazdasági tsz-tagok keresete, Ft	−0,064	−0,465	−0,124			
Az 1000 lakosra jutó orvosok száma, fő	−0,021	−0,148	−0,032			
Az 1000 lakosra jutó kórházi ágyak száma	−0,079	−0,445	−0,163			
A megfelelő szaktanár által leadott órák aránya a felső tagozatban	0,147	0,335	−0,221			
Az osztott iskolák aránya	−0,018	−0,204	0,215			
Az egy osztályteremre jutó tanulók száma, fő	0,024	0,342	−0,080			
Az 1000 lakosra jutó középfokú tanintézetek száma	0,186	0,111	−0,063			
Az 1000 lakosra jutó felsőfokú tanintézetek száma	0,107	0,138	0,170			
A kiépített utak és a tanácsi belterületi utak aránya				0,055	−0,386	0,117
Az egy lakásra jutó lakónépesség, fő				0,411	0,662	−0,317
Az 1000 lakosra jutó épített lakások száma nyaraló nélkül				0,125	−0,145	−0,847
A vízhálózatba bekapcsolt lakások aránya				0,041	−0,369	0,207
A zárt közcsatorna-hálózatba bekapcsolt lakások aránya				0,362	0,547	0,724
A rendszeres szemétgyűjtésbe bekapcsolt lakások aránya				0,317	−0,593	−0,486

8.1. táblázat. A kanonikus faktorsúlyok

Változó	Bal oldali kanonikus súlyok faktoronként			Jobb oldali kanonikus súlyok faktoronként		
	F_1	F_2	F_3	F_1	F_2	F_3
1000 fős népességre jutó vándorlási különbözet	0,488	0,044	−0,790			
A szocialista iparban foglalkoztatottak aránya a lakónépességben	0,707	0,126	0,398			
A munkás- és alkalmazott nők aránya a szocialista iparban	−0,432	0,008	−0,227			
A munkások aránya a szocialista szektorban	0,359	0,024	0,442			
A havi átlagbér, Ft	0,613	−0,186	0,389			
A mezőgazdasági tsz-tagok (nyugdíjasok nélkül) aránya a népességben	−0,751	0,406	0,010			
A mezőgazdasági tsz nő tagok (nyugdíjasok nélkül) aránya						
a termelőszövetkezetben	0,185	−0,288	0,069			
A mezőgazdasági tsz-tagok keresete, Ft	−0,182	−0,441	0,007			
Az 1000 lakosra jutó orvosok száma, fő	−0,149	0,042	−0,064			
Az 1000 lakosra jutó kórházi ágyak száma	0,049	−0,511	−0,356			
A megfelelő szaktanár által leadott órák aránya a felső tagozatban	0,464	0,071	−0,208			
Az osztott iskolák aránya	0,421	−0,067	−0,128			
Az egy osztályterembe jutó tanulók száma, fő	0,069	0,075	−0,011			
Az 1000 lakosra jutó középfokú tanintézetek száma	0,171	−0,195	−0,236			
Az 1000 lakosra jutó felsőfokú tanintézetek száma	0,094	0,093	0,258			
A kiépített utak és a tanácsi belterületi utak aránya				0,613	−0,372	0,347
Az egy lakásra jutó lakónépesség, fő				0,755	0,573	−0,057
Az 1000 lakosra jutó épített lakások száma nyaraló nélkül				0,295	−0,128	−0,866
A vízhálózatba bekapcsolt lakások aránya				0,672	−0,484	0,194
A zárt közcsatorna-hálózatba bekapcsolt lakások aránya				0,927	−0,045	0,257
A rendszeres szemétgyűjtésbe bekapcsolt lakások aránya				0,808	−0,514	0,039

8.2. táblázat. A változók faktorstruktúrája

Faktorok	Bal oldali		Jobb oldali	
	kiszűrt variancia	redundancia	kiszűrt variancia	rendundancia
1	0,168	0,125	0,499	0,373
2	0,055	0,026	0,164	0,078
3	0,100	0,031	0,163	0,050
4	0,058	0,013	0,042	0,010
5	0,061	0,006	0,087	0,009
6	0,047	0,003	0,045	0,003

8.3. táblázat. A kanonikus faktorok kiszűrt varianciái, redundanciák

Kiszűrt gyökök	Kanonikus R	R^2	χ^2	Szabadság-fok	Λ
0	0,8639	0,746	196,75	90	0,0602
1	0,6910	0,477	100,73	70	0,2372
2	0,5541	0,307	55,30	52	0,4538
3	0,4760	0,227	29,63	36	0,6549
4	0,3199	0,102	11,65	22	0,8467
5	0,2383	0,057	4,09	10	0,9432

8.4. táblázat. A szignifikancia vizsgálata

8.3. Kanonikus faktorelemzés

A kanonikus faktorelemzés a kanonikus korrelációelemzés egy érdekes alkalmazása a faktorelemzés elméletére. A kanonikus faktorelemzés esetén a kanonikus korrelációt arra használjuk, hogy megkeressük a megfigyelt változók olyan lineáris kombinációját, amely a megfigyelt változók főkomponensei (faktorai) lineáris kombinációjával maximálisan korrelál.

A kanonikus faktorelemzés módszerét C. R. Rao fejlesztette ki. Rao olyan faktorokat keresett, amelyek egymással korrelálatlanok, és a megfigyelt változókkal maximálisan korrelálnak. A módszer előnye, hogy invariáns a lineáris transzformációval szemben. A faktorelemzés eredménye megváltozik a változók standardizálása hatására, ha a korrelációmátrix helyett a variancia-kovarianciamátrixot használjuk. A kanonikus faktorelemzés eredménye ezzel szemben nem változik.

A kanonikus faktorelemzés kiinduló egyenlete a faktorelemzés modellje:

$$\mathbf{X} = \mathbf{F}\,\mathbf{S}' + \mathbf{E}, \tag{8.23}$$

vagyis a megfigyelt változók ($\mathbf{X}$) a nem megfigyelt változóknak, a faktoroknak ($\mathbf{F}$) lineáris függvénye, ahol $\mathbf{E}$ tartalmazza a véletlen komponenseket. A fenti egyenletet a következő formában is írhatjuk:

$$\mathbf{X} = \mathbf{M} + \mathbf{E}. \tag{8.24}$$

Eszerint minden változót két részre bonthatunk, az egyik rész a faktorokkal magyarázható, szisztematikus komponens (m_i), a másik rész a véletlen komponens: e_i; ($x_i = m_i + e_i$).

A megfigyelt változókról ($\mathbf{x}$) feltételezzük, hogy standardizáltak (ez nincs hatással az eredményekre). A következő feltételek fennállását írjuk még elő:

$$\frac{1}{n}\mathbf{M}'\mathbf{E} = \mathbf{0}, \qquad \frac{1}{n}\mathbf{F}'\mathbf{F} = \mathbf{I},$$

ahol $\mathbf{I}$ = az egységmátrix, vagyis a faktorok korrelálatlanok és standardizáltak, és

$$\frac{1}{n}\mathbf{E}'\mathbf{E} = \mathbf{U}^2,$$

ahol $\mathbf{U}^2$ diagonális mátrix, vagyis a véletlen komponensek korrelálatlanok.

Míg a kanonikus korreláció esetében a független (magyarázó) változók halmazának kapcsolatát vizsgáltuk a függő változók halmazával, most a függő változók szerepét a megfigyelt független változók faktorelemzéssel előállított lineáris kombinációi veszik át ($\mathbf{M}$).

Tehát a megfigyelt változók $\mathbf{X}$ és a faktorokkal becsült $\mathbf{M}$ szisztematikus komponens kapcsolatát vizsgáljuk. A többszörös korreláció esetében is hasonló a kérdésfeltevés. Ott a becslő regressziós függvény kapcsolatát vizsgáljuk a függő változóval. Így a kanonikus faktorelemzést tekinthetjük a többszörös korreláció egyfajta általánosításának.

A feladat tehát megtalálni a megfigyelt változók olyan transzformációját, amely maximálisan korrelál a szisztematikus komponensek lineáris kombinációjával. Matematikai formulákkal

$$\mathbf{v} = \mathbf{X}\,\mathbf{c}, \qquad \mathbf{w} = \mathbf{M}\,\mathbf{d}. \tag{8.25}$$

Maximalizáljuk $\mathbf{v}$ és $\mathbf{w}$ kanonikus változók korrelációját,

$$\frac{\mathbf{v}'\mathbf{w}}{n} \longrightarrow \max,$$

a következő feltételek mellett:

$\frac{1}{n}\mathbf{v}'\mathbf{v} = 1$ és $\frac{1}{n}\mathbf{w}'\mathbf{w} = 1$, azaz $\mathbf{v}$ és $\mathbf{w}$ standardizáltak, azaz

$$\begin{aligned}
\frac{1}{n}\mathbf{v}'\mathbf{v} &= \frac{1}{n}\mathbf{c}'\mathbf{X}'\mathbf{X}\,\mathbf{c} = \mathbf{c}'\mathbf{R}\,\mathbf{c},\\
\frac{1}{n}\mathbf{w}'\mathbf{w} &= \frac{1}{n}\mathbf{d}'\mathbf{M}'\mathbf{M}\,\mathbf{d} = \mathbf{d}'\mathbf{R}^*\mathbf{d},\\
\frac{1}{n}\mathbf{v}'\mathbf{w} &= \frac{1}{n}\mathbf{c}'\mathbf{X}'\mathbf{M}\,\mathbf{d} = \mathbf{c}'\mathbf{R}^*\mathbf{d},
\end{aligned} \tag{8.26}$$

ahol $\mathbf{R}^*$ = az ún. redukált korrelációmátrix, és a következőképpen értelmezzük. Mivel

$$\begin{aligned}
\mathbf{R} = \frac{1}{n}\mathbf{X}'\mathbf{X} &= \frac{1}{n}(\mathbf{F}\,\mathbf{S}' + \mathbf{E})'(\mathbf{F}\,\mathbf{S}' + \mathbf{E}) =\\
&= \frac{1}{n}\mathbf{S}\,\mathbf{F}'\mathbf{F}\,\mathbf{S}' + \mathbf{U}^2,
\end{aligned}$$

és mivel $\frac{1}{n}\mathbf{F}'\mathbf{F} = \mathbf{I}$, $\quad \mathbf{R} = \mathbf{S}\,\mathbf{S}' + \mathbf{U}^2$.

A redukált korrelációmátrixot

$$\mathbf{R} - \mathbf{U}^2 = \mathbf{R}^* \tag{8.27}$$

definiálja, feltéve, hogy a hiba szórásnégyzeteit ismerjük.

A feladat megoldása egy feltételes szélsőérték-feladat megoldását jelenti. Ehhez előállítjuk a Lagrange-függvényt:

$$L = \mathbf{c}'\mathbf{R}^*\mathbf{d} - \frac{1}{2}\mu(\mathbf{c}'\mathbf{R}\,\mathbf{c} - 1) - \frac{1}{2}\rho(\mathbf{d}'\mathbf{R}^*\mathbf{d} - 1). \tag{8.28}$$

A parciális deriváltakat egyenlővé tesszük 0-val:

$$\frac{\partial L}{\partial \mathbf{c}'} = \mathbf{R}^*\mathbf{d} - \mu\mathbf{R}\,\mathbf{c} = \mathbf{0}, \tag{8.29}$$

$$\frac{\partial L}{\partial \mathbf{d}'} = \mathbf{R}^*\mathbf{c} - \rho\mathbf{R}^*\mathbf{d} = \mathbf{0}. \tag{8.30}$$

Ha (8.29)-et balról megszorozzuk $\mathbf{c}'$-vel és (8.30)-at balról $\mathbf{d}'$-vel, láthatjuk, hogy

$$\mathbf{c}'\mathbf{R}^*\mathbf{d} = \mu\mathbf{c}'\mathbf{R}\,\mathbf{c} = \mu \quad \text{és} \quad \mathbf{d}'\mathbf{R}^*\mathbf{c} = \rho\mathbf{d}'\mathbf{R}^*\mathbf{d} = \rho,$$

így

$$\mu = \rho = \lambda,$$

ahol λ = a kanonikus korreláció $\mathbf{v}$ és $\mathbf{w}$ között.

A (8.30)-ból $\mathbf{c} = \lambda\mathbf{d}$, amiből

$$\mathbf{d} = \mathbf{c}\lambda^{-1},$$

és így a (8.29) egyenletet átírhatjuk:

$$\mathbf{R}^*\mathbf{c}\lambda^{-1} = \lambda\mathbf{R}\,\mathbf{c} \quad \text{vagy} \quad \mathbf{R}^*\mathbf{c} = \lambda^2\mathbf{R}\,\mathbf{c}. \tag{8.31}$$

Mivel

$$\mathbf{R} = \mathbf{R}^* + \mathbf{U}^2,$$

$$\mathbf{R}^*\mathbf{c} = (\mathbf{R}^* + \mathbf{U}^2)\mathbf{c}\lambda^2, \tag{8.32}$$

vagy átrendezve

$$\mathbf{R}^*\mathbf{c} = \mathbf{U}^2\mathbf{c}\lambda^2/(1 - \lambda^2).$$

Az egyszerűség kedvéért legyen

$$\alpha = \frac{\lambda^2}{1 - \lambda^2},$$

így a (8.32) egyenlet

$$\mathbf{R}^*\mathbf{c} = \alpha\mathbf{U}^2\mathbf{c}. \tag{8.33}$$

Vezessük be a *kanonikus korrelációelemzésnél is alkalmazott* segédvektort:

$$\mathbf{q} = \mathbf{U}\,\mathbf{c}, \quad \text{amiből} \quad \mathbf{c} = \mathbf{U}^{-1}\mathbf{q}.$$

Ezeket behelyettesítve (8.33)-ba:

$$\mathbf{R}^*\mathbf{U}^{-1}\mathbf{q} = \alpha\mathbf{U}\,\mathbf{q},$$

amit beszorzunk balról $\mathbf{U}^{-1}$-zel:

$$\mathbf{U}^{-1}\mathbf{R}^*\mathbf{U}^{-1}\mathbf{q} = \alpha\mathbf{q}, \tag{8.34}$$

vagy más formában

$$(\mathbf{U}^{-1}\mathbf{R}^*\mathbf{U}^{-1} - \alpha\mathbf{I})\mathbf{q} = \mathbf{0}.$$

Eljutottunk tehát a sajátérték, sajátvektor problémájához. Általában m különböző megoldást kapunk, m különböző sajátvektort, amit az $(m \times m)$ típusú $\mathbf{Q}$ mátrix tartalmazzon, és legyen az ehhez tartozó m sajátérték az $\mathbf{A}$ diagonális mátrix megfelelő diagonális eleme.

Megemlítjük, hogy az $\mathbf{U}^{-1}\mathbf{R}^*\mathbf{U}^{-1}$ sajátvektorai megegyeznek $\mathbf{U}^{-1}\mathbf{R}\,\mathbf{U}^{-1}$ sajátvektoraival, az $\mathbf{U}^{-1}\mathbf{R}\,\mathbf{U}^{-1}$ sajátértékeiből pedig le kell vonni egyet, hogy az $\mathbf{U}^{-1}\mathbf{R}^*\mathbf{U}^{-1}$ sajátértékeit megkapjuk. Feltesszük, hogy a sajátvektorok normalizáltak,

$$\mathbf{Q}'\mathbf{Q} = \mathbf{I}.$$

A $\mathbf{C}$ mátrixot a $\mathbf{Q} = \mathbf{U}\,\mathbf{c}$ egyenletből kaphatjuk.

A (8.34) általánosított formája

$$\mathbf{U}^{-1}\mathbf{R}^*\mathbf{U}^{-1}\mathbf{Q} = \mathbf{Q}\,\mathbf{A}, \tag{8.35}$$

ebből az egyenletből kifejezhetjük $\mathbf{R}^*$-t:

$$\mathbf{R}^* = \mathbf{U}\,\mathbf{Q}\,\mathbf{A}\,\mathbf{Q}'\mathbf{U}. \tag{8.36}$$

Tudjuk, hogy $\mathbf{R}^* = \mathbf{S}\,\mathbf{S}'$. A (8.36)-ot könnyen hozhatjuk ilyen formára:

$$\mathbf{R}^* = \left(\mathbf{U}\,\mathbf{Q}\,\mathbf{A}^{\frac{1}{2}}\right)\left(\mathbf{A}^{\frac{1}{2}}\mathbf{Q}'\mathbf{U}\right),$$

amiből

$$\mathbf{S} = \mathbf{U}\,\mathbf{Q}\,\mathbf{A}^{\frac{1}{2}}. \tag{8.37}$$

A teljes korrelációmátrixot ($\mathbf{R}$-t) is felbonthatjuk hasonló módon

$$\begin{aligned}\mathbf{R} = \mathbf{R}^* + \mathbf{U}^2 &= \mathbf{U}\,\mathbf{Q}\,\mathbf{A}\,\mathbf{Q}'\mathbf{U} + \mathbf{U}^2 = \\ &= \mathbf{U}\,\mathbf{Q}(\mathbf{A} + \mathbf{I})\mathbf{Q}'\mathbf{U},\end{aligned}$$

és ebből

$$\mathbf{G} = \mathbf{U}\,\mathbf{Q}(\mathbf{A} + \mathbf{I})^{\frac{1}{2}}, \tag{8.38}$$

azaz

$$\mathbf{R} = \mathbf{G}\,\mathbf{G}'.$$

Hogyan értelmezzük az $\mathbf{S}$ és a $\mathbf{G}$ mátrixokat? *Az* $\mathbf{S}$ *nem más, mint az* $\mathbf{x}$ *megfigyelt változók és a* $\mathbf{w}$ *nem megfigyelt kanonikus változók közötti korrelációkat tartalmazó mátrix.*

A kovarianciamátrix – figyelembe véve, hogy $\mathbf{L}$ diagonális mátrix a λ értékeket tartalmazza, valamint $\mathbf{D} = \mathbf{C}\,\mathbf{L}^{-1}$ – a következőképpen állítható elő:

$$\begin{aligned}\frac{1}{n}\mathbf{X}'\mathbf{W} = \frac{1}{n}\mathbf{X}'\mathbf{M}\,\mathbf{D} = \mathbf{R}^*\mathbf{D} = \mathbf{R}^*\mathbf{C}\,\mathbf{L}^{-1} = \\ = \mathbf{U}\,\mathbf{Q}\,\mathbf{A}\,\mathbf{Q}'\mathbf{U}\,\mathbf{C}\,\mathbf{L}^{-1} = \\ = \mathbf{U}\,\mathbf{Q}\,\mathbf{A}\,\mathbf{Q}'\mathbf{U}\,\mathbf{U}^{-1}\mathbf{Q}\,\mathbf{L}^{-1} = \\ = \mathbf{U}\,\mathbf{Q}\,\mathbf{A}\,\mathbf{Q}'\mathbf{Q}\,\mathbf{L}^{-1} = \mathbf{U}\,\mathbf{Q}\,\mathbf{A}\,\mathbf{L}^{-1}.\end{aligned} \tag{8.39}$$

Mivel $\mathbf{w}$ nem standardizált, a kovarianciamátrix még nem egyenlő a korrelációmátrixszal. Így meg kell határoznunk a $\mathbf{w}$ variancia-kovarianciamátrixát:

$$\begin{aligned}\frac{1}{n}\mathbf{W}'\mathbf{W} = \frac{1}{n}\mathbf{D}'\mathbf{M}'\mathbf{M}\,\mathbf{D} = \mathbf{D}'\mathbf{R}^*\mathbf{D} = \mathbf{L}^{-1}\mathbf{C}'\mathbf{R}^*\mathbf{C}\,\mathbf{L}^{-1} = \\ = \mathbf{L}^{-1}\mathbf{Q}'\mathbf{U}^{-1}\mathbf{R}^*\mathbf{U}^{-1}\mathbf{Q}\,\mathbf{L}^{-1} = \\ = \mathbf{L}^{-1}\mathbf{A}\,\mathbf{L}^{-1}.\end{aligned} \tag{8.40}$$

Mivel mindhárom mátrix diagonális, szorzatuk is diagonális lesz. Így (8.39)-et oszthatjuk (8.40) négyzetgyökével, és megkapjuk a korrelációmátrixot. Tehát $\mathbf{x}$ és $\mathbf{w}$ közötti korreláció

$$\mathbf{U}\,\mathbf{Q}\,\mathbf{A}\,\mathbf{L}^{-1}\left(\mathbf{L}\,\mathbf{A}^{\frac{1}{2}}\right) = \mathbf{U}\,\mathbf{Q}\,\mathbf{A}^{\frac{1}{2}} = \mathbf{S}.$$

Hasonló módon mutatjuk meg, hogy $\mathbf{G}$ *a megfigyelt* $\mathbf{x}$ *változók és a kanonikus* $\mathbf{v}$ *változók közötti korrelációkat artalmazza.*

Ezek után nézzük meg, hogy megoldásunk kielégíti-e a (8.26) szerinti feltételeket, vagyis $\mathbf{C}'\mathbf{R}\,\mathbf{C} = \mathbf{I}$. Mivel a sajátvektorokra fennáll $\mathbf{Q}'\mathbf{Q} = \mathbf{I}$, $\mathbf{C}'\mathbf{R}\,\mathbf{C} = \mathbf{A} + \mathbf{I}$ egyenlőséget kaphatjuk, ami azt mutatja, hogy a $\mathbf{v}$ változók ($\mathbf{V} = \mathbf{X}\,\mathbf{C}$) függetlenek, de nem standardizáltak. A normalizálást természetesen könnyen elvégezhetjük. Tekintsük a $\mathbf{v}$ változó előállítását:

$$\mathbf{V} = \mathbf{X}\,\mathbf{C} = (\mathbf{M} + \mathbf{E})\mathbf{C} = \mathbf{M}\,\mathbf{C} + \mathbf{E}\,\mathbf{C}.$$

Megmutathatjuk, hogy $\mathbf{v}$ variancia-kovarianciamátrixa $\left(\frac{1}{n}\mathbf{V}'\mathbf{V}\right)$ két tagra bontható:

$$\frac{1}{n}\mathbf{V}'\mathbf{V} = \frac{1}{n}\mathbf{C}'\mathbf{X}'\mathbf{X}\,\mathbf{C} = \mathbf{C}'\left[\frac{1}{n}\mathbf{M}'\mathbf{M}\right]\mathbf{C} + \mathbf{C}'\left[\frac{1}{n}\mathbf{E}'\mathbf{E}\right]\mathbf{C} = \mathbf{C}'\mathbf{R}^*\mathbf{C} + \mathbf{C}'\mathbf{U}^2\mathbf{C}.$$

Az első tag $\mathbf{C}'\mathbf{R}^*\mathbf{C}$ a szisztematikus részt képviseli, míg a második tag egységmátrix, azaz a hibaelemek egymástól függetlenek és egységnyi hosszúságúak:

$$\mathbf{C}'\mathbf{U}^2\mathbf{C} = \mathbf{Q}'\mathbf{U}^{-1}\mathbf{U}^2\mathbf{U}^{-1}\mathbf{Q} = \mathbf{Q}'\mathbf{Q} = \mathbf{I}.$$

Tekintsük most a (8.35) egyenletet: $\mathbf{U}^{-1}\mathbf{R}^*\mathbf{U}^{-1}\mathbf{Q} = \mathbf{Q}\,\mathbf{A}$. Transzformáljuk a megfigyelt változókat az $\mathbf{U}^{-1}$ diagonális mátrix felhasználásával:

$$\mathbf{X}_S = \mathbf{X}\,\mathbf{U}^{-1}.$$

Ekkor a transzformált megfigyelt változók variancia-kovarianciamátrixa:

$$\frac{1}{n}\mathbf{X}_S'\mathbf{X}_S = \mathbf{U}^{-1}\mathbf{R}\,\mathbf{U}^{-1} = \mathbf{U}^{-1}\mathbf{R}^*\mathbf{U}^{-1} + \mathbf{U}^{-1}\mathbf{U}^2\mathbf{U}^{-1} =$$

$$= \mathbf{U}^{-1}\mathbf{R}^*\mathbf{U}^{-1} + \mathbf{I}.$$

Az első tag a szisztematikus komponensek variancia-kovarianciamátrixa, a második tag a hibakomponens, egységmátrix. A (8.35) egyenletben $\mathbf{Q}$ tehát a variancia-kovarianciamátrix szisztematikus arányának főkomponens-megoldása.

Geometriailag a megfigyelési térből áttérve a szisztematikus komponensek $\mathbf{M}$ terébe (előtte elvégezve $\mathbf{X}_S = \mathbf{X}\,\mathbf{U}^{-1}$ transzformációt) főkomponens-elemzést végzünk, főtengely-transzformációt hajtunk végre. Az eddigiekben feltételeztük, hogy $\mathbf{U}$ ismert. A gyakorlatban $\mathbf{U}^2$ mátrixot általában nem ismerjük, így azt becsülni kell. A változók kommunalitásai (h_i^2) ismeretében végezhetünk becslést.

A becsléshez használjuk most fel a kanonikus faktorelemzés megoldását, amihez Lawley dolgozta ki a maximum likelihood módszer iteratív eljárását. Induljunk ki az egyfaktoros megoldásból ($\mathbf{R}\,\mathbf{S}_{11} = \mathbf{S}_{11}\lambda_{11}$)! A korrelációmátrixot ($\mathbf{R}$) két részre bonthatjuk: $\mathbf{R} = \mathbf{R}^* + \tilde{\mathbf{R}}$, a redukált és a reziduális korrelációmátrixra ($\tilde{\mathbf{R}} = \mathbf{R} - \mathbf{S}_{11}\mathbf{S}_{11}'$). A reziduális korrelációmátrix diagonális elemei lesznek $\mathbf{U}_{11}^2$ diagonális elmei. Ez lesz $\mathbf{U}^2$ első becslése. A kanonikus faktormegoldás

$$\mathbf{U}_{11}^{-1}(\mathbf{R} - \mathbf{U}_{11}^2)\mathbf{U}_{11}^{-1}\mathbf{q}_{11} = \mathbf{q}_{11}\alpha_{11}.$$

A kanonikus faktorelemzés szerint ebből megkaphatjuk az első faktor struktúrájának következő becslését ($\mathbf{S}_{12}\mathbf{U}_{11}\mathbf{q}_{11}\alpha_{11}^{\frac{1}{2}}$). Ennek alapján újabb reziduális mátrixhoz juthatunk ($\mathbf{R} - \mathbf{S}_{12}\mathbf{S}_{12}'$) ami $\mathbf{U}^2$ következő becslését adja. Ezt a becslést ($\mathbf{U}_{12}^2$) beírva a kanonikus faktoregyenletbe, az első faktorstruktúrára újabb becslést kapunk. Ezt az eljárást folytatjuk mindaddig, míg az i-edik és $(i + 1)$-edik becslés közötti különbség már elhanyagolható lesz.

Ezután a második faktor struktúráját becsüljük. Kiszámítjuk az $\mathbf{R}_1 = \mathbf{R} - \mathbf{S}_{11}\mathbf{S}'_{11}$ reziduális mátrixot, majd e reziduális mátrix legnagyobb sajátértékéhez tartozó sajátvektort. Ez lesz a második faktor első becslése:

$$\mathbf{R}_1\mathbf{S}_{21} = \mathbf{S}_{21}\lambda_{21}.$$

Az $\mathbf{U}^2$ új becslése $\mathbf{U}^2_{21} = \text{diag}(\mathbf{R} - \mathbf{S}_{21}\mathbf{S}'_{21})$, ahol $\mathbf{S}_{21} = [\mathbf{S}_{11}, \mathbf{S}_{21}]$ két oszlopvektort tartalmaz jelenleg ($\mathbf{S}_{21}\mathbf{S}'_{21} = \mathbf{S}_{11}\mathbf{S}'_{11} + \mathbf{S}_{21}\mathbf{S}'_{21}$). A kanonikus faktoregyenlet így

$$\mathbf{U}^{-1}_{21}(\mathbf{R} - \mathbf{U}^2_{21})\mathbf{U}^{-1}_{21}\mathbf{Q}_{21} = \mathbf{Q}_{21}A_{21}.$$

A faktormátrix becslése:

$$\mathbf{S}_{22} = \mathbf{U}_{21}\mathbf{Q}_{21}A^{\frac{1}{2}}_{21}.$$

Az $\mathbf{U}^2$ következő becslése:

$$\mathbf{U}^2_{22} = \text{diag}(\mathbf{R} - \mathbf{S}_{22}\mathbf{S}'_{22}).$$

Ennek alapján a faktormátrix újabb becsléséhez juthatunk ($\mathbf{S}_{23}$). Az eljárást addig folytatjuk, amíg a reprodukció kielégítő nem lesz. A második faktort így megkapva további faktort keresünk egészen addig, amíg meg nem kapjuk a p számú faktort. A faktorok számát $(p - t)$ vagy az eljárás elején határozzuk meg, vagy χ^2-próbával az eljárás során döntjük el, hogy hány faktort veszünk figyelembe. Az $\mathbf{U}$ mátrix becslésére léteznek más eljárások is, lásd pl. C. W. Harris és C. R. Rao munkáit [70; 114].

8.4. Koncentrációs elemzés

Tételezzük fel, hogy klaszterelemzéssel az n objektumot g számú homogén csoportba soroltuk. Csoportosítsuk a változókat is ettől függetlenül a klaszterelemzés segítségével, és az m számú változót soroljuk q számú csoportba. Így az $n \times m$ típusú $\mathbf{X}$ mátrixot egy $g \times q$ típusú táblázatba rendezzük:

	A változók csoportjai 1	2	...	q	Teljes
Az objektumok csoportjai					
1	f_{11}	f_{12}	...	f_{1q}	f_{10}
2	f_{21}	f_{22}	...	f_{2q}	f_{20}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
g	f_{g1}	f_{g2}	...	f_{gq}	f_{g0}
Teljes	f_{01}	f_{02}	...	f_{0q}	f_{00}

ahol f_{ij} az ij-edik blokkba tartozó objektumok száma, f_{0j} a változók j-edik csoportjába tartozó megfigyelések száma, f_{i0} az objektumok i-edik csoportjába tartozó esetek száma, f_{00} a minta elemszáma. ($f_{00} = n$)

Azt vizsgáljuk, hogy a blokkok szignifikánsan különböznek-e. Először meghatározzuk az adjusztált gyakoriságokat, vagyis az f_{ij} gyakoriságokat egyenlő blokkméretre transzformáljuk.

Az adjusztált gyakoriságok táblázata:

$$\mathbf{F} = \begin{array}{cccc|c} F_{11} & F_{12} & \dots & F_{1q} & F_{10} \\ F_{12} & F_{22} & \dots & F_{2q} & F_{20} \\ \vdots & \vdots & & \vdots & \vdots \\ F_{g1} & F_{g2} & \dots & F_{gq} & F_{g0} \\ \hline F_{01} & F_{02} & \dots & F_{0q} & F_{00} \end{array}$$

ahol $F_{ij} = \left\{ [f_{ij}/n_{ij}] \Big/ \sum_{s=1}^{g} \sum_{z=1}^{q} f_{sz}/n_{sz} \right\} f_{00}$, n_{ij} az ij-edik blokkban a cellák teljes száma, $F_{00} = f_{00} = n$.

Az általános nullhipotézis, hogy a blokkok gyakoriságai nem különböznek a várható gyakoriságoktól:

$$H_0 \colon E(\mathbf{F}) = \mathbf{F}^0.$$

Az általános H_0 hipotézist komponensekre bontjuk:

– homogenitás a változók csoportjai között:

$$H_{01} : E(F_{0j}) = F_{0j}^0 \quad \forall\ j\text{-re},$$

ahol $F_{0j}^0 = F_{00}/q$;

– homogenitás az objektumok csoportjai között:

$$H_{02} : E(F_{i0}) = F_{i0}^0 \quad \forall\ i\text{-re},$$

ahol $F_{i0}^0 = F_{00}/g$;

– homogenitás a $g \times q$ blokk között

$$H_{03} : E(F_{ij}) = F_{ij}^0 \quad \forall\ i,\ j\text{-re},$$

ahol $F_{ij}^0 = F_{i0}F_{0j}/F_{00}$.

A nullhipotézis próbafüggvényeit a 8.4. táblázat tartalmazza. Véletlen minta esetén, és ha F_{00} elég nagy, valamint ha a nullhipotézis igaz, a próbafüggvény χ^2-eloszlást követ.

Hipotézis	Szabadságfok	Próbafüggvény (χ^2)
H_{01}	$q-1$	$2\sum_{j=1}^{q} F_{0j} \ln[F_{0j}q/F_{00}]$
H_{02}	$g-1$	$2\sum_{i=1}^{g} F_{i0} \ln[F_{i0}g/F_{00}]$
H_{03}	$(q-1)(g-1)$	$2\sum_{i=1}^{g}\sum_{j=1}^{q} F_{ij} \ln[F_{ij}F_{00}/(F_{i0}F_{0j})]$
H_0	$qg-1$	$2\sum_{i=1}^{g}\sum_{j=1}^{q} F_{ij} \ln[F_{ij}qg/F_{00}]$

8.4. táblázat.

A nullhipotézist elvetjük, ha

$$\chi^2 \geq \chi^2_{\alpha,p}$$

választott α szignifikanciaszint mellett.

PÉLDA

Induljunk ki a következő gyakorisági táblázatból:

Az objektumok csoportjai	A változók csoportjai 1	2	3	Teljes
1	2	74	16	92
2	15	21	112	148
3	71	2	35	108
Teljes	88	97	163	348

Az egyes blokkokban a cellák számát a következő táblázat tartalmazza:

$$\mathbf{N} = \begin{bmatrix} 10 \times 14 & 9 \times 14 & 7 \times 14 \\ 10 \times 20 & 9 \times 20 & 7 \times 20 \\ 10 \times 11 & 9 \times 11 & 7 \times 11 \end{bmatrix}$$

Az adjusztált gyakorisági táblázat:

$$\mathbf{F} = \begin{array}{cccc} 1{,}7 & 71{,}0 & 19{,}8 & 92{,}5 \\ 9{,}1 & 14{,}1 & 96{,}8 & 120{,}0 \\ 78{,}1 & 2{,}4 & 55{,}9 & 136{,}4 \\ \hline 88{,}9 & 87{,}5 & 172{,}5 & 348{,}9 \end{array}$$

ahol pl. $F_{11} = [(2/140)/2{,}8767]348 = 1{,}7$.

A próbafüggvény értékei:

$$H_{01} : \chi^2 = 2\{88{,}9 \ln[(88{,}9)3/348{,}9] + 87{,}5 \ln[(87{,}5)3/348{,}9+$$
$$+ 172{,}5 \ln[(172{,}5)3/348{,}9]\} = 38{,}445.$$

Mivel

$$(\chi^2 = 38{,}445) > (\chi^2_{0{,}05,\ 2} = 5{,}991),$$

a H_{01} hipotézist elvetjük, a változók csoportjai között a heterogenitás szignifikánsnak tekinthető.

$$H_{02} : \chi^2 = 8{,}6474.$$

Mivel

$$(\chi^2 = 8{,}6474) > (\chi^2_{0{,}05,\ 2} = 5{,}991),$$

a H_{02} hipotézist is elvetjük.

$$H_{03} : \chi^2 = 2\{1{,}7 \ln[(1{,}7)(348{,}9)/(92{,}5)(88{,}9)] + \ldots +$$
$$+ 55{,}9 \ln[(55{,}9(348{,}9)/(136{,}4)(172{,}5)]\} = 266{,}7,$$

Mivel

$$(\chi^2 = 266{,}7) > (\chi^2_{0{,}05,\ 4} = 9{,}488),$$

a H_{03} hipotézist is elvetjük, így a blokkok közötti heterogenitás szignifikánsnak tekinthető.

Ha a H_{03} hipotézist elvetjük (a blokkok szignifikánsan különböznek), megvizsgálhatjuk, milyen latens dimenziók felelősek a blokkok különbözőségeiért. Keressük azokat a kanonikus latens változókat $(\mathbf{x}_i, \mathbf{y}_i)$, amelyek a változók, illetve objektumok csoportjai (klaszterei) különbözőségeit a lehető legjobban magyarázzák. Ezek alapján a χ^2 értékét komponensekre bonthatjuk:

$$\chi^2 = \chi^2_1 + \ldots + \chi^2_r = R^2_1 F_{00} + \ldots + R^2_r F_{00},$$

ahol $r \leq \min(q, g)$, $R^2_1, \ldots R^2_r$ kanonikus korrelációk négyzetei.

Az i-edik kanonikus faktorpár:

$$\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{iq}]',$$

$$\mathbf{y}_i = [y_{1i}, y_{2i}, \ldots, y_{gi}]',$$

ahol $\mathbf{x}_i$ elemei az i-edik kanonikus faktorpár változócsoportokra vonatkozó értékei, $\mathbf{y}_i$ elemei az i-edik kanonikus faktorpár objektumcsoportokra vonatkozó értékei.

„r" számú különböző kanonikus faktorpár korrelációi (R_i) alapján a faktorok fontosságát mérjük az

$$L_i = R_i F_{00}/\chi^2$$

mennyiséggel.

Az L_i értékekre igaz a következő összefüggés:

$$L_1 > L_2 > \ldots > L_r.$$

A kanonikus faktorpár értékeit egymásból átszámíthatjuk:

$$x_{ih} = \sum_{j=1}^{g} F_{jh} y_{ij}/F_{0h} R_i)$$

vagy

$$y_{ji} = \sum_{h=1}^{q} F_{jh} x_{ih}/F_{j0} R_i),$$

ahol x_{ih} a kanonikus faktornak a változók h-adik csoportjára vonatkozó értéke, y_{ji} az i-edik kanonikus faktornak az objektumok j-edik csoportjára vonatkozó értéke.

A kanonikus faktorok csoportokra vonatkozó értékei két mátrixba rendezhetők:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1q} \\ x_{21} & x_{22} & \ldots & x_{2q} \\ \vdots & \vdots & & \vdots \\ x_{r1} & x_{r2} & \ldots & x_{rq} \end{bmatrix} \quad \text{és} \quad \mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & \ldots & y_{1r} \\ y_{21} & y_{22} & \ldots & y_{2r} \\ \vdots & \vdots & & \vdots \\ y_{g1} & y_{g2} & \ldots & y_{gq} \end{bmatrix}$$

Az algoritmus a következő: Keressük az $\mathbf{S} = \mathbf{U}'\mathbf{U}$ mátrix sajátértékeit és sajátvektorait, ahol $S_{sz} = \sum_{j=1}^{g} U_{js} U_{jz}$ és

$$U_{js} = F_{js}/(F_{0s} F_{j0})^{\frac{1}{2}} - (F_{0s} F_{0j})^{\frac{1}{2}}/F_{00}.$$

Az $\mathbf{S}$ mátrix sajátértékei az $R_1^2, \ldots, R_r^2$ értékek, amelyek a kanonikus korrelációk négyzetei. A sajátvektorokat az $\boldsymbol{\alpha}_i$ vektorok tartalmazzák. A kanonikus faktor értékeit $\boldsymbol{\alpha}$-ból számítjuk:

$$x_{ih} = [F_{00}/F_{0h}]^{\frac{1}{2}} \alpha_{ih},$$

ahol $\boldsymbol{\alpha}_i$ elemei kielégítik a következő feltételt:

$$\sum_{h=1}^{q} \alpha_{ih}^2 = 1,$$

$$\sum_{h=1}^{q} F_{0h}^{\frac{1}{2}} \alpha_{ih} = 0.$$

PÉLDA

Az előző példát számítjuk tovább:

$$\mathbf{S} = \mathbf{U}'\mathbf{U} = \begin{bmatrix} 0{,}258850 & -0{,}210473 & -0{,}035920 \\ & 0{,}390458 & -0{,}127538 \\ & & 0{,}117006 \end{bmatrix}$$

Az **U** egy eleme pl.

$$U_{21} = F_{21}/(F_{01}F_{20})^{\frac{1}{2}} - (F_{01}F_{20})^{\frac{1}{2}}/F_{00} =$$

$$= 9{,}1/[(88{,}9)(120{,}0)]^{\frac{1}{2}} - [(88{,}9)(120{,}0)]^{\frac{1}{2}}/348{,}9 = -0{,}2079.$$

Az **S** mátrix alapján az első két kanonikus korreláció:

Kanonikus változó	R_i	$\chi_i^2 = F_{00}R_i^2$	$L_i\%$	szabadságfok $(q-1)+(g-1)-(2i-1)$
1	0,7490	195,2	73	3
2	0,4531	71,5	27	1
		$\chi^2 = 266{,}7$	100	4

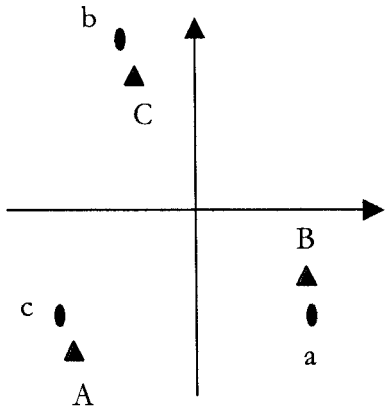
A kanonikus faktorok értékei:

	A változók csoportjai		
Kanonikus változók	1(A)	2(B)	3(C)
x_1	−1,08	1,63	−0,27
x_2	−1,32	−0,57	0,98

	Az objektumok csoportjai		
Kanonikus változók	1(a)	2(b)	3(c)
y_1	1,56	−0,15	−0,94
y_2	−0,56	1,37	−0,83

Az x_1, x_2 és az y_1, y_2 kanonikus faktorok értékeinek egyesített ábráját mutatja a 8.3. ábra. A pontábra alapján azt mondhatjuk, hogy a változók és az objektumok következő csoportjai mutatnak rokonságot egymással:

$A - c$,
$B - a$,
$C - b$.



8.3. ábra. Kanonikus faktorok tere

9. fejezet

Klaszterelemzés

A klaszterelemzés az alakfelismerés tanító nélküli tanuló algoritmusa. Egyszerűen úgy definiáljuk, hogy a klaszterelemzés megfigyelések egyedeit bontja viszonylag homogén csoportokba p változó értékeinek hasonlósága alapján. A klaszterelemzés az egyedek olyan csoportosítását keresi, amelyekre igaz, hogy egy egyed egy és csakis egy csoporthoz tartozik, és azokhoz az egyedekhez lesz hasonló, amelyekkel egy klaszterbe került, míg a többi klaszterbe tartozó egyedektől különbözik.

A klaszterelemzés túllép a klasszikus logika modelljein *egyrészt* azért, mert míg a klasszikus logika típusdefiniálása szigorúan monotetikus osztályozás, vagyis egy osztály minden eleme minden szempontból ekvivalens, addig a klaszterelemzés politetikus osztályokat definiál, amelyekben az egyedek egy vagy több jellemzőben nem feltétlenül ekvivalensek, de hasonlóak, így a sokdimenziós állapottérben is képes kevés számú csoport elkülönítésére, *másrészt* a klaszterelemzés nem definiál típusokat, mielőtt kijelölné az egyedeket, ugyanakkor a csoportosítás után megadhatja a típusjegyeket, így típusalkotásra képes.

A klasszikus logika modellje először definiálja a típusokat és utána sorolja az egyedet az osztályokhoz. Ezenkívül a klaszterelemzés nemcsak bináris változókat vehet figyelembe, így szélesebb körű lehet a matematikai modell.

A klaszterelemzést több tudományág, sokféle célból próbálta alkalmazni, ennek megfelelően különböző klasztertechnikák fejlődtek ki. Ball (1971) hét különböző lehetőséget sorol fel, ahol a klaszterelemzést sikerrel lehet alkalmazni:

1. Tipológia-alkotás
2. Modell illesztés
3. Csoportokon alapuló becslés
4. Hipotézis tesztelés
5. Adatstruktúrák felderítése
6. Hipotézis generálás
7. Adatredukció

Történetileg áttekintve az első kísérleteket a pszichológiában Zubin (1938) és Tryon (1939), az antropológiában Driver és Kroeber (1932) munkáiban találhatjuk.

A klaszterelemzés az 1950-es évek végéig számítási nehézségek miatt nem fejlődött.

Ezután a pszichológiában főleg McQuitty (1957, 1961, 1963, 1967) Pattern Analysis címen foglalkozott klasztertechnikákkal. Az antropológiában a klaszterelemzés fejlődése Driver (1965) nyomán indult meg újra. A biológiában numerikus taxonómia néven emlegetik. Összefoglaló jellegű könyv Sokal és Sneath (1963, 1973) munkája. Más tudományokban, így a közgazdaságtanban, Fischer (1969), a geográfiában Berry és Ray (1966), az irányítástudományokban Morrison (1967), Green, Frank és Robinson (1967), a politikai tudományokban Kaiser (1966), a pszichiátriában Lorr (1966), az urbanizációban Wingo (1967) munkásságát kell megemlíteni. Az utóbbi időben a matematikusok nagy száma igyekszik a klaszterelemzés technikáit általános keretbe foglalni, így különösen Jardine és Sibson (1971). Lazarsfeld latens struktúraelemzése is tulajdonképpen a klaszterelemzés egy fajtája.

A szociológusok a 60-as évekbe kezdtek foglalkozni a módszerrel; Nygreen (1969) és Dubin (1971). A szociológusok különösen a smallest space analysis Laumann (1966, 1969), Laumann és Guttman (1966), Blau és Duncan (1967) Bloombaum (1968, 1970), Guttman (1968), Nuth és Blooman (1968), Elizur (1970), Bailey (1972, 1975), McFarland és Brown (1973), és a multidimensional scaling Coleman (1957), Bloombaum (1968) és Coleman (1971)-féle megközelítésekkel kísérleteztek.

Ennyi talán elég annak illusztrálására, hogy az utóbbi időben milyen széles körben és milyen nagy számban próbálják alkalmazni a sokváltozós matematikai statisztikának ezt a fejezetét.

A klaszterező módszereket sokféle szempontból szokták típusokba sorolni. Everitt a technikákat öt típusba sorolja. Hierarchikus technikák – optimalizáló technikák – sűrűség-kereső technikák – átfedő technikák – egyéb technikák.

Bailey tizenkettő kritériumot ad az osztályozáshoz:

1. Agglomeratív–divizív
2. Monotetikus–politetikus
3. Természetes–mesterséges
4. A klaszterek száma előre meghatározott–nem meghatározott
5. Egyszintű–hierarchikus
6. Átfedő–egymást kizáró
7. Outliert megengedő–nem megengedő
8. A kapcsolás típusa szerint: egyszerű lánc–teljes lánc–átlagos lánc
9. Kombinatorikus–nem kombinatorikus
10. Teljes–nem teljes
11. Iteratív–nem iteratív
12. Klaszterező–átfedő (clumping)

9.1. A klaszterelemzés helye az alakfelismerés statisztikus módszerei között

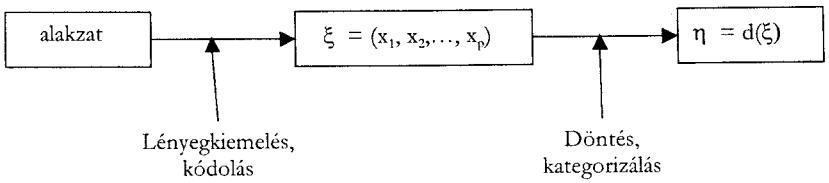
Az alakfelismerés feladata bizonyos egyedek osztályokba sorolása. Az egyedek mindig valamilyen mérhető halmaz elemei. Lehetnek személyek, a társadalom valamilyen alrendszerének egyedei stb., de lehetnek maguk a rendszer jellemzői is. Az egyedek osztályba sorolásán a mérhető halmaz (X) olyan felbontását értjük, hogy a kapott részhalmazok teljes rendszert alkossanak, vagyis a csoportok diszjunktak (egymást kizáróak) legyenek, és együttesen a teljes halmazt adják.

$$C_i \cap C_j = \emptyset \qquad \text{minden } i, j\text{-re}, (i \neq j)$$

$$\bigcup_{i=1}^{\infty} C_i = X$$

Az osztályozást mindig valamilyen döntésfüggvény (d) alapján végezzük el. Minden döntésfüggvény generál egy felbontást és minden felbontás generál egy döntésfüggvényt, vagyis egy osztályozást, tehát az osztályozás és felbontás megfeleltethetők egymással.

Az alakfelismerés folyamata két fő mozzanatra bontható:
1. lényegkiemelésre és
2. kategorizálásra.



9.1. ábra. Az alakfelismerés folyamata

9.2. Lényegkiemelés

A lényegkiemelés az egyedeknek a leképezését jelenti valamilyen jól kezelhető kódba, vagyis az egyedeknek

az $\mathbf{x} = [x_1, x_2, \ldots, x_p]'$ vektorral történő kódolása.

Ez a mozzanat több lényeges kérdést vet fel.

9.2.1. A jellemzők szelektálása

Az egyedek osztályozását az egyedek, jellemző változók alapján végezhetjük el. Mivel ezek száma végtelennek tekinthető, a kutatási célok szempontjából lényeges változókat kell kiválasztani, és törekedni kell arra, hogy a vizsgált kutatási terület minden fontosnak ítélt részét jellemezzék.

Az egyedeket jellemző változókat jelölje, ahol $[J = J_1, J_2, \ldots, J_p]$. A klaszterezés elvégzéséhez minden egyedre rendelkezésre kell állni a p jellemző megfigyelt értékének. Az x_{ij} jelöli az i-edik egyed j-edik jellemzőjének értékét.

A megfigyelt változók legtöbbször nem egyeznek meg elméleti változóikkal, valószínűségi változók lévén a véletlen hatást is magukban hordozzák. Feltételezésünk szerint a véletlen tényező mellett a szisztematikus hatást mutató rész egyezik meg azzal a jellemzővel, jelenséggel, amit mérni akartunk. Ez a szisztematikus tag azonban nagyon gyakran tovább bontható additív jellemző komponensekre. Feltételezésünk szerint tehát a megfigyelt változóinkat meghatározó szisztematikus komponensek, faktorok azonosíthatók elméleti változóinkkal, így célszerű lehet a megfigyelt jellemzőkről áttérni az őket meghatározó lényegi dimenziójegyeket tartalmazó faktorokra.

A változók választásánál különösen figyelni kell, hogy az egyedeket csak a megfelelően diszkrimináló változók jellemezzék. Az alakfelismerést ugyanúgy megzavarhatja, ha túlságosan instabil egy változó, vagyis indokolatlanul sok csoport kialakítását indukálhatja, vagy ha a változó konstans, nem különbözik az egyedeknél. Ezeket a változókat el kell hagyni.

A változók számának ésszerű csökkentésére ösztönöz az alakfelimerési algoritmusok számítógépes programjainak nagy memóriaigénye.

9.2.2. A jellemzők mérése

Amikor a változókat megfigyeljük és megmérjük az egyedeken, minden egyedhez kódokat rendelünk, $\mathbf{x}_i = [x_1, x_2, \ldots, x_p]'$. Az egyedekhez rendelt kódoktól természetszerűen elvárjuk, hogy az egyedek között meglévő valóságos relációt tükrözzék. Attól függően, hogy a hozzárendelt kódok (a gyakorlatban legtöbbször számok) között milyen tulajdonságokat tekintünk érvényesnek a változók kategóriái között fennálló relációnak megfelelően, a változók nominális, ordinális, intervallum- és aránymérési skáláját különböztetjük meg. A gyakorlatban legtöbbször az egyedeket különböző skálatípuson mért változókkal jellemezzük, ugyanakkor az egyedek összehasonlításakor feltételezzük a változók azonos mérési skáláját. Ezt a problémát a következőképpen oldhatjuk meg:

a) skálatranszformációt hajtunk végre. Ebben az esetben kiválasztunk egy domináns skálatípust, és az ettől eltérő mérési változókat transzformáljuk a kívántra.

b) Dichotomizálunk. A dichotom változók kezelésének rugalmasságát kihasználva a nem megfelelő típusú változókat dichotomizáljuk. Legyen egy változónak k különböző kategóriája. Ebben az esetben az adott változót $(k-1)$ dichotom változóval helyettesíthetjük úgy, hogy az i-edik dichotom változó 1 értéket vesz fel, ha az egyed a változó i-edik kategóriájába esik, 0-t különben.

c) Amennyiben több domináns skálatípusunk van, az elemzést a skálatípus szerint bontott változóhalmazra külön-külön elvégezhetjük, majd az eredményeket egységesítjük.

Azonos típusú mérési skálák esetén is probléma, hogy a változóknak különböző lehet a mértékegységük. Ez lehetetlenné teszi bizonyos hasonlóság vagy távolság mértékek használatát.

A különböző mértékegységű változók problémájának megoldása lehet, ha

1. – azonos mértékegységre számítjuk át a változókat;
2. – minden változót standardizálunk 0 várható értékű és 1 szórású változóvá.

9.2.3. A hasonlóság mérése

A jellemzők kiválasztása a vizsgálat szempontjából lényegesnek ítélt tulajdonságok számbavételét, e tulajdonságokhoz mérési skálával ellátott változók hozzárendelését jelenti. Bár tulajdonság lehet bármely ismérv, mennyiségi érték, minőségi állapot, földrajzi megjelölés stb., mégis a tulajdonságokhoz rendelt változó értékét tekintjük változónak. A változók típusától függnek elsősorban a kapcsolatok mérésének módszerei.

A hasonlósági mérőszámok általános (de nem minden esetben érvényes) tulajdonságai a következő formában írhatók fel: ha s_1, s_2 két összehasonlítandó objektum, és $A(s_1, s_2)$ a hasonlósági mérőszám, akkor

1. $A(s_1, s_2) = A(s_2, s_1)$ (szimmetria),
2. A értéke általában a $0 \leq A \leq 1$ vagy a $-1 \leq A \leq 1$ intervallumba esik,
3. $A(s_1, s_1) = 1$.

A páronként mért hasonlósági értékeket mátrixba rendezve szimmetrikus és a főátlójában egyeseket tartalmazó $\mathbf{A}$ mátrixot kapunk.

Ha adatmátrixunk $(n \times p)$ méretű, azaz n megfigyelési egységre p jellemző értékkel rendelkezünk, akkor az egyedek (a mátrix sorainak) összehasonlításával kapott $\mathbf{A}$ mátrix

($n \times n$) méretű. Ha a változók (oszlopok) páronkénti hasonlóságát tartalmazza az **A** mátrix, akkor mérete ($p \times p$).

A változópárok hasonlóságának mérésekor tekintettel kell lennünk a változók mérési skálájára, ezért a mutatókat a változók mérési szintje szerinti csoportosításban ismertetjük.

Nominális és ordinális változók hasonlósága

A mérés alapja a statisztikából ismert kereszttábla,

$A \setminus B$	1	2	...	q	összeg
1	f_{11}	f_{12}	...	f_{1q}	f_{10}
2	f_{21}	f_{22}	...	f_{2q}	f_{20}
⋮	⋮		⋮	⋮	
r	f_{r1}	f_{r2}	...	f_{rq}	f_{r0}
összeg	f_{01}	f_{02}	...	f_{0q}	n

ahol f_{ij} az i és j tulajdonság együttes előfordulásának – az n elemű mintából számított – gyakorisága.

A hasonlósági mértékek többsége a Pearson-féle χ^2-statisztikára épül, amely a változók közötti függetlenséget tételezi fel:

$$\chi^2 = n\left(\sum_{i=1}^{r}\sum_{j=1}^{q}\frac{f_{ij}^2}{f_{i0}f_{0j}} - 1\right) \tag{9.1}$$

Látható, hogy a χ^2 értéke közvetlenül függ a táblázat méretétől, és az n növekedésével minden határon túl nő. Ezért különféle normált értékeket célszerű használni. Pearson a P, Csuprov a T, Cramer pedig a V kontingenciaegyütthatót javasolta a hasonlóság mérésére.

$$P = \left(\frac{\phi^2}{1+\phi^2}\right)^{\frac{1}{2}}, \quad \text{ahol} \quad \phi^2 = \frac{\chi^2}{n};$$

$$T = \left[\frac{\chi^2}{n(r-1)(q-1)}\right]^{\frac{1}{2}},$$

$$V = \left\{\frac{\chi^2}{n \cdot \min[(r-1),(q-1)]}\right\}^{\frac{1}{2}}. \tag{9.2}$$

Kendall és Stuart mutatott rá a χ^2-statisztikán alapuló mértékek torzító hatásainak okaira. E mértékek arra a hipotézisre épülnek, hogy a kontingenciatáblázat olyan kétváltozós normális eloszlást reprezentál, amelyre teljesül, hogy

$$\lim_{n\longrightarrow\infty} P^2 = r^2$$

ahol r a korrelációs együttható.

A gyakorlatban ez a feltevés általában nem teljesül, így ezek a mértékek csak korlátozottan alkalmasak az asszociáció mérésére. Másik hiányosságuk az, hogy e mértékek alapján a változópárok egymás között nem összehasonlíthatók.

Erre rámutatva Goodman és Kruskal javasolta a Γ-statisztika bevezetését. Ez az asszociációs mérték az optimális osztály becslésén alapul.

A Γ-statisztika nominális változók esetén

Ha a kontingenciatáblázat minden elemét n-nel elosztjuk, a relatív gyakoriságokat kapjuk, amelyek n növelésével jól közelítik a megfelelő valószínűségeket. Indokolt tehát a következő jelölések bevezetése:

$$p_{ij} = \frac{f_{ij}}{n}\,; \quad p_{0j} = \frac{f_{0j}}{n}\,; \quad p_{i0} = \frac{f_{i0}}{n}. \tag{9.5}$$

Ha kiválasztunk az n elemű sokaságból véletlenszerűen egy elemet, és a lehető legkisebb hibával becsüljük, hogy melyik tartozik A_i, illetve B_j ismérvosztályba, akkor két feltételezéssel élhetünk:

1. csak azt tudjuk a kiválasztott elemről, hogy besorolható valamelyik két osztályba,
2. ismerjük a kiválasztott elem A_i osztályát.

Nyilvánvaló, hogy az utóbbi esetben több információnk van, az elkövetett hiba legfeljebb akkora lehet, mint az első esetben. Legyen

P_1 a besorolás hibájának valószínűsége az 1. esetben,

P_2 a besorolás hibájának valószínűsége a 2. esetben.

Ekkor a Γ_B asszociációs mérőszámot így definiáljuk:

$$\Gamma_B = \frac{P_1 - P_2}{P_1} \tag{9.6}$$

Γ_B a besorolási hibavalószínűség csökkenésének relatív értékét mutatja, amely az A_i osztály ismeretéből adódó információtöbbletből ered. Ha bevezetjük a $p_{0m} = \max\limits_j p_{0j}$ és a $p_{im} = \max\limits_j p_{ij}$ jelöléseket,

akkor $P_1 = 1 - p_{0m}$ és $P_2 = 1 - \sum\limits_{i=1}^{r} p_{im}$,

$$\Gamma_B = \frac{\sum\limits_{i=1}^{r} p_{im} - p_{0m}}{1 - p_{0m}}. \tag{9.7}$$

Ha nem az A_i, hanem egy B_j osztály azonosítható a véletlenszerűen kiválasztott elemmel, akkor a Γ_A hasonlósági mérőszám a fentivel azonos módon definiálható:

$$\Gamma_A = \frac{\sum\limits_{j=1}^{q} p_{mj} - p_{m0}}{1 - p_{m0}}. \tag{9.8}$$

Ezt a gondolatmenetet akkor is megismételhetjük, ha az A és a B ismérvváltozatok közötti kapcsolatoknak nincs kitüntetett iránya. Tehát egy tetszőleges elemet $1/2$ valószínűséggel az A_i vagy a B_j osztályba tudjuk sorolni. Ekkor ismeretlen besorolás esetén

$$P_1 = 1 - \frac{1}{2}(p_{0m} + p_{m0}),$$

és ismert besorolás esetén a hiba valószínűsége

$$P_2 = 1 - \frac{1}{2}\left(\sum_{i=1}^{r} p_{im} + \sum_{j=1}^{q} p_{mj}\right).$$

Az asszociációs mutató értéke

$$\Gamma = \frac{\frac{1}{2}\left(\sum_i p_{im} + \sum_j p_{mj} - p_{m0} - p_{0m}\right)}{1 - \frac{1}{2}(p_{0m} + p_{m0})}, \tag{9.9}$$

a $\Gamma_A \leq \Gamma \leq \Gamma_B$ intervallumba esik.

Az asszociációs mutató tulajdonságai a következők:

a) Γ akkor és csak akkor nem határozható meg, ha az egész sokaság egy osztályba tartozik, egyébként $0 \leq \Gamma \leq 1$;

b) $\Gamma = 1$ akkor és csak akkor, ha A_i ismerete egyértelműen definiálja a megfelelő B_j osztályt, azaz függvényszerű kapcsolat van a két változó között.

c) $\Gamma = 0$, ha a vizsgált osztályok statisztikailag függetlenek (nem megfordítható állítás);

d) Γ invariáns a kereszttábla sorainak (vagy oszlopainak) permutációjára.

A klaszterelemzésben a nominális változók közötti kapcsolatok jellemzésére jól használhatók még a kanonikus korreláción és az információelméleten alapuló mértékek is. (Részletesebben az 5. fejezetben tárgyaljuk.)

Γ-*statisztika ordinális változók esetén*

Most az A és a B ismérvváltozatok közül legalább az egyik természetes módon rendezhető. Így a kereszttábla sorainak és/vagy oszlopainak permutációjára Γ nem lehet invariáns. Válasszunk ki a sokaságból véletlenszerűen (visszatevéssel) két elemet! Tegyük fel, hogy az első $(A_{i_1}; B_{j_1})$, a második pedig $(A_{i_2}; B_{j_2})$ kategóriába tartozik, ahol $1 \leq i_k \leq r$ és $1 \leq j_k \leq q \quad (k = 1,2)$. Függetlenség esetén joggal várhatjuk, hogy az i_k indexek rendezettsége nincs összefüggésben a j_k indexek rendezettségével, míg kapcsolat esetén ez a rendezettség általában megegyezik. Jelöljük *a hasonló rendezettség valószínűségét* P_h-val:

$$P_h = P\{i_1 < i_2 \text{ és } j_1 < j_2 \text{ vagy } i_1 > i_2 \text{ és } j_1 > j_2\},$$

az eltérő rendezettség valószínűségét P_e-vel:

$$P_e = P\{i_1 < i_2 \text{ és } j_1 > j_2 \text{ vagy } i_1 > i_2 \text{ és } j_1 < j_2\},$$

valamint *az azonosság valószínűségét* P_a-val:

$$P_a = P\{i_1 = i_2 \text{ és } j_1 = j_2\}.$$

Az egyértelműség kedvéért az utóbbi esetet a mérőszám definiálásakor nem engedjük meg, vagyis P_h és P_e helyett az $\{i_1 = i_2$ vagy $j_1 = j_2\}$ esemény ellentettjére vonatkozó feltételes valószínűségeket tekintjük, pl. P_h helyett a $P_h/(1 - P_a)$ valószínűséget. Az asszociációs mutató:

$$\Gamma = \frac{P_h - P_e}{1 - P_a}. \tag{9.10}$$

A Γ mutató tulajdonságai a következők:

a) Γ nem határozható meg, ha a kontingenciatáblázat nem nulla elemei egy sorban vagy egy oszlopban vannak;

b) $-1 \leq \Gamma \leq 1$;

c) $\Gamma = 1$, ha a nem nulla elemek a $p_{11} \longrightarrow p_{rq}$ irányú átlóban vannak, ekkor $P_e = 0$;

d) $\Gamma = -1$, ha a nem nulla elemek a $p_{r1} \longrightarrow p_{1q}$ irányú átlóban vannak, ekkor $P_h = 0$;

e) $\Gamma = 0$, ha teljesül a függetlenség, ez az állítás nem megfordítható (kivétel a 2×2-es tábla).

Arány- és intervallumváltozók

Az n sorból és p oszlopból álló $\mathbf{X}$ mátrix elemei intervallum- vagy arányskálán mérhető értékek, feladatunk két tetszőleges oszlop hasonlóságának a mérése. Jelöljük az $\mathbf{X}$ mátrix két oszlopát $\mathbf{x}$ és $\mathbf{y}$ vektorokkal, $\mathbf{x}, \mathbf{y} \in R^n$. A két változó közötti hasonlóság mértékének a vektorok hajlásszöge és a korrelációs együttható tekinthetők:

$$A(\mathbf{x}, \mathbf{y}) = \cos\alpha = \frac{\mathbf{x}^*\mathbf{y}}{|\mathbf{x}||\mathbf{y}|}, \tag{9.11}$$

ahol α az $\mathbf{x}$ és $\mathbf{y}$ vektorok hajlásszöge, $\mathbf{x}, \mathbf{y} \neq \mathbf{0}$.

Képezzük az $\widehat{\mathbf{x}} = \mathbf{x} - \overline{\mathbf{x}}$ és $\widehat{\mathbf{y}} = \mathbf{y} - \overline{\mathbf{y}}$ nulla átlagú vektorokat! E vektorok korrelációs együtthatója:

$$r = r(\mathbf{x}, \mathbf{y}) = \frac{\mathrm{cov}(\widehat{\mathbf{x}}, \widehat{\mathbf{y}})}{\sqrt{\mathrm{var}(\widehat{\mathbf{x}}) \cdot \mathrm{var}(\widehat{\mathbf{y}})}}, \tag{9.12}$$

ahol $\mathrm{cov}(\widehat{\mathbf{x}}, \widehat{\mathbf{y}}) = \dfrac{\widehat{\mathbf{x}}^*\widehat{\mathbf{y}}}{n}$ és $\mathrm{var}(\widehat{\mathbf{x}}) = \dfrac{\widehat{\mathbf{x}}^*\widehat{\mathbf{x}}}{n}$.

Könnyen belátható, hogy $r(\mathbf{x}, \mathbf{y}) = A(\widehat{\mathbf{x}}, \widehat{\mathbf{y}})$, és $A(\widehat{\mathbf{x}}, \widehat{\mathbf{y}})$ invariáns a nyújtásra, $r(\mathbf{x}, \mathbf{y})$ pedig a nyújtásra és az eltolásra. Ebből adódik, hogy $A(\widehat{\mathbf{x}}, \widehat{\mathbf{y}})$ az arány, $r(\mathbf{x}, \mathbf{y})$ pedig az intervallumváltozók esetén alkalmazható eredményesen a hasonlóság mérésére.

Bináris változók hasonlósága

Sajátos tulajdonságuk miatt célszerű a bináris változókat külön kiemelni, mert

- az előző formuláknak bináris esetre általában létezik egyszerűbb alakjuk,
- a tulajdonságok asszociációs mérőszámai, bináris változók esetén, alkalmazhatók az objektumok összehasonlítására is.

A $\mathbf{T}$ mátrix most csak 0 és 1 számokat tartalmaz. Két tetszőleges oszlop összehasonlítása nyilvánvalóan minden esetben redukálható egy 2×2-es táblára:

$A \setminus B$	1	0	összesen:
1	a	b	$a+b$
0	c	d	$c+d$
összesen:	$a+c$	$b+d$	$a+b+c+d=n$

Példaként a már bevezetett Csuprov-együttható bináris alakját mutatjuk be:

$$A_{cs} = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}. \tag{9.13}$$

A bináris mérőszámok egyik problémáját a d érték figyelembevétele jelenti, ez ugyanis a közös tulajdonságok hiányát méri, és nem jelent feltétlenül hasonlóságot. A másik probléma az, hogy hogyan súlyozzuk az illeszkedéseket és a nemilleszkedéseket. A mutatókat a felvetett problémák szerint a 9.1. táblázat foglalja össze.

Súlyozás	0–0 illeszkedés a nevezőben	0–0 illeszkedés a számlálóban	
		nem szerepel	szerepel
Egyenlő súlyok	szerepel	1. Russel és Rao: $\frac{a}{a+b+c+d}$	2. Sokal és Michner: $\frac{a+d}{a+b+c+d}$
	nem szerepel	3. Jaccard: $\frac{a}{a+b+c}$	4. –
Dupla súlyozás a kapcsolódó pároknál	szerepel	5. (nem ajánlott) $\frac{2a}{2(a+d)+b+c}$	6. $\frac{2(a+d)}{2(a+d)+b+c}$
	nem szerepel	7. Dice: $\frac{2a}{2a+b+c}$	8. –
Dupla súlyozás nem kapcsolódó pároknál	szerepel	9. –	10. Rogers–Tanimoto: $\frac{a+d}{a+d+2(b+c)}$
	nem szerepel	11. $\frac{a}{a+2(b+c)}$	12. –
A kapcsolódó párok kizárva a nevezőből	–	13. Kulczynski: $\frac{a}{b+c}$	14. $\frac{a+d}{b+c}$

9.1. táblázat. Bináris asszociációs mérőszámok rendszere

Egyedek hasonlósága nominális változók terében

Az ismérvek közötti hasonlóság mérésére alkalmazott mutatók a megfigyelések közötti hasonlóság mérésére is alkalmasak. Ekkor az összehasonlítás azon alapul, hogy összeszámláljuk a közös és az eltérő ismérveket. Ha az ismérvek jelenléte vagy hiánya egyértelműen megállapítható, akkor a bináris változóknál bevezetett 2×2-es táblához jutunk. Előfordulhat, hogy egyes ismérvek nem jellemzőek a kérdéses megfigyelési egységre. Ezt figyelembe véve a két lehetséges alternatíva (0 és 1) helyett hármat bevezetve, a bináris változók kiterjesztéséről beszélhetünk. Legyen

n_{a+d} azon ismérvek száma, amelyekben a két egyed megegyezik,

n_d azon ismérvek száma, amelyekkel a két egyed nem jellemezhető,

n_{b+c} azon ismérvek száma, amelyekben a két egyed eltér egymástól.

A 9.1. táblázat képleteit felhasználva megkapjuk a 9.2. táblázatban összefoglalt kapcsolódási együtthatókat (a formulák sorszáma a 9.1. táblázat megfelelő számozására utal).

Súlyozás	0–0 illeszkedés a nevezőben	0–0 illeszkedés a számlálóban	
		nem szerepel	szerepel
Egyenlő súlyok	szerepel	1. $\frac{n_{a+d} - n_d}{n_{a+d} + n_{b+c}}$	2. $\frac{n_{a+d}}{n_{a+d} + n_{b+c}}$
	nem szerepel	3. $\frac{n_{a+d} - n_d}{n_{a+d} - n_d + n_{b+c}}$	4. –
Dupla súlyozás a kapcsolódó pároknál	szerepel	5. –	6. $\frac{2n_{a+d}}{2n_{a+d} + n_{b+c}}$
	nem szerepel	7. $\frac{2(n_{a+d} - n_d)}{2(n_{a+d} - n_d) + n_{b+c}}$	8. –
Dupla súlyozás nem kapcsolódó pároknál	szerepel	9. –	10. $\frac{n_{a+d}}{n_{a+d} + 2n_{b+c}}$
	nem szerepel	11. $\frac{n_{a+d} - n_d}{n_{a+d} - n_d + 2n_{b+c}}$	12. –
A kapcsolódó párok kizárva a nevezőből	–	13. $\frac{n_{a+d} - n_d}{n_{b+c}}$	14. $\frac{n_{a+b}}{n_{b+c}}$

9.2. táblázat. Kapcsolódási együtthatók nominális változók esetén

9.2.4. A távolság metrikus mértékei

A klaszterelemzés gyakorlati alkalmazásánál az egyik központi probléma a pontok, illetve ponthalmazok közötti távolság definiálása. A távolság megfelelő megválasztása legalább olyan körültekintést igényel, mint a megfelelő osztályozó algoritmus kiválasztása.

Tekintsük most az $\mathbf{X}$ $(n \times p)$-s adatmátrix sorait, mint megfigyeléseket egy-egy pontnak a p dimenziós térben.

E pontok között értelmezhetők metrikus tulajdonsággal rendelkező távolságmérő függvények. Jelöljük az $(\mathbf{x}, \mathbf{y})$ pontpár távolságát $d(\mathbf{x}, \mathbf{y})$-nal, amely minden $\mathbf{x}, \mathbf{y}, \mathbf{z} \in R^p$ esetén a következő tulajdonságokkal rendelkezik:

1. $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$,
2. $d(\mathbf{x}, \mathbf{x}) = 0$,
3. $d(\mathbf{x}, \mathbf{y}) > 0$, ha $\mathbf{x} \neq \mathbf{y}$,
4. $d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y})$.

Ezek a metrikus tér általános tulajdonságai, és az ezeket kielégítő $d(\mathbf{x}, \mathbf{y})$ függvényt metrikus függvénynek vagy röviden *metrikának* nevezzük. Ha a 3. feltétel nem teljesül, akkor d-t *pszeudometrikának* nevezzük.

Ha a 4. tulajdonság, az ún. háromszög-egyenlőtlenség nem teljesül, ezt az esetet *szemimetrikának* nevezzük. Azon metrikákat, amelyek kielégítik a $d(\mathbf{x}, \mathbf{y}) \leq \leq \max[d(\mathbf{x}, \mathbf{z}); d(\mathbf{z}, \mathbf{y})]$ egyenlőtlenséget, *ultrametrikáknak* nevezzük.

A pontpárok távolságából felírható a szimmetrikus, a főátlóban zérusokat tartalmazó D távolságmátrix.

Egy példán keresztül megmutatjuk, hogy milyen problémát okoz, ha nem teljesül a 4. tulajdonság.

Tegyük fel, hogy 5 pontunk van, az i-edik és a j-edik távolságát jelöljük d_{ij}-vel! Távolságaink legyenek

$$d_{12} = 2 \qquad d_{23} = 10 \qquad d_{15} = 1 \qquad d_{35} = 1{,}5$$
$$d_{13} = 10 \qquad d_{24} = 10 \qquad d_{25} = 100 \qquad d_{45} = 100$$
$$d_{14} = 10 \qquad d_{34} = 2$$

Az első két oszlop alapján két jól elkülöníthető osztályt kapunk, az $S_1 = \{\mathbf{x}_1, \mathbf{x}_2\}$ és az $S_2 = \{\mathbf{x}_3, \mathbf{x}_4\}$ osztályokat, de hova soroljuk az $\mathbf{x}_5$ pontot? Ha S_1-hez vesszük, akkor $\mathbf{x}_2$-től távolabb lesz, mint $\mathbf{x}_3$-tól, pedig az előbbivel egy osztályba tartozik, de ugyanígy ésszerűtlen S_2-be is sorolni, mert akkor $\mathbf{x}_1$-hez lesz aránytalanul közelebb, mint $\mathbf{x}_4$-hez. Marad még egy lehetőség, $\mathbf{x}_5$ külön osztályt alkot, de ez sem kielégítő megoldás, mert $\mathbf{x}_1$-től is és $\mathbf{x}_3$-tól is kisebb távolságra van, mint a velük egy osztályba tartozó $\mathbf{x}_2$, illetve $\mathbf{x}_4$ pontok.

A Minkowski-metrika és speciális esetei

Az egyik legáltalánosabb metrikaosztály, amely tetszőleges $1 \leq r < \infty$ értékek mellett egy-egy metrikát ad, a következőképpen definiálható:

$$d_r(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{p} |x_i - y_i|^r \right)^{\frac{1}{r}}. \tag{9.14}$$

A Minkowski-metrika $r = 2$ esetén az ismert euklideszi távolsággal azonos:

$$d_2(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}, \tag{9.15}$$

$r = 1$ esetén a távolság a koordinátánkénti eltérések[1] abszolút értékeinek összege:

$$d_1(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{p} |x_i - y_i|. \tag{9.16}$$

A klaszterelemzésben leggyakrabban e két utóbbi mértéket szokás alkalmazni.

Az ismertetendő klaszterelemző módszerek szempontjából közömbös, hogy az adott objektumok között távolságot vagy hasonlóságot értelmezünk, csak arra kell ügyelni, hogy a minimális távolság maximális hasonlóságnak felel meg és fordítva.

Ha egy, a pontpárok távolságán értelmezett monoton csökkenő függvényt adunk meg, amelynek értékei 0 és 1 közé esnek, akkor a metrikához egy hasonlóságot rendelünk hozzá. A legegyszerűbb hasonlósági mérték a (9.12) szerinti korrelációs együttható, illetve ennek a $(0; 1)$ tartományba transzformált értéke:

$$r' = (1 + r)/2 \tag{9.17}$$

Két pont hasonlóságát méri a (9.11) alapján felírt vektorok hajlásszögének koszinusza is, amelyből

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{1 - \cos^2(\mathbf{x}, \mathbf{y})} \tag{9.18}$$

képlettel származtatható távolság, ha $\mathbf{x}, \mathbf{y} \neq \mathbf{0}$.

[1] City-blokk vagy Manhatten távolságként is említik.

A $d(\mathbf{x}, \mathbf{y})$ ebben az esetben pszeudometrika, vagyis a 3. feltétel itt nem teljesül. Két pont távolsága akkor lesz 0, ha a vektorok egy egyenesre esnek, így $\mathbf{x} \neq \mathbf{y}$ pontok távolsága is lehet 0.

A távolságmértékekből mindig képezhető hasonlóság, például az alábbi képlettel:

$$A(\mathbf{x}, \mathbf{y}) = \frac{1}{1 + d(\mathbf{x}, \mathbf{y})}.$$

Így biztosan teljesül a $0 < A(\mathbf{x}, \mathbf{y}) \leq 1$ feltétel.

Fontos azonban megjegyezni, hogy a hasonlósági mértékből számított távolság csak akkor elégíti ki az 1–4. feltételeket, ha az $A(\mathbf{x}, \mathbf{x}) = 1$ teljesül, és a hasonlósági mátrix nemnegatív definit. Ekkor a távolság értéke:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{2(1 - A(\mathbf{x}, \mathbf{y}))}$$

Bináris változók esetében – az eddigieken kívül – más távolságot is használhatunk. A kétdimenziós tábláknál szokásos jelölésekkel

$$d_{ij} = \frac{b + c}{a + b + c + d} \tag{9.19}$$

egy lehetséges metrika. De mérhető a pontok távolsága

$$d_{ij} = 1 - \frac{2a}{2a + b + c} \tag{9.20}$$

értékkel is. Ez utóbbi metrikánál alaposan mérlegelni kell a klaszterezés alkalmazási szempontjait, mert a dichotom változók 0 és 1 értékeinek megválasztása gyakran esetleges (pl. a nemeknél férfi=0, nő=1 vagy fordítva), ez pedig azt jelenti, hogy a jelölésektől függően más lesz a pontok távolsága. Ha pl.

$$x = (1,\ 0,\ 0,\ 0,\ 0,\ 0) \quad \text{és} \quad y = (1,\ 1,\ 1,\ 0,\ 0,\ 0),$$

akkor $a = 1$, $b = 0$, $c = 2$, $d = 3$ alapján (9.20) szerint: $d_{ij} = \frac{1}{2}$. Ha a változók értékeiben felcseréljük a nulla és egy jelölést, akkor

$$x = (0,\ 1,\ 1,\ 1,\ 1,\ 1) \quad \text{és} \quad y = (0,\ 0,\ 0,\ 1,\ 1,\ 1),$$

vagyis $a = 3$, $b = 2$, $c = 0$, $d = 1$, és ennek megfelelően: $d_{ij} = \frac{3}{4}$.

A korábban tárgyalt metrikáknál ilyen probléma nincs, ott a távolság független a változók értékeihez rendelt számoktól.

A Minkowski-metrika alkalmazásakor figyelemmel kell lenni arra, hogy a mérőszám a változók közötti függetlenséget tételezi fel. Így előfordulhat, hogy a változók közötti kapcsolatok esetén egyfajta hatást többszörösen veszünk figyelembe.

Ha ismerjük adatrendszerünk *valószínűségeloszlását*, vagy a minta elég nagy ahhoz, hogy a variancia-kovarianciamátrixot ($\mathbf{S}$) kielégítő pontossággal becsüljük, akkor pontjaink távolságát a Mahalanobis-metrikával mérhetjük:

$$d_M(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^* \cdot \mathbf{S}^{-1} \cdot (\mathbf{x} - \mathbf{y})} \tag{9.21}$$

Ebben az esetben a távolság nemcsak az $\mathbf{x}$ és $\mathbf{y}$ pontok koordinátáitól függ, hanem az $\mathbf{S}$ inverzén keresztül az összes többi ponttól is. A variancia-kovarianciamátrix felhasználásával figyelembe vesszük a változók kapcsolatát is. Ha *a változók páronként korrelálatlanok*, akkor $\mathbf{S}$ diagonális mátrix, és inverzének főátlójában az egyes változók

szórásnégyzetének a reciproka áll. Ha erre bevezetjük a $w_j = 1/s_j^2$ jelölést, akkor a Mahalanobis-távolság egyszerűbb alakban is felírható:

$$d_M(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{j=1}^{p} w_j (x_j - y_j)^2} \tag{9.22}$$

Az egyes *változók súlyozását* valamennyi metrikánál el kell végezni, de annak megítélése, hogy milyen súlyrendszert alkalmazzunk, elsősorban a kutató feladata. Az **X** mátrix két oszlopának (két ismérv) összehasonlításakor két, egyenként homogén koordinátájú vektorunk van, amíg két sor (két egyed) összevetésekor az egyes koordináták gyakran különböző tartalmú és típusú változókat reprezentálnak. Ebből adódik, hogy a mértékeket befolyásolják a nagyságrendek, és felmerül a különböző mértékek additivitásának kérdése is. E problémák megoldására szolgál a súlyozás. Ha a vizsgálat körébe bevont változók nem egyformán fontosak az adott kérdés szempontjából, akkor pl. szubjektív súlyozást alkalmazunk.

Más jellegű probléma, amikor a változók különböző dimenziójából adódó esetlegességet kívánjuk kiszűrni. A statisztikában leggyakrabban használt *standardizálást* itt is alkalmazhatjuk, ekkor minden változó azonos súlyú. Euklideszi távolság esetén ez a *normálás* $w = \dfrac{1}{s^2}$ súlyozást jelent, ami azzal a veszéllyel jár, hogy éppen azon – nagy szórású – ismérvek szerepét csökkentjük, amelyek a legalkalmasabbak a csoportok megkülönböztetésére. Lényegesen elfogadhatóbb eredményre jutnánk, ha a teljes szórás helyett a csoporton belüli szórással *normálnánk*; a vizsgálat előtt ez persze nem ismert.

Az *a priori* ismeretek hiánya miatt bírálhatók azok a módszerek is, amelyek a súlyozással a korreláció hatását kívánják kikapcsolni, ilyen a *Mahalanobis*-távolságfogalom is. Megfelelően szűri ki a korrelációnak a kapcsolatok mérését torzító hatását a faktorelemzés (főfaktormódszer), ami azzal a további előnnyel is jár, hogy az eredeti adatmátrix mérete jelentősen csökken. Meg kell azonban jegyezni, hogy a súlyozás a különböző típusú változók együttes megjelenéséből adódó problémákat nem oldja meg, ehhez a megfelelő skálatranszformációt kell elvégezni.

Nem metrikus távolságmértékek

A nem metrikus mértékek egyik típusa az objektumok között relációkat definiál, és ezek relációelméleti alapon való feldolgozásával csoportosít. A másik típus tulajdonképpen feltételez valamilyen – az előzőektől eltérő – metrikát, amelyre támaszkodva az objektumok rendezését végzi el, de a rendezésnek már nincsenek meg a metrikától megkövetelt tulajdonságai.

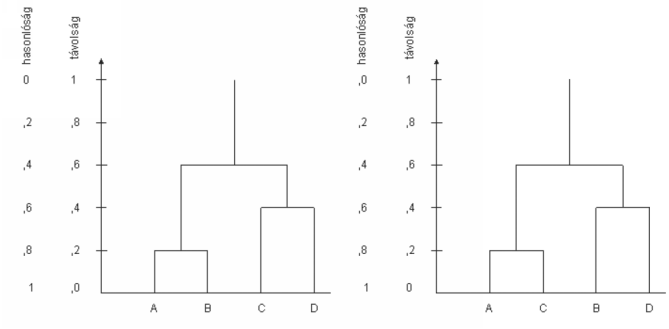
Intervallum változók: Calhoun-távolság

Ez a távolságfogalom a kérdéses két pont és a koordinátatengelyek iránya által meghatározott hiperfelületek közé eső többi pontra épül. Pl. az $\mathbf{x}_1$ és $\mathbf{x}_2$ pont közötti távolságot a bevonalazott területbe eső pontok segítségével határozhatjuk meg (9.2. ábra).

Két pont Calhoun-távolsága a definíció szerint

$$d_c = 6N_i + 3N_b + 2N_z, \tag{9.23}$$

ahol



9.2. ábra. A Calhoun-féle távolságfogalom szemléltetése

N_i = az azon pontok száma, amelyek legalább egy változójuk szerint a két pont által meghatározott hipersíkba vagy meghosszabbításába esnek, legalább egy változójuk szerint,

N_b = az azon pontok száma, amelyek egyetlen dimenzióban sem esnek a két pont közé, de egy vagy több változó szerint határra esnek,

N_z = az azon pontok száma, amelyek (egy vagy több változó szerint mindkét ponttal azonos értékűek, de) nem esnek a hipersík belsejébe vagy a határra.

Ha N az alappontok száma, akkor d_c maximális értéke $6(N-2)$.

A 9.2. ábra pontjait vizsgálva d_c maximuma $6 \cdot (7-2) = 30$. Ax x_1 és x_2 Calhoun-távolságát mérve $N_i = 3$, $N_b = 0$ és $N_z = 1$, így $d_c(x_1, x_2) = 6 \cdot 3 + 2 \cdot 1 = 20$.

A Colhoun-távolság mint mérték nem felel meg a metrika követelményeinek, mert két pont távolsága akkor is lehet 0, ha a két pont nem esik egybe. Ez a mérték hasznos lehet olyan esetekben, ha a klaszterek egy vagy több változó szerint átfedik egymást.

Lance és Williams mértéke[2]

Ez a mérték két objektum távolságát arányként definiálja

$$d_{LW} = \sum_i \frac{|x_i - y_i|}{(x_i + y_i)}, \text{ ahol } \quad x_i, y_i \geq 0 \tag{9.24}$$

A számláló a Minkowski-metrika $r = 1$ esetén, a nevező pedig a maximális kiterjedést méri. Bináris változókra a következő alakot kapjuk:

$$d_{LW} = \frac{b+c}{2a+b+c} = 1 - \frac{2a}{2a+b+c} \tag{9.25}$$

A továbbiakban közelséget említünk, ha a hasonlósági vagy távolsági mérték megkülönböztetése elhagyható.

[2] A szakirodalomban Canberra mértéke néven is szerepel.

9.2.5. Információs mérték

Jardine és Sibson a Q-típusú elemzéseknél a távolság és hasonlóság problémáját elkerülendő az információs mérték számítását javasolja, amely a Shannon-féle entrópiából vezethető le.

A Shannon-féle entrópia:

$$H = -\sum_{i}^{n} p_i \log_2 p_i = \sum_{i}^{n} p_i \log_2 \frac{1}{p_i},$$

ahol p_i a változó x_i értékéhez tartozó valószínűség.

Ez onnan ered, hogy egy X-esemény bekövetkezése $\log_2 \frac{1}{p_i}$ információt hordoz (annak egysége a bit, minthogy $\log_2 \frac{1}{0{,}5} = \log_2 2 = 1$ a kettes alapú logaritmus esetén). A többszöri megfigyelés során nyert információ átlagos értéke az entrópia.

Az entrópia lehetséges értéke $0 \leq H \leq \log_2 n$ intervallumban mozog. A maximumot akkor veszi fel, amikor minden lehetséges kimenetel egyformán valószínű, 0 akkor, ha csak egyetlen biztosan bekövetkező kimenetel lehet. Ezért mondhatjuk, hogy az entrópia a rendezetlenség mértéke.

A két változó X és Y együttes entrópiáján a

$$H(X, Y) = -\sum_{i}^{n} \sum_{j}^{k} p_{ij} \log_2 p_{ij}$$

mennyiséget értjük (ennyi információra van szükségünk ahhoz, hogy a két változó együttes kimenetelét meg tudjuk mondani). A p_{ij} az $(\mathbf{x}_i, \mathbf{y}_j)$ értékpárhoz tartozó valószínűség.

A kölcsönös információ értéke

$$I(X, Y) = H(X) + H(Y) - H(X, Y)$$

Az $I(X, Y)$, akkor és csak akkor egyenlő nullával, ha X és Y függetlenek.

Ha a kölcsönös információs mértéket normáljuk a két változó entrópiájának szorzatával, egy hasonlóság jellegű mértéket kaphatunk (azért hasonlósági jellegű, mert a felső határa $\frac{1}{H(Y)}$ vagy $\frac{1}{H(X)}$).

Az információs mérték:

$$IM = \frac{I(X, Y)}{H(X)H(Y)}$$

Ha az egyedeket bináris változókkal jellemezzük, két egyed, az i-edik és j-edik hasonlóságát az alábbi táblázatból számíthatjuk:

		i-edik egyed		
		1	0	
j-edik egyed	1	n_{11}	n_{01}	$n_{\cdot 1}$
	0	n_{10}	n_{00}	$n_{\cdot 0}$
		$n_{1\cdot}$	$n_{0\cdot}$	p

Az i-edik egyed entrópiája:

$$H(X_i) = -\left[\frac{n_{1.}}{p}\log_2\frac{n_{1.}}{p} + \frac{n_{0.}}{p}\log_2\frac{n_{0.}}{p}\right]$$

A j-edik egyed entrópiája:

$$H(X_j) = -\left[\frac{n_{.1}}{p}\log_2\frac{n_{.1}}{p} + \frac{n_{.0}}{p}\log_2\frac{n_{.0}}{p}\right]$$

Az együttes entrópia:

$$H(X_i, X_j) = -\left[\frac{n_{11}}{p}\log_2\frac{n_{11}}{p} + \frac{n_{01}}{p}\log_2\frac{n_{.10}}{p} + \right.$$
$$\left. + \frac{n_{10}}{p}\log_2\frac{n_{10}}{p} + \frac{n_{00}}{p}\log_2\frac{n_{00}}{p}\right]$$

Az információs mérték:

$$IM = \frac{H(X_i) + H(X_j) - H(X_i, X_j)}{H(X_i)H(X_j)}.$$

9.2.6. A változók súlyozása

A változók megfigyelési értékeinek közvetlen felhasználása az alakfelismerési eljárásban bizonyos problémákat vet fel.

– A változók különböző mértékegysége az egyedek hasonlóságát mérő mutatók számítását megkérdőjelezheti.
– A változók értékeinek különböző nagyságrendje szintén befolyásolhatja a felhasznált mértékek nagyságát, és így az eredményeket.

A problémák feloldására célszerű a változókat súlyozni. A súlyozással egyrészt kiküszöbölhetjük a mérés különbözőségéből adódó problémákat, másrészt a vizsgálat céljának megfelelő fontossággal vehetjük figyelembe a változókat.

1. A nagyságrendi eltérések kiküszöbölésére használatos eljárás lehet, ha a változót valamilyen statisztikája felhasználásával transzformáljuk.

Ilyen egyenlősítő faktor lehet

- a változó átlaga
- a változó maximális terjedelme
- a változó szórása

Ezzel a súlyozási eljárással a változókat azonos nagyságrendűvé transzformáljuk. A változók szórását is egységesítjük, ha standardizált változóvá alakítunk át minden változót (a változó minden lehetséges értékéből kivonjuk a változó átlagát, és osztjuk a szórással).

A standardizált változóknak egységesen 0 az átlaguk és 1 a szórásuk.

Egy speciális egyenlősítő eljárás lehet a főkomponens-elemzés. Ebben az esetben a mintatérből az ún. faktortérbe térünk át, ahol a változók az adatrendszer belső struktúrájában játszott szerepük szerint kapnak súlyt.

2. A súlyozásnak egy másik módja, ha valamilyen, a kutatás céljának megfelelő kritériumváltozót választunk és a minta változóit aszerint súlyozzuk, ahogy a változók a kritériumváltozót a legkisebb négyzetek módszere értelmében legjobban becslő regressziós súlyokat kapják.

Ez a módszer szubjektív elemként kezeli a kritériumváltozó kijelölését, viszont az adatrendszer objektív struktúrájának megfelelő regressziós súlyokat rendel a változókhoz.

3. Nominális változók esetén a súlyozást a változók osztályainak a súlyozására alkalmazhatjuk. Természetes gondolat, hogy az egyes osztályok súlya előfordulásuk relatív gyakoriságának a függvénye (vagyis valószínűségének a függvénye) legyen. Ha n_i az i-edik osztályba eső egyedek száma, annak valószínűsége, hogy egy tetszőlegesen választott egyed az i-edik osztályba esik n_i/n.

Két egyedet összehasonlítva annak valószínűsége, hogy mindkettő az adott változó i-edik osztályába esik. $p_{ii} = (n_i/n)^2$. Annak valószínűsége, hogy az egyik az i-edik, a másik a j-edik osztályba esik $p_{ij} = 2n_i n_j/n^2$.

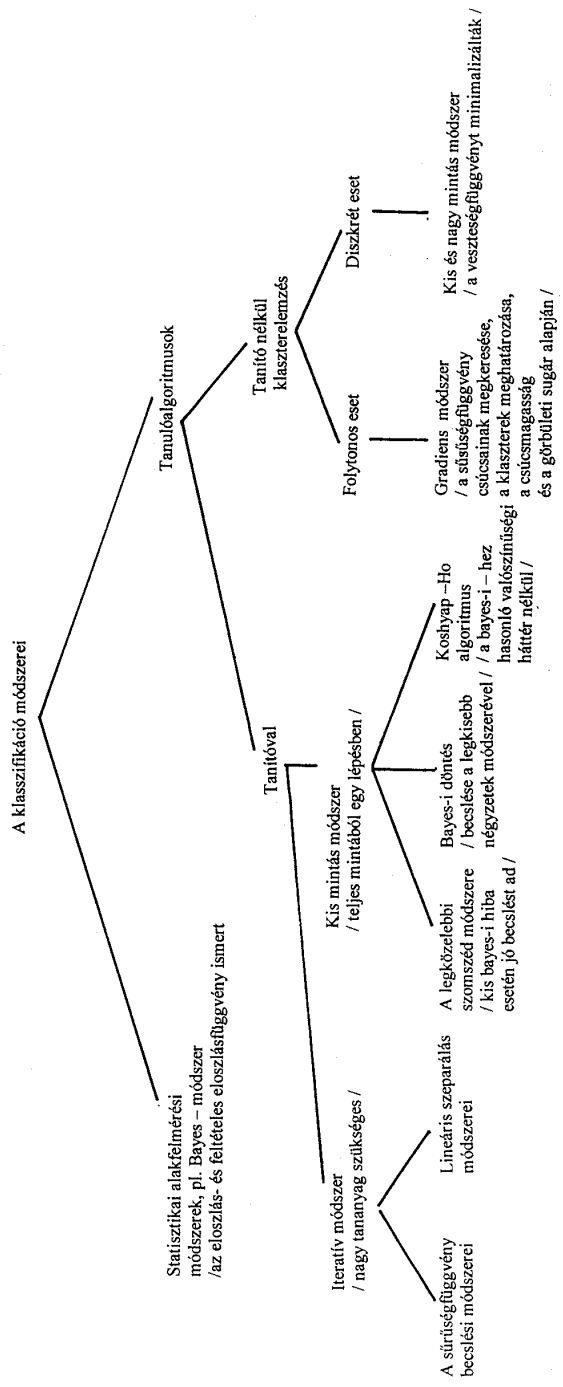
A valószínűség helyett célszerű annak valamilyen függvényét használni, pl. $(1-p_{ij})$ vagy $1/p_{ij}$. Az $(1-p_{ij})$ kevésbé diszkriminál, így az $1/p_{ij}$ vagy az $(1-p_{ij})$ valamilyen hatványa látszik megfelelőnek. Mivel a változók kategóriáinak a száma különböző lehet és a fenti súly függ a kategóriák számától, célszerű, ha az átlagos súllyal osztjuk az osztályok súlyait, így minden változónál egységnyi átlagú súlyokat kapunk.

9.3. Kategorizálás

Az alakfelismerés folyamatának másik lényeges mozzanata a *kategorizálás*.

A kategorizálás különbségei szerint az alakfelismerés következő módszereit különíthetjük el (lásd 9.3. ábra).

Az alakfelismerés módszereit két nagy csoportra bonthatjuk. Az egyik csoportban azokat a módszereket találhatjuk, amelyek feltételezik a csoportok típusainak ismeretét, a második csoport módszerei pedig megpróbálják megkeresni a típusokat a mintából. A második csoportban a módszerek egy része a tanuló mintát úgy bontja csoportokba, hogy egy kiértékelt mintapéldát, tananyagot kap az osztályok definiálásával, amelyet az osztályozásnál figyelembe vesz. A tanító nélküli algoritmusoknak előzetes értékelés nélkül kell megtalálni a mintában található diszjunkt osztályokat, típusokat.



9.3. ábra. Az alakfelismerés módszerei

9.4. A klaszterelemzés módszerei

Tudományos osztályozás, kategorizálás esetén a besorolás objektív kritériumok alapján végezhető el. Egy osztályozási sémáról önmagában nem lehet eldönteni, hogy jó-e vagy rossz, ez azon múlik, hogy a felosztás a vizsgálat szempontjából mennyire megfelelő. Meghatározhatunk ugyanakkor általános elvárásokat, mint például objektivitás, stabilitás és prediktivitás.

Az *osztályozás objektivitása* alatt azt értjük, hogy az adott terület szakemberei a vizsgált egyedeket jellegükben azonos módon csoportosítják. A *stabilitás* azt jelenti, hogy egy-egy új adat kevéssé befolyásolja az osztályozást. A *prediktivitás* a gyakorlatban ritkán teljesül, csak olyan magas szintű osztályozás sajátja, mint a Mengyelejev-féle periódusos rendszer.

A klasszikus logikára épülő osztályozásban két lépés különböztethető meg:

- típus- és koncepcióalkotás, kategóriák definiálása,
- az események kijelölése a már definiált kategóriákhoz.

A klasszikus logika azonban csak olyan kategóriákat definiál, amelyeknek minden egyede minden szempontból ekvivalens. Az ilyen elven alapuló osztályozást monotetikus osztályozásnak nevezzük. Ez a módszer sokváltozós, nagyméretű adatrendszer esetén az eredmények áttekinthetetlen felaprózottsága miatt a gyakorlatban használhatatlan lenne. Ezért nagy jelentőségű a klaszterelemzés, amely lehetővé teszi a sokváltozós nagy minták áttekinthető értékelését.

A klaszterelemzés három szempontból is eltér a klasszikus osztályozástól:

a) Nem definiál típusokat mielőtt kijelölné a mintaelemeket, de feltételezi, hogy

- léteznek típusok,
- a típusfogalom ismerete nélkül is létezik olyan kritérium, amelynek felhasználásával a klaszterek felismerhetők,
- a felismert klaszterhez az egyedek ismérvei alapján megadhatók a típusjellemzők.

Mindez szemléletesen azt jelenti, hogy az n-dimenziós térben az egyes típusokat elkülönítő hipersíkok akkor válnak láthatókká, ha az azonos klaszterbe tartozó elemeket meghatároztuk. Így a csak empirikusan előforduló típusok is felismerhetők.

b) A klaszterelemzés megengedi a politetikus osztályokat. Politetikusnak tekintünk egy osztályt, ha elemei több, de nem minden jellemző szerint ekvivalensek vagy hasonlóak. Ezáltal jelentősen csökkenthető az osztályok száma.

c) A klasszikus modellek csak diszkrét változókkal dolgoznak, a klaszterelemzés megenged folytonos, sőt vegyes változókat is.

A klaszterelemzést sikerrel alkalmazhatjuk akkor, ha célunk

- az adatstruktúra feltárása
- a típusalkotás és a reprezentatív elemek kiválasztása
- a csoportokon alapuló becslések elvégzése.

A klaszterelemzés összefoglaló elnevezése azoknak a módszereknek és eljárásváltozatoknak, amelyek az objektumok hasonlósága vagy távolsága alapján végzik el a csoportosítást.

A szerteágazó klasszifikációs eljárásokat összefoglaló 9.3. ábrán már szerepelt a klaszterelemzés folytonos és diszkrét esetének megkülönböztetése. Mivel a megfigyelések vagy a változók klaszterezésére többféle elv alapján sokféle algoritmus készült, az eljárások bemutatása előtt érdemes a további csoportosítási lehetőségeket ismertetni.

Fontos különbséget tenni az átfedéses és az *átfedésmentes kategorizálás* között. Elméleti kidolgozottsága és gyakorlati jelentősége miatt részletesen a diszjunkt, átfedéseket nem tartalmazó osztályozási módszerekkel foglalkozunk.

A kategorizálás alapja az elemző által kiválasztott döntési kritérium vagy függvény. Döntési kritériumon azt az elvet értjük, amely szerint a kategóriák kialakulnak, ha a p-dimenziós térben rendezzük a vizsgálandó n egyedet. A döntésfüggvény mérheti

- a klaszteren belüli elemek hasonlóságát, vagy
- a klaszterek közötti különbséget.

Az előző alfejezetekben láttuk, hogy két objektum közötti hasonlóság vagy távolság is többféleképpen értelmezhető. A döntésfüggvény feladata ennél jóval összetettebb, mert több elemet tartalmazó csoportok közötti hasonlóságot, illetve különbözőséget kell mérnie vagy becsülnie. A különböző döntési kritériumokhoz különböző klaszterfogalmak kapcsolódnak, így léteznek:

1. sűrűségfüggvény-becslésen alapuló eljárások,
2. valószínűségeloszlások keverékének szétválasztásán alapuló eljárások,
3. csoporton belüli variancia becslését felhasználó eljárások,
4. a csoportok közötti különbséget becslő, és
5. gráfelméleten alapuló eljárások.

A homogén egyedek átfedéseket nem tartalmazó klaszterezése

- hierarchikus kategorizálással és/vagy
- nemhierarchikus kategorizálással végezhető el.

A hierarchikus vagy nemhierarchikus klaszterezés menete alapvetően a kategorizáláshoz felhasznált döntési függvénytől függ, de a gráfelméleti alapú egyszerű láncelvű klaszterezés például több eljárás algoritmusában is szerepel.

A *hierarchikus kategorizálás* két eljárásváltozatot foglal magába. Az összevonó (agglomeratív) hierarchikus eljárás kezdetben minden elemet külön osztálynak tekint, majd az osztályok összevonásával lépésről lépesre újabb kategorizálási szinteket alakít ki, mindaddig, amíg az összes elem egyetlen osztályba nem kerül. A hierarchikus agglomeratív eljárások $(n-1)$ lépésben elvégzik azt az összevonás-sorozatot, amely grafikusan – 2 dimenzióban – megjeleníthető. Ha az adott lépésben k csoport van, akkor a következő összekapcsolást maximum $(k-1)(k-2)/2$ távolság összehasonlításával lehet kiválasztani.

A felosztó (divizív) hierarchikus eljárás minden egyes lépésben – valamilyen döntési kritérium alapján – kettéosztja a megfigyeléseket, így az eljárás $(2^{n-1}-1)$ felosztás megvizsgálása után fejeződik be.

Tegyük fel, hogy a szakmai ismereteink alapján k számú klaszter létét tételezzük fel, ilyenkor nemhierarchikus eljárást választunk. A *nemhierarchikus kategorizálás* előre adott k számú osztályra bontja – alkalmasan megválasztott döntésfüggvények alapján – a mintát.

Az n számú elem[3] k nem üres csoportba $\frac{1}{k!}\sum_{j=1}^{k}(-1)^{k-j}\binom{k}{j}j^n$-féleképpen sorolható be. A képlet alapján 8 megfigyelést 2 csoportba $1/2(-2+2^8) = 127$ változatban, 3 csoportba pedig 966 felosztás szerint lehet besorolni.[4]

[3] Hasonlóan írható fel a Stirlingtől származó formula akkor is, ha p számú változót sorolunk k csoportba.

[4] A hierarchikus klaszterezés felosztó eljárása $n = 8$ esetén 127 lehetséges partíciót jelent, az agglomeratív klaszterezés az első lépésben maximum 28, a másodikban 21, összesen nem több, mint 84 összevonási lehetőség vizsgálatát igényli.

Az csak az egyik probléma, hogy minden lehetséges felosztást elkészíteni kis minta esetén is nagy számításigényt jelent. Az elkészített felosztások közül a „legjobb" kiválasztása az igazi kihívás a kutató számára, mert nincsen olyan objektív mérőszám, amely minősíti a klaszterezés eredményét.

Ha a struktúrafeltárás kezdetén a csoportok számát nem ismerjük, akkor minden $1 \leq k \leq n$ számra el kellene végezni a felosztást, hogy a k elfogadható értékét megtaláljuk. Nagy méretű feladatok esetében ez az út járhatatlan, ezért ilyenkor a hierarchikus klaszterezés felosztó vagy összevonó változatát elvégezve „tájékozódhat" a kutató.

9.5. Hierarchikus módszerek

9.5.1. Agglomeratív eljárások

A hierarchikus módszerek sajátossága az, hogy a csoportosításhoz nem kell megadni a mintában létező (vagy feltételezett) csoportok számát. Hierarchikus klaszterezést a hasonlósági vagy távolság-mátrixból[5] kiindulva két változatban készíthetünk. A megfigyelések egyre finomabb felosztását végző (divizív) eljárás mellett a legközelebbi elemeket összevonó (agglomeratív) eljárások sora áll az elemző rendelkezésére.

Az összevonó algoritmusok során származtatott távolsággal mérjük a már egy klaszterbe sorolt egyedek távolságát a többi egyedtől vagy klasztertől. A származtatott távolságokra teljesülnie kell az előző alfejezetben bevezetett ultrametrikus egyenlőtlenségnek,[6] azaz az összevonás során a származtatott távolságok monoton nőnek. E feltételnek eleget tevő elgondolás az *„egyszerű lánc"* klaszterezés, amelynek számos numerikus változata létezik. Van agglomeratív eljárás (legközelebbi szomszéd) és a fordított utat követő, divizív klaszterezés is, amely az egyszerű láncot követi. A legtöbb szerző csak agglomeratív eljárásként tárgyalja az egyszerű lánc klaszterezést, míg mások, például Jardine és Sibson (1971), Sibson (1973) rámutattak arra, hogy gráfelméleti alapon[7] is végezhető egyszerű lánc klaszterezés. A pontok által kifeszített minimális fa (MST: minimum spanning tree) készítése során növekvő sorba rendezzük a pontok közötti távolságokat.[8] Minden lépésben két olyan pontot kötünk össze, amelyeket eddig nem kötött össze él, és az élek hosszának összege a legkisebb. Minden pont egyszer és csak egyszer szerepel a láncban, nem képződik hurok. A gráf összefüggő részei a klaszterek. Nagy elemszám esetén nehezen áttekinthető a minimális fa.

Az összevonó eljárások egyelemű klaszterekből indulnak ki, és az összevonások sorozata után végül az egész mintát egy csoportba sorolják.

Az általános eljárás a következő:

---

[5] A mátrix elemei közvetlenül is rendelkezésre állhatnak, vagy az eredeti adatokból származtatott „közelségi" vagy „különbözőségi" mértékeket tartalmazzák.

[6] A megfigyelt adatokból számolt legtöbb távolság nem ultrametrikus.

[7] Ez az algoritmus se nem agglomeratív, se nem divizív.

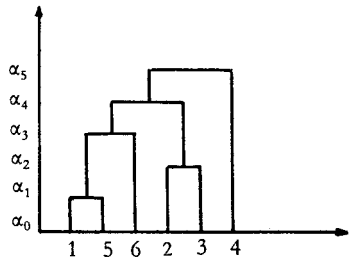
[8] Ha egyező távolságok fordulnak elő, akkor nem kapunk egyértelmű megoldást.

1. kiindul n db egyelemű csoportból (klaszterből);
2. megkeresi a hasonlósági (távolság-) mátrix maximális (minimális) elemét, vagyis a két leghasonlóbb klasztert;
3. a két klasztert összevonja, ezzel a klaszterek számát eggyel csökkenti. Az új klaszter többitől mért távolságát (hasonlóságát) újraszámítja;
4. a 2. és 3. lépést $(n-1)$-szer elvégezve minden egyed egy klaszterbe kerül.

A módszerek abban különböznek, hogy hogyan definiáljuk a csoportok hasonlóságát (távolságát).

Az összevonó eljárások eredménye a klaszterek hierarchikus elrendezését tükröző kétdimenziós ábrán is megjelenik. Ha nem engedünk meg átfedő klasztereket, akkor *dendrogramnak* nevezzük ezt az ábrát. A dendrogram vízszintes tengelyén az egyedek sorszámait, függőleges tengelyén pedig a klaszterek összevonásának szintjeit (α) tüntetjük fel. A függőleges tengelyen hasonlósági vagy távolságmértékek találhatók az input adatoknak megfelelően.

Szemléltetésül tekintsük a 9.4. ábrán látható dendrogramot:



9.4. ábra. A hierarchikus klaszterezés ábrázolása dendrogrammal

Az α értéke 0-tól kezdődik és monoton növekszik, ha távolságmértékkel dolgozunk. Minden klaszter (kivéve az első sort) az előző szint klasztereinek az összevonásából származik. Minden egyedpárhoz hozzárendelhetjük azt az x értéket, amikor először kerültek közös klaszterbe. Ezekből az értékekből a dendrogrammal ekvivalens távolságmátrixot szerkeszthetünk. Erre a távolságra érvényesek a következő tulajdonságok:

— $d(x, x) = \alpha_0 = 0$,
— $d(x, y) = 0$ akkor és csak akkor, ha $x = y$,
— $d(x, y) = d(y, x)$, azaz szimmetrikus a távolság, és
— $d(x, z) < \max [d(x, y), d(y, z)]$.

Ezt az utóbbi tulajdonságot ultra-metrikus egyenlőtlenségnek hívják. Ez erősebb megszorítást jelent, mint a háromszög-egyenlőtlenség:

$$d(x, z) < d(x, y) + d(y, z).$$

A dendrogramon az összevonási szinteknél bekövetkező nagyobb ugrások jelzik az elkülönülő csoportokat. Az induló adatmátrixból a dendrogramot különböző klaszterelemző módszerekkel állíthatjuk elő. A módszerek ismertetése után visszatérünk a dendrogramok értelmezésére.

Legközelebbi szomszéd módszer (egyszerű lánc elve)

Az egyszerű láncmódszer két klaszter távolságát a két klaszter legközelebbi tagjai közötti távolságként értelmezi:

$$D_1(I, J) = \min\ d(\mathbf{x}_i, \mathbf{x}_j), \tag{9.26}$$

ahol $\mathbf{x}_i$ és $\mathbf{x}_j$ az i és j objektumok megfigyelési értékeit tartalmazó két vektor,

$i = 1, \ldots, n_I$,

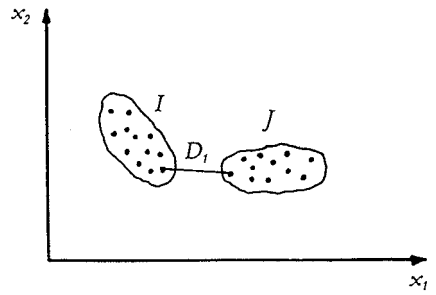
$j = 1, \ldots, n_J$,

n_I az I klaszterbe tartozó objektumok száma,

n_J a J klaszterbe tartozó objektumok száma,

$d(\mathbf{x}_i, \mathbf{x}_j)$ az elemek távolsága.

Ezt a kritériumot mutatja a 9.5. ábra.



9.5. ábra. Távolság értelmezése a legközelebbi szomszéd módszernél

Az I és J csoportokat összevonó lépést követően bármely K csoport távolsága az (IJ) klasztertől az alábbi képlettel határozható meg:

$$D_1(IJ, K) = \frac{1}{2}[D_1(I, K) + D_1(J, K) - |D_1(I, K) - D_1(J, K)|] \tag{9.27}$$

Az egyszerű láncmódszer minden lépésben azt a két klasztert vonja össze, amelyek legközelebbi elemeinek a távolsága (vagyis a klaszterek távolsága) a legkisebb. Ez a módszer a közbülső pontok miatt összekapcsolhat különálló klaszereket. Ez a láncha-tásnak nevezett probléma megnehezíti a kapott eredmények értelmezését.

Az egyszerű láncmódszer a klaszter területének kiterjedését eredményezi. Egyike azoknak a módszereknek, amelyek képesek nem ellipszis alakú klaszterek felismerésére. A csoportok között fekvő közbülső pontok miatt a módszer összefűzhet különböző tulajdonságú klasztereket, így a csoportok nagyon heterogének lehetnek. Itt tehát a klaszterek összekötése, nem pedig a klaszterek homogenitása a cél.

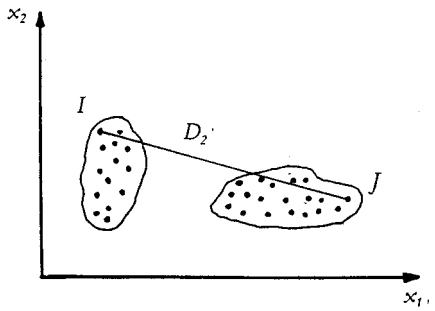
Legtávolabbi szomszéd módszer (teljes lánc-elv)

Ez az eljárás a klaszterek távolságát a két klaszter legtávolabbi tagjai közötti távolságként értelmezi (9.6. ábra):

$$D_2(I, J) = \max\ d(\mathbf{x}_i, \mathbf{x}_j). \tag{9.28}$$

Az (IJ) klaszter távolsága egy tetszőleges K csoporttól:

$$D_2(IJ, K) = \frac{1}{2}[D_2(I, K) + D_2(J, K) + |D_2(I, K) - D_2(J, K)|] \tag{9.29}$$



9.6. ábra. Távolság értelmezése a legtávolabbi szomszéd módszernél

A hierarchikus eljárás egy-egy lépésében azokat a klasztereket vonjuk össze, amelyek legtávolabbi elemei közötti távolság a legkisebb.

A teljes lánc módszer használata akkor célszerű, ha kis átmérőjű csoportok vannak a mintában. Ezzel a kritériummal az összevonásra kerülő klaszterek területének átmérője minimális lesz.

Csoportátlag-módszer (átlagos lánc-elve)

Ez a módszer a klaszterek távolságát úgy számítja, mint a két csoport elemei közt mért távolságok átlagát:

$$D_3(I, J) = \frac{1}{n_I n_J} \sum_{i,j} d(\mathbf{x}_i, \mathbf{x}_j), \quad \text{ahol} \quad \mathbf{x}_i \in I \quad \text{és} \quad \mathbf{x}_j \in J \tag{9.30}$$

Az összevonás kritériuma az, hogy az új csoporton belüli távolságok átlagának növekedése minimális legyen.

A K csoport távolsága a korábban összevont $(I\,J)$ klasztertől a külön-külön mért távolságok súlyozott átlaga:

$$D_3(I\,J, K) = \frac{n_I}{n_I + n_J} \cdot D_3(I, K) + \frac{n_J}{n_I + n_J} D_3(J, K) \tag{9.31}$$

A csoportátlag-módszernek két számítógépes változata van. Az alapértelmezés szerint a *csoportok közötti átlagos láncot* ($n_I \cdot n_J$ távolság átlagát) minimalizáljuk. A csoporton belüli átlagos lánceljárásnál ($n_I + n_J$) elem közti távolság átlagát határozzuk meg.

Centroid-módszer

A centroid-módszer két klaszter távolságát a két klaszter centroidjának, átlagvektorának a távolságaként definiálja:

$$D_4(I, J) = d(\overline{\mathbf{x}}, \overline{\mathbf{y}}), \tag{9.32}$$

ahol $\overline{\mathbf{x}} = [n_I^{-1} \Sigma \mathbf{x}_{ik}]$ $(k = 1, \ldots, p)$ az I klaszterhez tartozó objektumok átlagvektora,

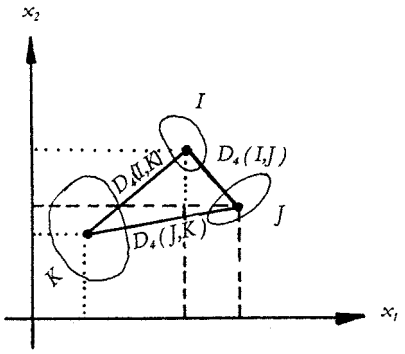
$\overline{\mathbf{y}} = [n_J^{-1} \Sigma \mathbf{y}_{jk}]$ $(k = 1, \ldots, p)$ a J klaszter elemeinek átlagvektora.

Egy tetszőleges K csoport távolsága a közös (I, J) klaszterben levő elemek középpontjától az alábbi képlettel határozható meg:

$$D_4(I\,J, K) = \frac{n_I}{n_I + n_J} D_4(I, K) + \frac{n_J}{n_I + n_J} D_4(J, K) - \frac{n_I n_J}{(n_I + n_J)^2} D_4(I, J) \tag{9.33}$$

A centroid-módszer hátránya, hogy lényegesen különböző számú egyedet tartalmazó klaszterek összevonása után az új centroid nagyon közel kerül a nagyobb elemszámú klaszterhez, így a kisebb méretű klaszter jellege elvész az egyesítés során.

A centroidok távolságán alapuló klaszterezést a 9.7. ábra mutatja.



9.7. ábra. Centroidok távolsága

Medián-módszer

A centroid módszernek az aránytalan elemszámú klaszterek összevonásából adódó problémáját igyekszik feloldani a Gower (1967) által javasolt módszer.

Az átlagpontok távolsága helyett két klaszter távolságát a következőképpen definiáljuk.

$$D_5(I, J) = d(Me_i, Me_j), \tag{9.34}$$

ahol Me_i és Me_j az i-edik és j-edik klaszter mediánjai.

Ha a csoportbeli elemek eloszlása szimmetrikus, akkor a centroid és a medián módszer eredménye közel azonos. A K csoport mediántávolsága (9.29) speciális alakjával határozható meg:

$$D_5(IJ, K) = \frac{1}{2}D_5(I, K) + \frac{1}{2}D_5(J, K) - \frac{1}{4}D_5(I, J) \tag{9.35}$$

Orlóci (1967) módszere a csoportátlag-módszer változata. A csoportok összevonásának kritériuma, hogy a csoporton belüli távolságok átlagának növekedése minimális legyen. Eszerint egy-egy cikluson belül az a két csoport kerül összevonásra, amelyik esetén az összevont csoport egyedei közötti átlagos távolság ($\overline{d}_i$) minimálisan növekszik. A hierarchikus összekapcsolódásokat mutató dendrogram függőleges tengelyén a csoporton belüli átlagos távolságokat ($\overline{d}_i$) a teljes minta távolságai átlagának ($\overline{d}_t$) százalékában mérjük.

Orlóci javasolt egy mutatót a végső csoportok osztályozási hatékonyságának mérésére:

$$E = \frac{\overline{d}_t - \sum^{g} \overline{d}_i}{\overline{d}_t} = 1 - \frac{\sum^{g} \overline{d}_i}{\overline{d}_t}$$

amely tehát a klaszterek közötti eltérések mértékét fejezi ki. Ha minden egyed külön klaszterbe tartozik, E értéke 1.

Orlóci módszere, amikor a csoporton belüli heterogenitást minimalizálja, az E értékét minimalizálja.

Ward-módszer

Ward abból indult ki, hogy a csoportok összevonásával információveszteség keletkezik. A csoportosítás döntésfüggvénye ezt az információveszteséget minimalizálja. Az információveszteséget Ward úgy definiálta, mint a megfigyelések csoportátlaguktól való eltéréseinek a négyzetösszegét. Ez az eltérés-négyzetösszeg nem más, mint a csoporton belüli variancia. Ha T-vel jelöljük a teljes minta varianciáját, akkor érvényes a következő felbontás, ahol B a csoporton belüli varianciák összege és K a csoportok közötti varianciák összege

$$T = B + K.$$

A T egyértelműen meghatározott, és B és K a csoportosítástól függően változik. A cél olyan klaszterek keresése, amelyek esetén B minimális.

Az első lépésben, amikor minden klaszter egyelemű, a belső – átlagtól való – eltérések nullával egyenlők. Bármely két elem összevonása növeli[9] a belső eltérések négyzetösszegét, az SSB-t, amely az I csoportra az alábbiak szerint írható fel:

$$SSB_I = \sum_{i=1}^{n_I} \sum_{j=1}^{p} (x_{ij} - \overline{x}_{Ij})^2 \tag{9.36}$$

Két elem vagy csoport összevonása információvesztéssel jár együtt, de Ward-eljárását követve az SSB lehető legkisebb növekedése az információveszteséget is minimalizálja.

Ha az I és J csoportokat összevonjuk egy klaszterbe, akkor új átlagvektort kell meghatároznunk ($\overline{\mathbf{x}}_{IJ}$), és $n_I \cdot n_J$ elemre összegzünk (9.36) szerint. A csoportok távolsága, azaz a belső eltérés-négyzetösszeg növekedése az összevonás következtében a centroidok távolság-négyzetének és a csoportok elemszámának a függvénye.[10]

$$D_6(I, J) = SSB_{IJ} - (SSB_I + SSB_J) = \frac{n_I n_J}{n_I + n_J} d^2(\overline{\mathbf{x}}_I, \overline{\mathbf{x}}_J) \tag{9.37}$$

Az I és J összevonására csak akkor kerül sor, ha $D_6(I, J)$ kisebb, mint bármely más összevonással kapott D_6 érték. Bármely K csoport és az (IJ) klaszter távolsága a következő képlettel számolható ki:

$$D_6 = (IJ, K) = \frac{1}{n_I + n_J + n_K} \{(n_I + n_K) D_6(I, K) + (n_J + n_K) D_6(J, K) - n_K D_6(I, J)\} \tag{9.38}$$

A Ward-technika fogyatékossága, hogy nem ad minden esetben a B minimális értékére optimális megoldást. Így előfordulhat pl., hogy három csoportos felbontás minimális B értéket adó varianciája a következő lépésben, két csoport esetén nem lesz optimális. Biztosítani csak a lokális optimum megkeresését tudja.

---

[9] Az SSB maximumát akkor éri el, ha a klaszterezés utolsó lépésében minden elem egy csoportba kerül. Ekkor $SSB = SST = \sum_{i=1}^{n} (x_i - \overline{x})^2$, minden egyes változóra.

[10] (9.37) bizonyítása megtalálható: Jobson II. 513. old.

Lance és Williams flexibilis módszere

A (9.26)–(9.38) képletekből látható, hogy a hat agglomeratív eljárás különböző-képpen méri a csoportok közötti távolságot. Lance és Williams (1966) megmutatta, hogy e különbözőségek ellenére a klaszterek távolsága az alábbi közös képlettel írható fel:

$$D(IJ, K) = \alpha_I D(I, K) + \alpha_J D(J, K) + \beta D(I, J) + \gamma |D(I, K) - D(J, K)| \qquad (9.39)$$

Az összevonás kezdetén $D(I, J)$ két eredeti megfigyelés közötti minimális távolság, és a további lépésekben az α, β, γ paraméterek megválasztásával bármelyik eljárás számítógépes programja elkészíthető. A 9.3. táblázatban az egyes hierarchikus összevonó eljárások és a távolság-paraméterek megfeleltetése látható.

Eljárás	α_I	α_J	β	γ
1. Egyszerű lánc	1/2	1/2	0	–1/2
2. Teljes lánc	1/2	1/2	0	1/2
3. Átlagos lánc	$\frac{n_I}{n_I + n_J}$	$\frac{n_J}{n_I + n_J}$	0	0
4. Centroid	$\frac{n_I}{n_I + n_J}$	$\frac{n_J}{n_I + n_J}$	$-\alpha_I \alpha_J$	0
5. Medián	1/2	1/2	–1/4	0
6. Ward	$\frac{n_I + n_K}{n_I + n_J + n_K}$	$\frac{n_J + n_K}{n_I + n_J + n_K}$	$\frac{-n_K}{n_I + n_J + n_K}$	0

9.3. táblázat. Agglomeratív eljárások Lance–Williams együtthatói

Minimális feszített fa módszer

Zahn (1971) írt le egy módszert az automatikus irányítású klaszterek keresésére. A módszer a teljes összefüggő gráf minimális feszített fa konstrukcióján alapul. A minimális feszített fa módszerének használata a pontábrák könnyű átláthatósága miatt kézenfekvő előnnyel jár. A fa gráfelméleti fogalom. Egy gráfot fának nevezünk, ha

- bármely két pontja között létezik egy lánc (a pontokat összekötő élek sorozata), vagyis a gráf összefüggő
- ciklusmentes
- legalább két csúcsa (pontja) van.

A fa hosszúságán a pontokat összekötő élek hosszának összegét értjük. Minimális feszített fa alatt a minimális hosszúságú fát értjük.

A minimális feszített fa módszer alapgondolat: az adatok belső lényegi szeparációját úgy fedezi fel, hogy a minimális feszített fából eltöröl olyan éleket, amelyek szignifikánsan hosszabbak, mint a szomszédos élek. Ezeket az éleket inkonzisztens éleknek nevezzük.

A minimális feszített fát a következő algoritmussal szerkeszthetjük meg.

Sorbarendezzük az egyedek közötti távolságokat.

Két pontot akkor kötünk össze éllel, ha

- eddig még nem köti össze él,
- az összekötött pontokon keresztül nem juthatnak el az egyikből a másikba,
- az előző két feltételt kielégítő pontszámok közül az adott két pont távolsága a legkisebb.

Az eljárás lépéseiként kapott gráf összefüggő komponensei az egyes csoportok.

Az algoritmus során a csoportok száma fokozatosan csökken, az eljárás végeredményeként pedig minden egyed egy csoportba kerül. Az így elkészített minimális fából az inkonzisztens éleket kell kitörölni.

Zahn egy élet inkonzisztensnek nevez, ha szignifikánsan nagyobb, mint a szomszédos élek átlaga. Zahn szignifikánsan nagyobbnak nevez egy élet, ha

- hosszúsága több mint f-szerese a szomszédos élek átlagának,
- hosszúsága nem kisebb, mint a szomszédos élek szórásának s-szerese.

Az f és s paramétereket a felhasználó határozza meg.

Célszerűnek látszik eddig végzett futtatásaink alapján f értékét minimálisan 2 vagy 3-nak megválasztani; az s értéke pedig jó, ha 1-nél nagyobb.

A „szomszédos" értelme a következő:

egy P pont szomszédos egy Q ponttal, ha a P pontot nyilakkal összekötve a Q ponttal d vagy kevesebb él szükséges[11] a minimális kifeszített fában. A d szintén paraméter, a felhasználó adja meg az értékét.

Az inkonzisztens éleket a fából kitörölve a fát részfákra osztjuk, amelyek az adatok egy-egy klaszterét adják.

9.5.2. Divizív eljárások

A hierarchikus módszereknek ez a fajtája n egyedet oszt egymás után kisebb elemszámú csoportokba egészen addig, amíg minden egyed külön klasztert alkot. A kezdő lépésben n egyedet kell kétfelé bontani. Ezt $2^{n-1} - 1$ féle módon végezhetjük el. Ez mutatja, hogy még kis adathalmaz esetén is nagy számítógépre van szükség az összes lehetséges felbontás megvizsgálására. Ezért a módszereknél törekedni kell arra, hogy minél kevesebb eset megvizsgálása után kerüljön sor a döntésre.

A divizív eljárásokat két nagy csoportra bonthatjuk:

- politetikus és
- monotetikus módszerek.

Politetikus módszer

NacNaughton és Smith (1964) használható politetikus módszert közöl. A teljes mintából kiválasztjuk azt a pontot, amelynek a többi ponttól mért távolságainak átlaga maximális. Ez lesz a splinter csoport első egyede. Ezt a teljes mintából hasított csoportot azzal az egyeddel kell bővíteni, amelyiknek a fő csoport egyedeivel mért átlagos euklideszi távolságának s a splinter csoport egyedeivel mért átlagos távolságának a különbsége maximális. Egészen addig kell végezni az eljárást, amíg ez a különbség minden fő csoporthoz tartozó egyed esetében negatívvá válik. Ekkor a kapott két csoporttal külön-külön folytatjuk az eljárást mindaddig, míg vagy megfelelő számú csoportot nem kaptunk, vagy már tovább nem lehet bontani, minden csoport egyelemű.

Egy másik eljárás a diszkriminanciaelemzésen alapul. Az algoritmus indításához egy kezdő felosztásra van szükség. Ekkor számítjuk ki a diszkriminanciafüggvényt. Ezután iterációval újra kijelöljük a pontokat, és az új csoport beosztásra számítjuk ki a diszkriminanciafüggvény aktuális értékét és a Wilks-féle lambdát. Az iterációt mindaddig

[11] Az összekötéshez.

folytatjuk, amíg meg nem találjuk a legjobban elkülönülő csoportokat. Casetti, Hung és Dubes írtak hozzá számítógépes programot.

Monotetikus módszer

Az adatrendszert bináris változókkal jellemezzük, és a feladat az, hogy olyan csoportokat keressünk, amelyeken belül az egyedek minden változó szerint azonosak (0 vagy 1 értékűek). Lambert és Williams (1962, 1966) közöl asszociációs elemzés címen ilyen eljárást. Az n megfigyelési egységet m bináris változóval jellemezzük. A j-edik és k-adik változó asszociációs kapcsolatát a χ^2 (khi négyzet) együtthatóval jellemezhetjük

$$\chi^2_{jk} = \frac{(ad - bc)^2 n}{(a+b)(a+c)(b+d)(c+d)}$$

ahol az a, b, c, d jelentése a szokásos kontingenciatábla gyakoriságát jelöli.

		j változó		
	1	a	b	$a+b$
k változó				
	0	c	d	$c+d$
		$a+c$	$b+d$	n

Az m változó alapján egy $m \times m$-es χ^2 együtthatókat tartalmazó mátrixot állíthatunk elő.

A minták kettébontását azon k változó szerint végezzük, amelyre $\sum_{j \neq k} \chi^2_{jk}$ minimális.

Az egyik csoportba a k jellemzővel rendelkező egyedek kerülnek, a másikba azok az egyedek, amelyek nem rendelkeznek a k tulajdonsággal. Így a kettéosztást mindig egy változó szerint végezzük, ezért nevezik monotetikusnak a módszert.

A bontási kritérium lehet még

$$\max\{\Sigma|ad - bc|\}$$

$$\max\{(ad - bc)^2\}.$$

AID (Automatic Interaction Detector) módszer

Songuist és Morgan (1964) módszerét eredetileg nem klasszifikációs módszerként ismerték, a kapcsolat csak később derült ki. Ez a többváltozós technika azokat a változókat keresi meg, amelyek maximálisan elkülönítenek néhány kritériumváltozó szerint képzett csoportot. Az eljárás lényegében monotetikus felosztó technika.

A csoport szétválasztásának kritériuma: a kritériumváltozó csoporton belüli négyzetes hibaösszeg minimumának keresése vagy a csoportok közötti és a teljes eltérés négyzetösszeg hányadosának maximalizálása.

Minden változóra kiszámítjuk ezt az arányt (a kritériumváltozó csoportok közötti és a teljes eltérés négyzetösszegének hányadosát) és a maximális hányadossal rendelkező bináris változó szerint osztjuk a megfigyeléseket csoportba. Minden csoportra az eljárást megismételjük.

9.5.3. A hierarchikus eljárások összehasonlítása

Lance és Williams flexibilis módszere mutatja meg ezt a „közös gyökeret", amely a hierarchikus eljárások egyik legvonzóbb tulajdonsága, és egyúttal ez okozza az alkalmazások során a legnagyobb nehézséget. A módszerek változékonyságából eredő hátrány azzal magyarázható, hogy az eltérő eljárások[12] eltérő felosztást, és így eltérő dendrogramot eredményeznek. A szakirodalomban egyes szerzők (pl. Johnson és Wichern [1998]) azt javasolják, hogy több változatban célszerű elvégezni a klaszterezést, és stabilabb az eredmény, ha egymással összhangban levő felosztások adódnak. Mivel a hierarchikus módszereknél a korábban besorolt elemek áthelyezése nem valósítható meg, a kezdeti lépések döntő jelentőségűek.

Más szerzők (pl. Krzanowski [2000]) amellett érvelnek, hogy a csoportosítandó elemek természetét tanulmányozva előre kell módszert választani. Így elkerülhető a sok fölösleges futtatás, valamint az, hogy az előzetes elvárásainknak legjobban megfelelő eredményt választjuk. Mindkét megközelítés megfontolandó, ezért a módszerválasztás megkönnyítése érdekében tekintsük át részletesebben a klaszterező eljárások főbb jellemzőit. Ha a klaszteréljárások matematikai tulajdonságait vizsgáljuk, akkor fontos megjegyezni, hogy az egyedek közötti d_{ij} távolságok monoton traszformációjára csak az egyszerű lánc és a teljes lánc módszerek invariánsak. Az *invariancia* fontosságára Jardine és Sibson (1971) mutattak rá. A távolságok logaritmusát véve például eltérő felosztás és eltérő dendrogram adódik, ha nem a legközelebbi vagy a legtávolabbi szomszéd-elvet követjük.

A klaszterek geometriai alakja eltérő az egyes eljárásoknál. Az egyszerű lánc módszer jellemzője a *lánchatás*, vagyis az a tulajdonság, hogy bizonyos elemeket közbeeső elemek láncolata révén kapcsol össze. A közös klaszterbe kerüléshez elegendő az is, ha a csoport egyetlen tagjához hasonlít a vizsgált egyed, így az eljárás *tér-összehúzó* hatású. A lánchatás érvényes a medián-módszernél is, ahol az utoljára kapcsolódó pontnak döntő hatása lehet a klaszterezés további menetére.

Viszonylag zárt, „gömbölyű" klaszterekét kapunk, ha teljes lánc-, átlagos lánc-, vagy centroid-módszerekkel végezzük az osztályozást. Ekkor egy-egy klaszter elemei egymással nagyon homogének. A legtávolabbi szomszéd-elv alapján inkább új klaszterek képződnek egy-egy következő lépésben, nem a meglevő csoportokhoz kapcsolódnak az újabb egyedek. Ezt *tér-tágító* hatásnak nevezik a szakirodalomban, míg az átlagos láncelv *tér-konzerváló* hatásúnak tekinthető. A teljes lánc-módszer egyenlő átmérőjű, a Ward-módszer pedig egyenlő elemszámú klaszterek kialakítására törekszik.

Ha az adatok klasztereződése nem egyértelmű, akkor a centroid- és a medián-módszer alkalmazása során problémát okozhat az *inverzió* előfordulása. Ekkor az összevonás későbbi lépésében kisebb távolság adódik, mint a korábbi szintek klaszterei közötti legkisebb távolság: $D(IJ, K) \leq D(I, J) \leq \min [D(I, K), D(J, K)]$. Az inverziót a dendrogramon egymást keresztező összekapcsolódások vagy a vízszintes tengelyen (és az agglomeráció szintjét megadó táblázatban) nem monoton növekvő távolságok jelzik.

További – bár a klaszterezésben nem lényegi – problémát okoz az, ha a távolsági vagy hasonlósági mátrixban megegyező elemek vannak. Ekkor – különösen az összevonás elején – többféle felosztás adódhat, és ez az értelmezést nehezíti.

A hierarchikus eljárással kapott felosztás „illeszkedésének jósága" nem mérhető, de az eredmény *stabilitása* egyszerűen vizsgálható. Ha a csoportok jól elkülöníthetők,

[12] Emlékeztetünk arra, hogy a sokféle hasonlósági és távolságmérték közötti választás lehetősége még további klaszterkombinációkat eredményezhet.

akkor a kiválasztott eljárást az eredeti adatokra és azok perturbációjára[13] is alkalmazva a klaszterezés megegyező, azaz stabil eredményt ad.

Ha többféle távolságmértékkel és/vagy eltérő eljárásokkal is elvégezzük a klaszterezést, akkor nagy valószínűséggel különböző dendrogramokat kapunk, amelyek hasonlóságát meg kell vizsgálni.

Dendrogramok értékelése, összehasonlítása

A hierarchikus összevonó eljárások közös tulajdonsága, hogy az n számú egyedet $(n-1)$ lépésben összevonják egyetlenegy csoportba. Ezt a folyamatot tükrözi a dendrogramnak nevezett kétdimenziós ábra, melynek vízszintes tengelyén az egyedek szerepelnek. A függőleges tengelyen a kiválasztott agglomeratív eljárástól függő származtatott különbözőség[14] értéke jelenik meg.

A dendrogramon látható tehát az összevonás teljes folyamata, de további elemzést igényel a megfelelő klaszterezési eredmény leolvasása. Ehhez információt nyújt az összevonás rendjét és távolságszintjeit mutató táblázat. Mivel egy csoporton belüli egyes egyedek és egy másik csoport egyedeinek távolsága egyetlen szám lesz a továbbiakban, a táblázatban azok a távolságszintek szerepelnek, amelyek mellett először került összevonásra a két csoport. Az összevonás menetét követve $(n-1)$ távolságot kapunk az eredeti $n(n-1)/2$ adat helyett, ezért a származtatott távolságmátrixban sok érték ismétlődik.

- A kapott felosztás érvényességét, azaz a klaszterek validitását a Pearson-féle korrelációs együtthatóval[15] mérhetjük. Ha az osztályozás megfelelő, akkor az eredeti távolságok és a származtatott távolságok közötti korreláció nagyon szoros, azaz a mutató értéke közel egy. A korrelációs együttható kiszámítása révén összehasonlíthatjuk két különböző klaszterezési eljárás eredményét is.
- A két eredmény hasonlóságát mérő korrelációs együttható helyett számíthatjuk az eredeti és a származtatott $(\widehat{\alpha})$ távolságok közötti eltérést is. Az összehasonlítást így a sokdimenziós skálázás (14. fejezet) stress-mértékéhez hasonlóan végezzük el. Az S mutató nemnegatív, maximuma egy, és nullához közeli értéke jelzi a jó illeszkedést.

$$S = \sum_{j=2}^{n} \sum_{i=1}^{j-1} (d_{ij} - \widehat{d}_{ij})^2 / \sum_{j=2}^{n} \sum_{i=1}^{j-1} \widehat{d}_{ij}^2 \tag{9.40}$$

- A klaszter-középpontok közötti távolság mérésével is jellemezhetjük a felosztást. Az összevonó eljárás minden egyes lépésében meghatározhatjuk a kapott klaszterek centroidjait, és a centroidok[16] között mért távolságokat tekinthetjük a klaszterek közötti származtatott távolságnak. Ekkor a származtatott távolságok nem elégítik ki az ultrametrikus egyenlőtlenséget.
- Két dendrogramot összehasonlíthatunk úgy is, hogy a származtatott távolságok értéke helyett az összekapcsolódásokat vetjük egybe. Az $n(n-1)/2$ pontpárra meghatározzuk, hogy az egyes dendrogramokban hányadik összekapcsolódás után kerültek egy csoportba és a két összevonási adatsorra korrelációt számítunk.

[13] Perturbáció esetén tetszőlegesen kicsi ($\varepsilon > 0$) értéket adunk hozzá az adatokhoz.

[14] Az összevonásnál kapott klasztertávolság vagy hasonlóság az ún. származtatott különbözőség. E helyett gyakran az átskálázott érték szerepel, azaz a maximális távolság helyett 25, a többinél pedig arányosan kisebb érték olvasható a tengelyen.

[15] A szakirodalomban ez „cophenetic" korreláció néven szerepel.

[16] Ez nem azonos a centroid-módszer szerinti klaszterezéssel.

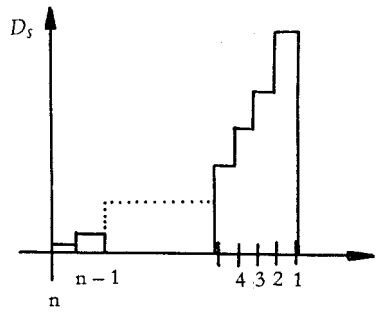
- Ha a dendrogramokat szintek szerint hasonlítjuk össze, akkor k/s arány méri a hasonlóságot, ahol s az összes szeletek száma, és k azonos dendrogramszelet fordul elő összesen.

A fentiekből látható, hogy nemcsak a dendrogramok előállítását, hanem összehasonlítását is többféleképpen végezhetjük. Fontos azonban megjegyezni, hogy sem az eljárások, sem az összehasonlító mérőszámok nem adnak egyértelmű választ arra a kérdésre, hogy hány csoportba sorolható a vizsgált adathalmaz.

A klaszterek számának megállapítása

A hierarchikus agglomeratív eljárások alkalmazása során az n pontot $(n-1)$ lépésben összevonjuk egy klaszterbe, majd a származtatott közelségi mértékek és a dendrogram tanulmányozása alapján megállapítjuk az optimális (k) klaszterszámot. Az eddigiekben ismertetett eljárások egyike sem tartalmaz optimalitási kritériumot, ezért a lehetséges megközelítéseket ebben az alfejezetben ismertetjük.

A klaszterszámot *grafikusan* és statisztikai tesztekkel állapíthatjuk meg. A grafikus megközelítés egyik változatát jelenti az, ha koordinátarendszerben ábrázoljuk az összevonási folyamatot úgy, hogy a vízszintes tengelyen a csoportok száma,[17] a függőleges tengelyen pedig a klaszterek származtatott távolsága (D_s) szerepel. (9.8. ábra) Ekkor a görbe a kezdeti lassú növekedés után exponenciálisan fog emelkedni. A folyamat lépcsős függvénnyel ábrázolható, mert a lépések számát diszkrét változó jelzi.



9.8. ábra. A klaszterek összevonási folyamata

Ha a származtatott távolság értéke nagyon megnő, amikor k helyett $(k-1)$ csoportot kapunk, akkor k a csoportok optimális száma.

Az optimális klaszterszámot kereshetjük úgy is, hogy a grafikonon a klaszterszám függvényében ábrázoljuk a származtatott távolságok megváltozását $(\triangle D_s)$. A kezdeti lassú növekedést követő gyors változás jelzi, hogy elértük az optimális klaszterszámot.

A klaszterek számának vizsgálatára számos *numerikus* lehetőséget is ajánl a szakirodalom. Ha az algoritmus kielégíti az ultrametrikus egyenlőtlenséget, azaz nem lép fel inverzió, akkor a származtatott távolságok monoton nőnek. Ezen D_s értékek sorozatára *Mojena* (1977) az alábbi tesztet javasolja. Ha azt tételezzük fel, hogy a megfigyelések nem alkotnak klasztereket, akkor a származtatott távolságok sorozata a normális eloszlás

[17] Ez eltér a tengely szokásos értelmezésétől, mert itt balról jobbra haladva monoton csökkenő érték szerepel.

jobb oldali feléből származik. Ezen értékek átlagával és szórásával standardizáljuk az egyes D_s értékeket:

$$z_D = \frac{D_s - \overline{D}_s}{S(D_s)}$$

és a normális eloszlás táblázata alapján végezzük el a tesztelést. Ha egy adott lépésben a standardizált származtatott távolság túl nagy, akkor az itt leolvasott klaszterszám nem optimális.

MANOVA-típusú statisztikai tesztekkel is megállapíthatjuk a klaszterek optimális számát. Jelölje a teljes eltérés-négyzetösszeg[18] mátrixát $\mathbf{T}$, egy adott k csoportszám mellett a csoportok közötti eltérés-négyzetösszeg mátrixa legyen $\mathbf{K}$, és a csoporton belüli eltérés-négyzetösszeg mátrixa $\mathbf{B}$, ahol $\mathbf{T} = \mathbf{K}+\mathbf{B}$. Ekkor a varianciák aránya *pszeudo-F teszttel* vizsgálható, ahol $tr\mathbf{K}$ és $tr\mathbf{B}$ a mátrixok nyoma, azaz főátlóbeli elemeik összege:

$$F^* = (tr\mathbf{K}/(k-1))/(tr\mathbf{B}/(n-k)) \tag{9.41}$$

Többváltozós normális eloszlás feltételezése mellett a próba a szóráselemzéshez hasonlóan a klaszter-átlagok egyezését vizsgálja. Ha az eloszlásra tett szigorú feltevés fennáll, akkor az F^* $p(k-1)$ és $p(n-k)$ szabadságfokú F-eloszlást követ, ahol p a változók száma. Mivel a csoportszám csökkenésével a belső eltérések négyzetösszege rohamosan nő, és így az F^* monoton csökken, a nullhipotézis elfogadása azt jelzi, hogy túl kevés csoportot képeztünk. Az optimális k csoportszámot elértük, ha e nullhipotézis elvetése után először $(k-1)$-re fogadjuk el az átlagok egyezését. Az F^* teszt jól használható, ha kevés számú és jól elkülönülő, gömb alakú klaszter van az adathalmazban.

Ha két felosztás is elfogadható a grafikus vagy numerikus vizsgálat alapján, akkor a kettő között a *Beale-féle F hányados* (F') alapján választhatunk:

$$F' = \frac{\frac{[tr\mathbf{B}_1 - tr\mathbf{B}_2]}{tr\mathbf{B}_2}}{\left[\frac{n-k_1}{n-k_2}\cdot\left(\frac{k_2}{k_1}\right)^{2/p} - 1\right]} \tag{9.42}$$

ahol $\mathbf{B}_1$ és $\mathbf{B}_2$ a két felosztás belső eltérés-négyzetösszegeinek mátrixa, és $k_2 > k_1$. Beale F-hányadosa közelítően[19] $p(k_2-k_1)$ és $p(n-n_2)$ szabadságfokú F eloszlást követ. Ha a nagyobb csoportszámú megoldás szignifikánsan jobb, akkor F' értéke viszonylag nagy, és ekkor elvetjük az eltérés-négyzetösszegek egyezését kimondó nullhipotézist.

Ha a hierarchikus klaszterezés két egymás utáni lépésében ($k_2 = 1 + k_1$) kapott felosztást hasonlítjuk össze, akkor

$$F^* = \frac{\frac{[SSB_{12} - SSB_1 - SSB_2]}{tr\mathbf{B}_2}}{\left[\frac{n-k_1}{n-k_1-1}\cdot\left(\frac{k_1+1}{k_1}\right)^{2/p} - 1\right]} \tag{9.43}$$

a számlálóban a belső eltérések növekedésének a Ward-eljárással (9.37) megegyező értéke szerepel.

A *többszörös korreláció négyzete*, a determinációs együttható is felhasználható a klaszterek elkülönülésének mérésére. A csoportok által magyarázott eltérés aránya:

$$R_k^2 = tr\mathbf{K}/tr\mathbf{T} \tag{9.44}$$

[18] Az angol elnevezés rövidítése alapján egyváltozós esetben SST, SSB és SSK jelölés is elfogadott. Most a mátrix formájú felírást használjuk, azaz például SSB helyett $\mathbf{B}$ áll.

[19] A többváltozós normális eloszlás fennállása esetén jogos az F eloszlás feltételezése.

a klaszterek közötti **K** eltérés-négyzetösszeg mátrix nyomának és a **T** mátrix nyomának a hányadosa. Ahogy csökken a klaszterek száma, úgy csökken az R^2 értéke. A hirtelen csökkenés azt jelzi, hogy az utolsó lépésben már két nagyon különböző csoportot kapcsolt össze az eljárás.

A *szemi-parciális* R^2 mutató a klaszterezés két egymást követő lépésében méri az R^2 változását:

$$\triangle R^2 = R_k^2 - R_{k-1}^2 \tag{9.45}$$

A szemi-parciális mutató úgy is előállítható, mint egy olyan hányados, amely a (9.36) első tényezőjének számlálóját a teljes eltérés-négyzetösszeg nyomához viszonyítja:

$$\triangle R^2 = (SSB_{12} - SSB_1 - SSB_2)/tr\mathbf{T} \tag{9.46}$$

Ez a mutató is a Ward-módszerhez hasonlóan méri a belső eltérések változását, és akkor is jól alkalmazható a klaszterszám megállapítására, ha nem a négyzetösszeg-eljárást alkalmazzuk. A szemi-parciális mutatót használhatjuk két különböző eljárással kapott felosztás összehasonlítására is.

A korrelációs együttható is alkalmas a klaszterezés minőségének jellemzésére, ha abból a megközelítésből indulunk ki, hogy az egy csoportba kerülő egyedek közötti eredeti távolság biztosan kisebb, mint a különböző klaszterbe soroltak közötti. Így egy felosztás minőségét mérhetjük úgy, hogy a származtatott távolságmátrixban az egy csoportba került egyedekhez nullát, a különböző csoportban levő párokhoz pedig egyet rendelünk, majd az $n(n-1)/2$ eredeti és az így származtatott (0 vagy 1) távolságok között korrelációt számolunk. A speciális értelmezés miatt ezt a mutatót *point-biszeriális korrelációnak* nevezzük. A mutató magas (egyhez közeli) értéke azt jelzi, hogy az egyes kóddal megjelölt párok közötti távolság eredetileg is magas volt, míg a zérushoz közeli korreláció arra utal, hogy az egy klaszterben levő párok egymáshoz közeliek voltak.

Konkordanciaegyüttható[20] is mérhetjük a klaszterezés jóságát:

$$\gamma = (S + D)/(S - D) \tag{9.47}$$

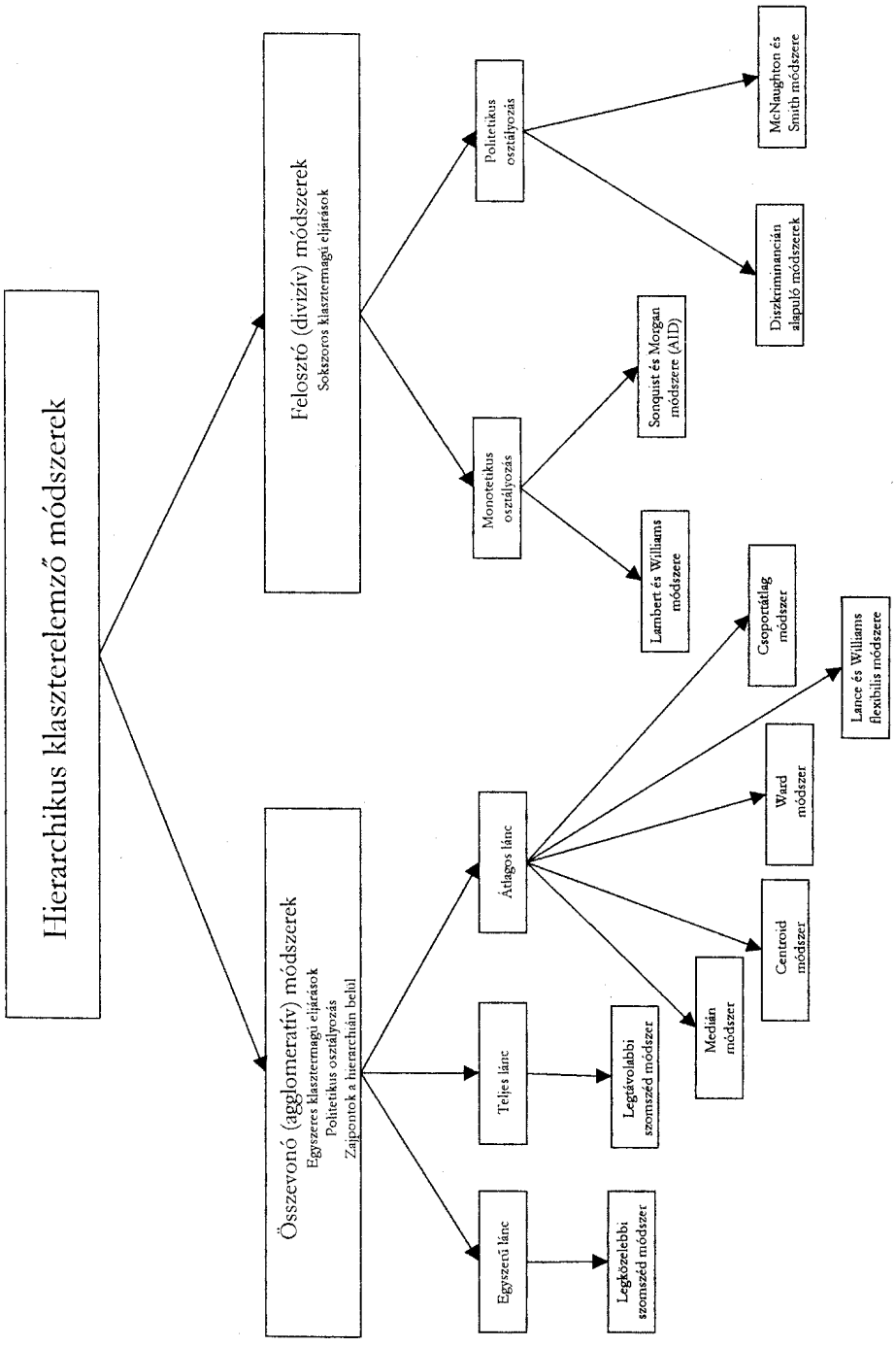
ahol S méri azt, hogy az egy csoportba sorolt elempárok közötti eredeti hasonlóság hány esetben haladja meg a különböző klaszterekben levő pontpárok közötti eredeti hasonlóságot, D pedig a fordított esetek számát adja meg. A gamma-mutató egyhez közeli értéke az eredeti hasonlóság és a besorolás közötti jó megfelelést fejezi ki. Az egyezéseket nem vesszük figyelembe.

Az egyezés a $D/(S-D)$ aránnyal is mérhető, de ekkor a nullához közeli érték fejezi ki azt, hogy a klaszterekbe sorolás jól követi az eredeti közelségeket.

A fejezetben ismertetett mutatók egy részét a statisztikai programcsomagok közvetlenül kiszámítják, a többi előállítása néhány lépésben (például korrelációszámítás vagy szóráselemzés segítségével) megoldható. A különböző eljárásokkal eőállított, és a dendrogramokról leolvasott klaszterszámok melletti „legjobb” felosztás egy vagy több mutató kiszámításával kiválasztható.

Ha szakmai ismeretek alapján előre tudjuk, hogy hány csoport van a vizsgált mintában, akkor ne alkalmazzuk az agglomeratív eljárásokat, mert azok nem alkalmasak egy várt felosztás reprodukálására. Ilyen feladatok megoldására a konkrét céltól függően számos más sokváltozós statisztikai eljárás választható.

[20] A mutató megegyezik a kereszttábláknál definiált (5.45) gamma mértékkel.



9.9. ábra. Hierarchikus klaszterező módszerek

9.6. Nemhierarchikus módszerek

Bár a hierarchikus módszerek alkalmazása a leggyakoribb, számos más klaszterező módszer is létezik. Ezek további csoportokba sorolhatók, így megkülönböztetünk

- Fuzzy-módszereket,
- optimalizáló,
- partícionáló és
- sűrűségkereső módszereket.

A *Fuzzy-módszerek* alapvetően különböznek a többi klaszterezéstől abban, hogy megengedik a klaszterek közötti átfedést,[21] míg a többi módszer diszjunkt csoportokat eredményez. Az eljárás a hasonlósági vagy távolságmátrix előállításával és az összevonáshoz szükséges minimális hasonlósági szint megadásával kezdődik. Két klasztert összevonunk, ha legalább k számú pontjuk közös. A k értékét előre rögzítjük, így az átfedés nagysága szabályozható. Ha $k = 1$, akkor átfedés nem fordul elő, és az eljárás az egyszerű lánc-klaszterezéssel azonos eredményt ad.

9.6.1. Optimalizáló módszerek

Az *optimalizáló módszerek* családjába tartozó *Q-típusú módszer* a nevét a statisztikai módszerek Catell-féle osztályozása nyomán kapta. A klaszterezés során általában a megfigyelési egységeket csoportosítjuk, azaz az induló $n \times p$ méretű $\mathbf{X}$ adatmátrix sorait vizsgáljuk. A klaszterezés kezdetén számított páronkénti hasonlósági vagy távolsági mérőszámok egy $(n \times n)$ méretű mátrix elemei. Ez a mátrix $\mathbf{X}\mathbf{X}'$ alakban írható fel, és dekompozíciója a Q-típusú[22] elemzés. A megfigyelések száma általában nagyobb, mint a változók száma, ezért az $\mathbf{X}\mathbf{X}'$ mátrix rangja legfeljebb p. Így a sajátérték-sajátvektor felbontással $k(\leq p)$ sajátvektort kapunk. A sajátvektorok azokat az együtthatókat adják meg, amelyekkel az eredeti egyedek lineáris kombinációját felírva "tiszta típusokat" kapunk. Az osztályozással az n pont mindegyikét a hozzá legközelebbi típus klaszterébe soroljuk be. Ideális esetben egy egyed csak egy típus lineáris kombinációjában szerepel nagy[23] súllyal.

A többi nemhierarchikus optimalizáló módszer a többváltozós szóráselemzés[24] alapösszefüggését használva választ döntési függvényt.

Eszerint a teljes minta szórásnégyzete két részre bontható: a csoportokon belüli átlagos eltérésre, és a csoportok közötti átlagos eltérésre.

---

[21] Ez fontos tulajdonság lehet például akkor, ha a szavakat jelentésük szerint vizsgáljuk. A többjelentésű szavak csak átfedéssel osztályozhatók.

[22] Főkomponens-elemzéskor lényegében az $\mathbf{X}'\mathbf{X}$ szorzatmátrix sajátérték-sajátvektor felbontására kerül sor, és ezt R-típusú elemzésnek hívjuk. Standardizált adatok esetén az $\mathbf{X}'\mathbf{X}/n = \mathbf{R}$, ahol $\mathbf{R}$ a változópárokra számolt korrelációk mátrixa. Részletesen a 14. fejezetben tárgyaljuk.

[23] Ha ez nem teljesül, akkor a megoldást rotáljuk. Részletesen a 14. fejezet tárgyalja a rotált megoldás előállítását.

[24] A 4. fejezet tárgyalja a szóráselemzést.

Ha **T** jelöli a teljes eltérésnégyzetösszeg-mátrixot, **B** a csoportokon belüli, **K** a csoportok közötti eltérések négyzetösszegeinek a mátrixát, érvényes a következő egyenlőség:

$$\mathbf{T} = \mathbf{B} + \mathbf{K}.$$

Mivel egy adott mintában a teljes eltérések összege állandó, kézenfekvő, hogy a klaszterező kritériumok a csoportokon belüli varianciát minimalizálják (illetve a csoportok közötti varianciát maximalizálják), ezzel homogén klaszterek kialakítására törekednek.

A **B** *mátrix nyomának*[25] *minimalizálása*

Ez a kritérium a felosztáshoz tartozó csoportokon belüli átlagtól mért eltérések négyzetösszegeinek minimalizálását jelenti. A hierarchikus eljárások közül a Ward-eljárás is a $tr(\mathbf{B}) \longrightarrow \min$ döntésfüggvényt használja.

Azok a módszerek, amelyek az egyedek besorolásánál a legközelebbi középpontokat (euklideszi értelemben) veszik figyelembe, implicit módon ezt a kritériumot alkalmazzák. Ez a módszer jól felismeri a szférikus (gömbölyű) struktúrát, az ellipszis alakú csoportokat viszont szétbontja.

A **B** *mátrix determinánsának minimalizálása*

A $|\mathbf{B}| \longrightarrow \min$ kritériummal ekvivalens a $|\mathbf{B}|/|\mathbf{T}| \longrightarrow \min$ kritérium. Ez a mutató Wilks-féle lambdaként (λ) vált ismertté, és a diszkriminanciaelemzésben használatos.

A fentivel ekvivalens kritérium még az

$$|\mathbf{I} + \mathbf{B}^{-1}\mathbf{K}| \longrightarrow \max \tag{9.48}$$

vagy a

$$\prod_i (1 + \lambda_i) \longrightarrow \max \tag{9.49}$$

kritérium (ahol

$$\lambda_i \text{ a } |\mathbf{K} - \lambda\mathbf{B}| = 0 \tag{9.50}$$

determinánsegyenlet gyöke).

A $|\mathbf{B}| \longrightarrow \min$ kritérium az ellipszis alakú klasztereket ismeri fel jól.

Általában az azonos alakú klaszterek elkülönítését végzi jól el ez az eljárás.

A $(\mathbf{K}\,\mathbf{B}^{-1})$ *mátrix nyomának maximalizálása*

Ez a kritérium a Hotelling-féle nyomkritérium, és egyenlő a $\sum_i \lambda_i$ maximalizálásával, ahol a λ_i-k a (9.50) gyökei. Használatát Friedman és Rubin javasolta 1967-ben, mert a sajátértékek invariánsok az adatok nemszinguláris lineáris transzformációjára. A $(\mathbf{K}\,\mathbf{B}^{-1})$ mátrix sajátértékei azt mérik, hogy a csoportok közötti változékonyság hogyan aránylik a csoportokon belüli eltérésekhez, ezért a legjobban elkülönülő csoportokhoz nagy sajátértékek tartoznak. A sajátértékek segítségével tehát a felosztás jóságát tudjuk

[25] Egy mátrix nyomán a mátrix diagonális elemeinek összegét értjük, jelölése $tr(\mathbf{B})$.

mérni. Az összes lehetséges felosztás előállítása (és a maximális sajátértékűek kiválasztása) azonban nagyon nagy számításigényű. Ezért a globálisan optimális megoldás keresése helyett egy induló felosztásból[26] kiindulva lépésről lépésre javítjuk a fenti célfüggvények egyike szerint az egyedek klaszterbe sorolását.

9.6.2. Sűrűségkereső módszerek

A *sűrűségkereső módszerek* akkor alkalmazhatók, ha a mintában természetes klaszterek fordulnak elő. Ekkor a metrikus tér egyes részein a pontok nagyon közel vannak egymáshoz, és ezeket a klasztereket kisebb sűrűségű területek választják el egymástól. A típusalkotás célja olyan klaszterek körülhatárolása, ahol a térben a pontok koncentrációja viszonylag sűrű. A klaszterezéshez használt döntésfüggvény itt az ismeretlen elméleti sűrűségfüggvény becslésén alapul. A sűrűségfüggvényt a mintából becsülve gradiens-módszerrel keressük a lokális maximumokat, és a sűrűségfüggvény csúcsai adják a klaszterek súlypontjait.

A sűrűségkereső módszerek többsége, így a móduszelemzés és az osztályozó térkép-(taxmap) módszer is az egyszerű lánc-elvet követi a sűrűségfüggvény becslése helyett. Mindkét eljárás során a lánchatás legyőzése a cél.

A *móduszelemzés*[27] a sűrűségfüggvény lokális maximumhelyei környezeteként értelmezi a klasztereket, és a kezdeti felosztást a sűrűsödési pontok alapján határozza meg. Ehhez minden pont körül r sugarú környezetet határozunk meg, és megszámoljuk, hogy hány pont esik az egyes tartományokba. Amelyik pont környezetében legalább s számú pont fordul elő, azt sűrűsödési pontnak tekintjük. Azok a pontok, amelyek minden sűrűsödési ponttól r-nél távolabb vannak, saját klasztert alkotnak. Az r és az s értékét előre rögzítjük, így kialakul a kezdeti klaszterezés. Ha $s > k$, azaz a sűrű pontok száma meghaladja a klaszterek számát, vagy a klaszterek közötti távolság egy küszöbszám alatt marad, akkor folytatjuk az összevonást. A felosztás r növelésével megismételhető, így mind több pont válik sűrűvé.

Minden új sűrű pont után meg kell vizsgálni, hogy a többi sűrű ponttól milyen távol van:

- ha az új pont távolsága a többi sűrű ponttól nagyobb mint r, akkor az új sűrű pont egy új klaszter magpontja lesz, így a klaszterek száma eggyel növekszik;
- ha az új pont egy vagy több azonos klaszterbe tartozó sűrű ponttól r-nél kisebb távolságra van, akkor az új pont ahhoz a klaszterhez fog tartozni;
- ha az új pont egy vagy több különböző klaszterbe tartozó sűrű ponthoz esik r-nél közelebb, akkor a szóban forgó klasztereket összevonjuk;
- minden ciklusban megvizsgáljuk, hogy a különböző klaszterekhez tartozó sűrű pontok közötti legkisebb távolság kisebb-e, mint egy küszöbérték. Ha igen, akkor a két klasztert összevonjuk.

Az *osztályozó térkép-módszer* a két legközelebbi egyed összevonása után az egyszerű lánc-elv alapján meghatározza a többi egyed és a klaszter távolságát. Ebből a mát-

[26] Önkényesen vagy a hierarchikus klaszterezés eredményét tanulmányozva választjuk meg a csoportok számát, k-t.

[27] A móduszelemzést kidolgozójáról Wishart-eljárásként is említi az irodalom. Egyváltozós esetben a leggyakrabban előforduló érték a módusz.

rixból kiválasztja azt a pontot, ami az első klaszterhez a legközelebb van, és a három pont közötti távolságok átlagát veszi. Ha ez az átlagos távolság egy előre megadott küszöbszámnál[28] jobban meghaladja az első két pont közötti távolságot, akkor a pontok nem alkotnak egy klasztert. A harmadik pont új klaszter kezdőpontja lesz, és a folyamat az új klaszterre megismétlődik.

Vannak olyan módszerek is, ahol a döntési függvény a *valószínűségeloszlások keverékének szétválasztásán* alapuló osztályozást eredményez. A kevert modellek nagy minta esetén alkalmazhatók jól, amikor ismert típusú, de a paraméterekben különböző sűrűségfüggvényt tételezünk fel az egyes csoportokban. A feltételezés szerint a teljes minta többváltozós eloszlása a csoportok többváltozós eloszlásainak konvex lineáris kombinációja:

$$f(\mathbf{x}) = \sum_{s=1}^{k} \lambda_s \cdot f_s(\mathbf{x}), \quad \text{ahol } 0 \le \lambda_s \le 1 \text{ és } \sum_{s=1}^{k} \lambda_s = 1. \tag{9.51}$$

Wolfe algoritmusa a klaszterstruktúrát többváltozós normális eloszlás feltételezése mellett maximum likelihood becsléssel tárja fel. Az egyedeket ahhoz a csoporthoz sorolja, amelyhez a legnagyobb valószínűséggel tartoznak. A

$$P(s \mid \mathbf{x}) = \frac{\lambda_s f_s(\mathbf{x}, \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)}{f(\mathbf{x})} \tag{9.52}$$

annak a valószínűsége, hogy $\mathbf{x}$ egyed az s-edik klaszterbe tartozik. A klaszter várhatóérték-vektora $\boldsymbol{\mu}_s$, kovarianciamátrixa $\boldsymbol{\Sigma}_s$, a sűrűségfüggvénye f_s:

$$f_s(\mathbf{x}, \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s) = (2\pi)^{-1/2} \cdot |\boldsymbol{\Sigma}_s|^{-1/2} \exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_s)' \boldsymbol{\Sigma}_s^{-1} (\mathbf{x} - \boldsymbol{\mu}_s)\} \tag{9.53}$$

Érdemes megemlíteni a fenti modellhez szervesen kapcsolódó latens struktúraelemzés (Lazersfeld) modelljét. A klasszikus logika osztályozási modelljéhez hasonlóan ez a modell is lineárisan kódolt kvantitatív tulajdonságokkal foglalkozik. A latens struktúraelemzésben az egyedek egy osztályát homogénnek tekintjük, ha az egyedek változóértékei korrelálatlanok.

Tehát a populáció a latens osztályok keverékének számít itt is. Az algoritmus feladata szétválasztani minden homogén osztályt. Wolfe megmutatta, hogy a latens struktúraelemző modell az általános „eloszlások keveréke” modell diszkrét eloszlásokra alkalmazott változata:

$$f_s(\mathbf{x}, \boldsymbol{\mu}_s) = \prod_{i=1}^{p} \mu_{si}(1 - \mu_{si})^{1-x_i}, \quad \text{ahol } x_i = 0 \text{ vagy } 1, \text{ és } \mu_{si} = P(x_i = 1|s) \tag{9.54}$$

9.6.3. Legközelebbi centroid-módszerek

A nemhierarchikus módszerek közül a leggyakrabban alkalmazott és a hierarchikus klaszterezéshez a leghasonlóbbak a diszjunkt klasztereket előállító legközelebbi centroid-módszerek. A különböző eljárások általános menete a következő:

- a kezdő klaszterek kialakítása és az egyedek szétosztása a kezdő klaszterekbe,
- az egyedek átsorolása a klaszterek között.

[28] Ez a szám a diszkontinuitás mértéke, szokásos értéke 0,5.

Az első és a második lépés végrehajtása többféleképpen történhet, ezért több eljárásváltozat ismert.

A *kezdő klaszterek kialakítását* a csoportok k számának és a k klaszterközéppontnak a megadásával kezdjük. A megfelelő k megválasztása szakmai tapasztalaton vagy korábbi statisztikai elemzésen (pl. hierarchikus klaszterezésen) alapulhat.

- McQueen a megfigyelések közül az első k egységet választja magpontnak.
- Megszámozhatjuk az egyedeket, és véletlenszerűen választhatunk belőlük k számút.
- Választhatunk k mesterséges egységet, amelyek nem feltétlenül elemei az adatrendszernek.
- Valamilyen módszer szerint (és ez lehet valamilyen hierarchikus eljárás) felosztjuk a mintát k csoportra, és a csoportok középpontjait tekintjük magpontnak.
- Astrahan-módszere komplexebb:
  - megszámolja, hogy az egyedektől d_1 távolságon belül hány egyed helyezkedik el, vagyis kiszámítja a megfigyelési egységek szerinti sűrűséget,
  - kiválasztja a legnagyobb sűrűségű egyedet, ez lesz az első magpont,
  - kiválasztja a csökkenő sűrűségnek megfelelően azt az egyedet, amelynek az előzőektől mért távolsága legalább d_2,
  - ezt az eljárást folytatja mindaddig, amíg már csak zéró sűrűségű pontok maradnak.

    Ha a magpontok száma nagyobb a kívánatosnál, hierarchikus klaszter-módszerrel csökkentjük a klasztermagok számát. Astrahan módszerében problémát okoz d_1 és d_2 helyes megválasztása.

A kezdő klaszterek kialakításának *második szakasza* a magponthoz rendeli az egységeket. McQueen módszere az egyedeket a legközelebbi centroid-módszer szerint rendezi a magpontokhoz. A kapott klasztereknek a centroidjait tekinti azután magpontnak, és így ismétli az eljárást. Ezzel függetleníti az első véletlenszerű magpontválasztástól a felosztást.

Legközelebbi centroid-módszerek – állandó számú klaszterrel

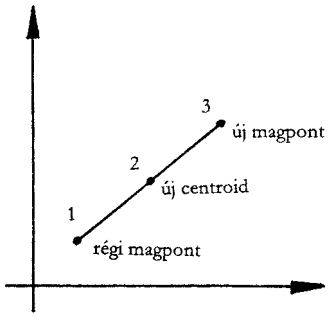
A klaszterező algoritmusok következő lépése a kezdő klaszterek egyedeinek átcsoportosítása valamilyen kritérium szerint, azzal a feltétellel, hogy minden egyed egyszerre egy és csak egy klaszterbe tartozzon (vagyis a klaszterek egymást kizáróak legyenek), és a klaszterek száma az eljárás során ne változzon. Az egyszerű iteratív módszerek kétlépéses folyamatból állnak. Először a magpontok kiszámítása a klaszterek centroidjaként, majd a klaszterek kialakítása az egyedeknek a legközelebbi klasztermaghoz való besorolásával. Ez az eljárás mindaddig folytatódik, amíg a felosztás nem állandósul.

Forgy módszere nagyon egyszerű, és általában kevés iterációs lépésre van szükség az optimum eléréséig.

1. Kezdő magpontokkal indulunk.
2. Minden egyedet hozzásorolunk a legközelebb lévő magpontú klaszterhez.
3. Kiszámítjuk a klaszterek centroidjait, és ezeket tekintjük az új magpontoknak.
4. A 2. és 3. lépést addig ismételjük, amíg a folyamat nem konvergál, vagyis amíg a klasztertagság állandó nem marad.

Jance Forgytól függetlenül javasolt egy módszert, amely csak a 3. lépésben tér el az előzőtől.

Jance a régi magpontból az új centroidhoz húzott egyenes (a gradiens) mentén mozdul el. A javulás gyorsítható, ha az új magpontot erre az egyenesre tükrözzük (9.10. ábra). Az eredmény itt is független lesz a kezdő magpontoktól, mivel az új centroidokat csak az egyedek besorolása után számítjuk. A két módszer implicit módon minimalizálja a klaszteren belüli eltérés-négyzetösszegeket.



9.10. ábra. Jance-féle magponttükrözés módszere

McQueen-féle k középpontú módszer

A McQueen-féle eljárás a következő:

1. Kiindulunk az első k adatból, mint magpontból.
2. Az egyedeket ahhoz a klaszterhez soroljuk, amely középpontjához a legközelebb esnek. A klaszterközéppontokat minden egyed besorolása után újra számítjuk.
3. Miután minden egyedet besorolunk valamelyik csoportba, az új középpontokat megfeleltetjük a magpontoknak, és az adatokat újra hasonlítjuk a magpontokhoz.

A 2. és 3. lépést addig folytatjuk, amíg a klasztertagság nem változik.

Láthatjuk, hogy ez az eljárás annyiban különbözik Forgy módszerétől, hogy minden egyes egyed valamelyik csoporthoz csatolása után az illető klaszter centroidja (és azé is, amelyikből az egyedet átsoroltuk) megváltozik. A McQueen-féle módszer egyszerű, gazdaságos és nagyon népszerű. A statisztikai programcsomagok is általában ezt tartalmazzák.

Legközelebbi centroid-módszerek – változó számú klaszterrel

A klaszterek számának előzetes megadása az előző három módszer alkalmazási körét szűkíti, ugyanis a gyakorlatban sok esetben az adatok természetes struktúrájának ismerete hiányában ezt pontosan nem tudjuk megadni, így egy rosszul definiált klaszterszámmal hamis klaszterstruktúrát adhatnak a módszerek:

- felbonthatunk homogén klasztereket, vagy
- összevonhatunk két elkülönülő klasztert.

Ezért hasznos a következő három módszer, amely bizonyos rugalmasságot biztosít a klaszterek számának kialakítására.

McQueen-féle k középpontú módszer paraméterek becslésével

A már ismert k középpontú módszer módosított változatáról van szó. A módszer a következő lépésekre bontható:

1. Három paraméternek választunk értéket: k a kezdő klaszterek száma, c a klaszterek legnagyobb sugara, r az új klaszter legkisebb távolsága.
2. Vesszük az első k egyedet, mint a kezdő klaszterek egy-egy elemét.
3. Kiszámítjuk a k egyed közötti távolságokat. Ha a legközelebbi távolság kisebb, mint a megadott c érték, akkor összevonjuk a két klasztert, és kiszámítjuk az új centroidot. Az összehasonlítást addig végezzük, amíg minden centroid legalább c egységre van egymástól.
4. A megmaradó $n - k$ számú egyedet a legközelebbi középpont klaszterébe soroljuk. Minden besorolás után a klaszter centroidját újraszámoljuk, a többi klasztercentroidoktól mért távolságokat is. Ha két klaszter középpontja közelebb kerül, mint c, akkor a két csoportot összevonjuk. Ha a legközelebbi középpontú klaszter és a besorolandó egyed közötti távolság nagyobb, mint az r paraméter, akkor ezt az egyedet egy új klaszter magpontjaként kezeljük.

Ha a 4. lépést végrehajtjuk az $n - k$ egyedre, akkor megkapjuk az adatrendszerben természetesen előforduló klaszterek középpontjait és a hozzájuk közel eső egyedeket. A módszer ezzel a beosztással véget ér, iteratív javítást nem alkalmaz. A c és r paraméterek helyes megválasztásán sok múlik, így célszerű azokat módosítva az eljárást újra végrehajtani.

A k középpontú módszer Wishart-féle változata

1. Kiindulunk egy kezdő felosztásból, és kiszámítjuk az induló klaszterek középpontjait.
2. Legyen k a klaszterek aktuális száma. Számítsuk ki valamennyi egyed valamennyi centroidtól mért távolságát:
   a) ha a legkisebb távolság ezek közül nagyobb, mint egy t küszöbérték, akkor ezt az egyedet outliernek tekintjük, és annak a csoportnak újraszámítjuk a középpontját, ahonnan kivettük;
   b) ha a legközelebbi centroid nem a saját klaszterének a középpontja és nem nagyobb a távolság, mint t, akkor az egyedet átsoroljuk a legközelebbi középpontú klaszterbe, és mindkét klaszter középpontját újraszámítjuk;
   c) ha az outlierek közül van olyan egyed, amelyiknek a távolsága a legközelebbi középponttól kisebb, mint t, akkor ezt az egyedet besoroljuk az adott klaszterbe, és a centroidot újraszámítjuk.
3. Miután minden egyedet elosztottunk a 2. lépésben definiáltak szerint, az outlierek között klasztereket alakítunk ki úgy, hogy minden kialakított klaszter elemszáma legfeljebb s legyen.
   A lépés célja: az elkülönülő klaszterek számának csökkentése.
4. A 2. és 3. lépést addig ismételjük, míg a felosztás konvergens nem lesz, vagyis egyetlen egyed sem változtatja meg a helyét a 3. lépésben, vagy a lépések száma el nem ér egy maximális c_1 küszöbértéket.
5. Kiszámítjuk a klaszterek közötti hasonlóságokat, és összevonjuk a két leghasonlóbbat. Ezután a 2–5. lépést ismételjük addig, amíg új felosztást nem kapunk. A klaszterek összevonását addig végezzük, amíg a klaszterek száma el nem ér egy minimális c_2 küszöbértéket.

Kanonikus klaszterelemzés[29]

Ebben a fejezetben az eddigiektől eltérően értelmezett kanonikus távolságfogalom alapján végezzük a klaszterezést.

Kanonikus távolságon a két halmaz pontjaira felírt lineáris kombináció minimális távolságát értjük:

$$D_n(C_i, C_j) = \min_{V,W} d \left\{ \mathbf{V} = \sum_{m=1}^{n_i} b_m \mathbf{x}_{im},\ \mathbf{W} = \sum_{k=1}^{n_j} c_k \mathbf{y}_{jk} \right\}, \tag{9.55}$$

ahol $\mathbf{x}_{im} \in C_i,\ m = 1, \ldots, n_i,$
$\mathbf{y}_{jk} \in C_j,\ k = 1, \ldots, n_j,$
$\sum_{m=1}^{n_i} b_m^2 = 1$ $(b_m \geq 0)$ és $\sum_{k=1}^{n_j} c_k^2 = 1$ $(c_k \geq 0)$ a normalizálási feltételek.

A kanonikus távolság a két ponthalmaz ún. kanonikus pontjai közötti távolságot méri. A halmaz kanonikus pontja:

a) a halmaz pontjainak a normalizálási feltétellel számított lineáris függvénye,
b) minimális távolságra van egy másik halmaz kanonikus pontjától.

A kanonikus távolság tulajdonságai a következők:

- az egyik halmaz kanonikus pontja a másik halmazhoz tartozó pontok függvényében változik, így a kanonikus pontok közötti távolság a halmazok struktúráját tükrözi,
- a kanonikus távolság az előző távolságok általánosításának tekinthető,
- a kanonikus távolság nem lehet kisebb, mint a legközelebbi szomszéd-távolság, és nem lehet nagyobb, mint a legtávolabbi szomszéd-távolság,
- a p dimenziós térben mért kanonikus távolság leképezhető a $(p+1)$ dimenziós térben meghatározott, a (ponthalmazok közötti) kanonikus korrelációból. Innen származik az elnevezés.

A kanonikus klaszterelemzés tulajdonságai:

- a klaszterezés során kialakuló csoportok a lánchatáshoz hasonló sávhatást tartalmaznak „szerpentin” jelleggel,
- folytonos leképezés, mivel a pontok kis változtatása kis változást eredményez a klaszterek struktúrájában,
- a kanonikus klaszterelemzés a többi klaszterező módszerrel szemben a ponthalmaz elemeihez különböző súlyokat rendel a viszonyítási halmaz pontjai függvényében, így a pontok halmazon belüli „értékelése” (a reprezentáns pontok előállításában játszott szerepe, súlya) lépésenként különböző lehet.

A kanonikus távolság számítása

A két ponthalmaz kanonikus pontjai közötti távolság minimumát keressük. A két kanonikus pont $\mathbf{V}$ és $\mathbf{W}$, közöttük az euklideszi távolság négyzetének minimumát:

$$d^2 = (\mathbf{V} - \mathbf{W})'(\mathbf{V} - \mathbf{W}) = (\mathbf{X}\,\mathbf{b} - \mathbf{Y}\,\mathbf{c})'(\mathbf{X}\,\mathbf{b} - \mathbf{Y}\,\mathbf{c}) \longrightarrow \min, \tag{9.56}$$

[29] Ezt az eljárást Füstös László dolgozta ki

a lineáris kombináció együtthatóira vonatkozó feltételek esetén

$$\mathbf{b}'\mathbf{b} = 1, \quad b_i \geq 0,$$
$$\mathbf{c}'\mathbf{c} = 1, \quad c_j \geq 0.$$

A feladatot a Lagrange-féle multiplikátor-módszerrel oldjuk meg. A Lagrange-függvény

$$F = (\mathbf{X}\mathbf{b} - \mathbf{Y}\mathbf{c})'(\mathbf{X}\mathbf{b} - \mathbf{Y}\mathbf{c}) - \rho(\mathbf{b}'\mathbf{b} - 1) - \mu(\mathbf{c}'\mathbf{c} - 1), \tag{9.57}$$

vagy a szorzást elvégezve:

$$F = \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} - \mathbf{c}'\mathbf{Y}'\mathbf{X}\mathbf{b} - \mathbf{b}'\mathbf{X}'\mathbf{Y}\mathbf{c} + \mathbf{c}'\mathbf{Y}'\mathbf{Y}\mathbf{c} - \rho(\mathbf{b}'\mathbf{b} - 1) - \mu(\mathbf{c}'\mathbf{c} - 1). \tag{9.58}$$

A szélsőértéket megkapjuk, ha a parciális deriváltakat nullával tesszük egyenlővé:

$$\frac{\partial F}{\partial \mathbf{b}'} = 2\mathbf{X}'\mathbf{X}\mathbf{b} - 2\mathbf{X}'\mathbf{Y}\mathbf{c} - 2\rho\,\mathbf{b} = \mathbf{0}. \tag{9.59}$$

$$\frac{\partial F}{\partial \mathbf{c}'} = 2\mathbf{Y}'\mathbf{Y}\mathbf{c} - 2\mathbf{Y}'\mathbf{X}\mathbf{b} - 2\mu\,\mathbf{c} = \mathbf{0}. \tag{9.60}$$

A (9.59) és (9.60) egyenletekben a Lagrange-függvény konstansait λ-val helyettesítve:

$$\left(\begin{bmatrix} \mathbf{X}'\mathbf{X} & -\mathbf{X}'\mathbf{Y} \\ -\mathbf{Y}'\mathbf{X} & \mathbf{Y}'\mathbf{Y} \end{bmatrix} - \lambda\mathbf{I}\right)\begin{bmatrix} \mathbf{b} \\ \mathbf{c} \end{bmatrix} = \mathbf{0},$$

vagyis a kanonikus változók együtthatóit (a két halmaz pontjainak súlyait) egy sajátérték-, sajátvektor-feladat megoldásaként kapjuk. A Lagrange-függvény lokális minimumhelyei közül a nem negatív sajátvektorok adják feladatunk megoldásait.

A k-középpontú klaszterezés értelmezése

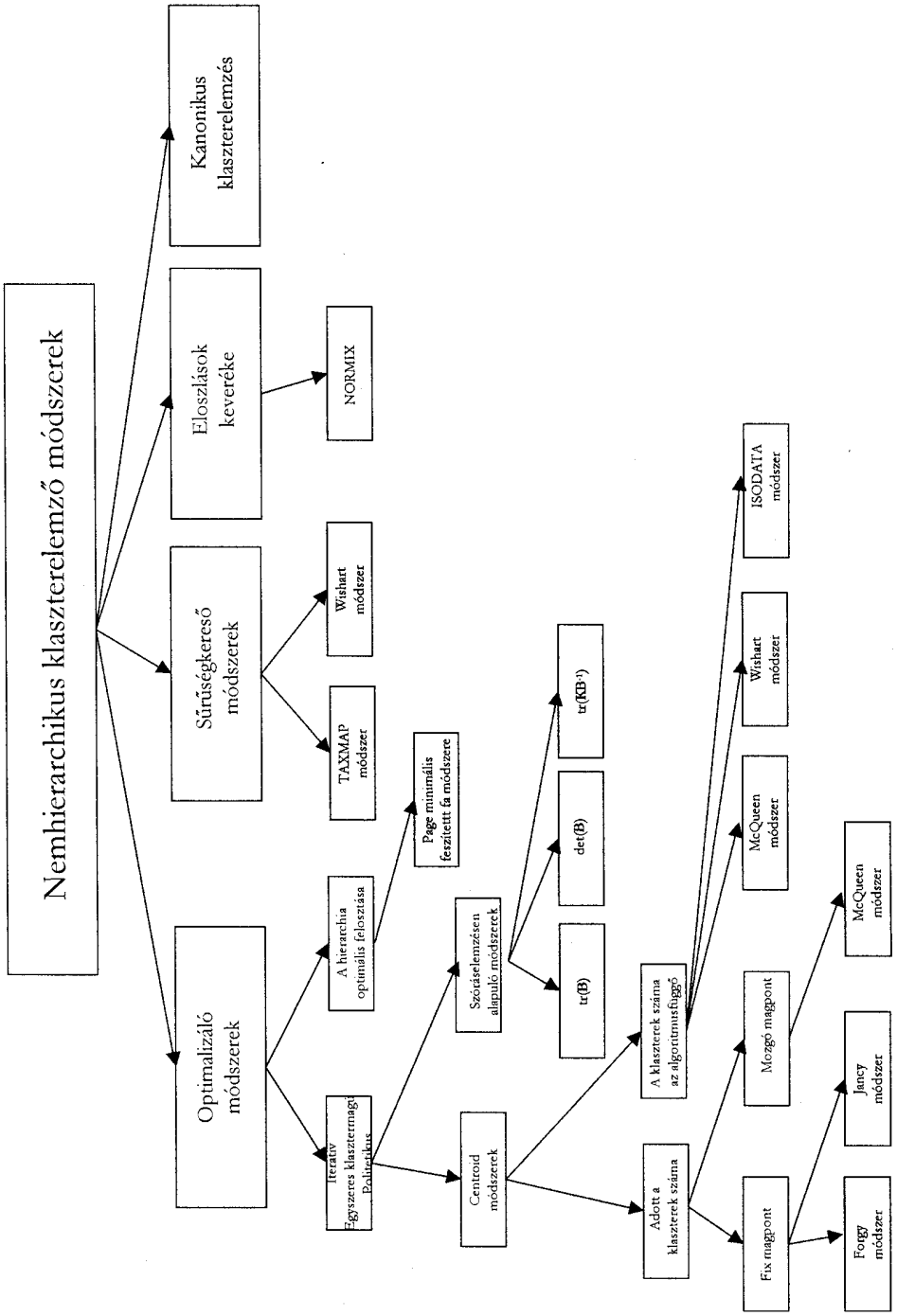
A feltételezett klaszterszám mellett elkészített felosztás értelmezését az egyedek és a változók klaszterenkénti vizsgálatával végezhetjük.

Az egyedek arányos szétosztása a klaszterek között nem követelmény, de a nagy aránytalanság fontos információt hordoz. Az egyelemű klaszterek a kilógó, a többiektől nagyon eltérő tulajdonságú egyedek létére figyelmeztetnek. A nagy elemszám pedig arra hívja fel a figyelmet, hogy érdemes a csoportszám növelésével megismételni a klaszterezést.

A klaszterközéppontok és a köztük levő távolságok előállítása is segíti az értelmezést és a klaszterek megkülönböztetését. Ezt kiegészíthetjük azzal, hogy az egyes egyedeknek a saját klaszterük középpontjától mért távolságát is meghatározzuk. A távolságok alapján dönthetünk az egyes csoportok szétvágásáról vagy összevonásáról, azaz a k növeléséről vagy csökkentéséről.

Az egyedek osztályozásán túl vizsgálható az is, hogy a figyelembe vett p változó mindegyike jelentős szerepet játszott-e a klaszterek megkülönböztetésében. Az egyes klaszterek varianciáit kiszámolva a csoportok alakját hasonlíthatjuk össze, mivel az azonos variancia-kovarianciamátrix azonos alakot jelez. A szóráselemzés (F-próba)[30] segítségével kiválaszthatjuk a csoportokat elkülönítő változókat, és így akár dimenziócsökkentést is végrehajthatunk a következő lépésben.

[30] Csak leíró, és nem tesztként való alkalmazásról van szó, mert a matematikai előfeltételek nem teljesülnek.



9.11. ábra. Nemhierarchikus módszerek

9.7. A klaszterelemzés eredményének értékelése

A klaszterelemzés eredményének értékelését alapvetően az nehezíti, hogy többdimenziós térben az adathalmaz nem ábrázolható. A klaszterek meghatározását úgy végezzük, hogy nem ismerjük a csoportok számát, sőt még létezésüket is csak feltételezzük. A klaszterezés során tehát exploratív elemzést készítünk, nem következtetünk annak a sokaságnak a szerkezetére, amiből a megfigyelések származnak.

A klaszterezés eredménye nagymértékben függ a kiválasztott változóktól és mérőszámoktól, az alkalmazott eljárástól, és – ha van egyáltalán – a háttérben rejlő klaszterstruktúra jellegétől. Az egyes eljárások implicit feltevéseket fogalmaznak meg a minta struktúrájáról, és ezt próbálják meg előállítani ahelyett, hogy a valódi tagozódást tárnák fel.

Sem a hierarchikus, sem a nemhierarchikus klaszterezéshez nem tartozik célfüggvény, amivel az osztályozás jósága egyértelműen mérhető, és nincsenek szigorú matematikai feltételek, amelyek teljesülése ellenőrizhető.

A fentiek ellenére számos szakkönyv foglalkozik a klaszterelemzés és a klaszterek értékelésének problémájával, és általában csak annyit állítanak, hogy bizonyos módszerek jól alkalmazhatók bizonyos adatokra. A továbbiakban összefoglaljuk azokat a követelményeket,[31] amelyek támpontot jelenthetnek a kapott eredmények értékelésében.

- Nyilvánvaló kívánalom, hogy a klaszterezés eredménye független legyen a megfigyelések sorrendjétől.
- Jól definiáltak legyenek a klaszterek abban az értelemben, hogy azonos megfigyelt adatokból azonos felosztást kapjunk. Az egyenlő távolság, illetve hasonlósági értékek közötti önkényes választás miatt ez a tulajdonság több eljárásnál nem teljesül.
- A *folytonosság* követelménye is megfogalmazódik abban az értelemben, hogy az adatokban bekövetkező kis változások kis változást eredményezzenek a felosztásban. Itt külön kell vizsgálnunk a különbözőségi mérőszámok kis változását, valamint azt, ha az eredeti adatokban fordul elő „hiba”, mert kis eltérések is nagy változást okoznak egyes távolságmértékekben.
- A *stabilitás* követelménye azt jelenti, hogy ha egy egyedet elveszünk vagy hozzáadunk a megfigyelésekhez, akkor az osztályozásban nagyon kis változás következzen be. Ez lánchelyzetű pont esetében nem feltétlenül teljesül. A stabilitási követelmény részének tekinthető az az elvárás is, hogy ha egy klaszter minden egyedét (hierarchikus esetben a dendrogram egy ágát) kihagyjuk, akkor a többi elem tagozódása invariáns legyen erre a változtatásra. Williams és szerzőtársai (1971) a kilógó pontok elhagyására való érzéketlenséget is a stabilitás részének tekintik.
- Gyakori követelmény, hogy az osztályozás eredménye invariáns legyen a különbözőségek monoton transzformációjára. Itt említjük meg az adatok lineáris transzformációjára való invariancia követelményét is, amely például a standardizált adatok használatát teszi lehetővé. Ha a vektorok hajlásszögének koszinuszából számítunk távolságot, akkor a pontok közötti távolság nem arányosan változik.
- A klaszterek *érvényessége* (validitása) négy kritérium alapján vizsgálható. *Külső* követelményként értelmezhető az, ha ismert csoportokba tartozó egyedekből veszünk mintát, és arra végezzük el a klaszterezést. *Belső* követelménynek tekint-

[31] A követelmények többségét Jardine és Sibson (1971) fogalmazta meg. Más szerzők nevét külön megemlítjük.

hetők azok a mutatók, amelyekkel az eredeti és a származtatott távolságok illeszkedését mérjük.[32] Harmadik megközelítést jelent a *megismételhetőség* kritériuma, amelynek lényege a kettéosztott megfigyelések klaszterezése és a felosztások összevetése. A klaszterek érvényességének *relatív* kritériuma az adatmátrix több eljárás szerinti klaszterezését és a felosztások közötti egyezés mérését[33] fogalmazza meg.

Ha több eljárás egyidejű alkalmazásával azonos csoportosítást kapunk, akkor ez nagymértékben fokozza az eredmények érvényességébe vetett hitet. Fontos azonban ismét aláhúzni azt az empirikus elemzésekkel bizonyított tényt, hogy csak jól elkülönülő és gömb alakú struktúrák esetében tekinthetjük az egyező felosztásokat úgy, mint amelyek a természetes csoportok létét igazolják.

- A *robusztusság* követelménye a kilógó pontok hatásának csökkentését jelenti. Ha több nem tipikus, „távoli" pont van a mintában, akkor ezek jelentősen befolyásolhatják a felosztást olyan eljárások esetében, amelyek a belső eltérés-négyzetösszeget minimalizálják, vagy a kanonikus távolságot állítják elő. Ilyenkor a csoportokon belüli azonos kovariancia-struktúra feltevése téves lehet, pedig az optimalizáló eljárások csak azonos alakú csoportok feltárására alkalmasak.

A klaszterelemző módszerek és a számítógépes eljárásváltozataik alkalmazásával kapott csoportosítások értelmezése és értékelése nagy szakmai felkészültséget és körültekintést igényel. Érdemes más sokváltozós módszereket, például a sokdimenziós skálázást és a diszkriminanciaelemzést is végezni, hogy a minta szerkezetéről megbízható megállapításokat fogalmazhassunk meg.

9.7.1. A klaszterelemző módszerek problémáiról

Szólni kell az outlierek szerepéről az egyes módszereknél.

Outlier az adatrendszer olyan egyede, amely minden más egyedtől, illetve csoporttól elkülönül. Az outliereket szokták zajelemeknek is nevezni, mivel a homogén struktúra kialakítását megnehezítik. Kérdés, hogy az egyes eljárások mennyire képesek kiszűrni az outlierek hatását, illetve magukat az outliereket az elemzés során.

Az egyszintű felosztó módszerek minden egyedet besorolnak valamilyen csoportba, így nem szűrik ki az outliereket.

Az egyszintű optimalizáló módszerek egyedülálló magpontokként, illetve a paraméterbecsléssel dolgozó eljárások maradékként elkülönítik az outliereket. (Éppen emiatt az előnyös tulajdonságuk miatt nem biztosított ezen módszerek konvergenciája.)

A hierarchikus agglomeratív módszerek végeredményeként egy csoportba egyesítenek minden egyedet. Az adatrendszer struktúrájáról a hierarchia egyes szintjeit elemezve kapunk képet. Amennyiben a hierarchiát mutató dendrogramot a kívánt szinten elvágjuk, és van más egyeddel vagy csoporttal össze nem kapcsolt egyed, az outlier.

Természetes, az outlierek száma itt függ az elvágott dendrogram hasonlósági szintjétől, hiszen kezdetben minden egyed outlier.

---

[32] Részletesen a 9.4.1.-ben tárgyaltuk. Millian (1981) vizsgálatai alapján a point-biszeriális és a gamma-mutató a leghasználhatóbb belső kritériumként.

[33] A mutatókat a 9.4.1. fejezet ismertette.

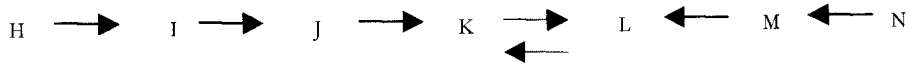
A divizív eljárások általában nem szűrik ki az outliereket, mivel egyetlen, minden elemet tartalmazó csoport felosztásával dolgoznak, és az eljárás az outlierek elkülönítése előtt befejeződhet. Így a zavaró pontok torzíthatják a struktúrát.

A kutató egyik igen fontos döntése, hogy a hierarchikus módszerek milyen típusát választja a csoportok közötti kapcsolatok értelmezésére. A két csoport összevonása a hierarchián belül a két csoport közötti „összekötő lánc" definiálását jelenti.

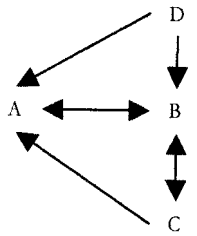
Ez a lánc lehet

- egyszerű lánc,
- teljes lánc,
- átlagos lánc.

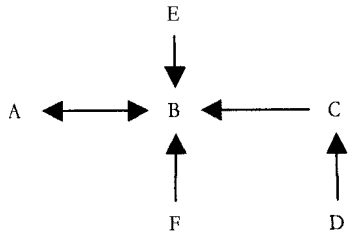
Az egyszerű lánc módszerek az összekapcsoláskor használható láncok közül a minimális lánccal kötik össze a két csoportot. (Ezek a single linkage módszerek.) A kialakuló klaszterek „szerpentin" jellegűek, mivel minden egyednek elég csak a klaszter egy egyedéhez hasonlónak lenni.



A teljes lánc módszerek a csoportok közötti maximális hosszúságú lánc alapján választják ki az összevonandó csoportokat. Így zárt, jól körülhatárolható osztályokat kapnak. (Ezek a komplett linkage módszerek.)



Az átlagos lánc módszerek a csoportok minden elemét figyelembe véve, átlagos összekötő láncot definiálnak a klaszterek között. Így az „amőboid" jellegű, viszonylag zárt klasztereket kapnak. (Ezek az módszerek.)



Az irodalomban elég eltérően értékelik a fenti három kapcsolódási típust.

Az ausztráliai iskola képviselői, akik közül a legkiemelkedőbb Lance és Williams, elvetik az egyszerű lánc módszert „tér-összehúzó" hatása miatt. Tér-összehúzó hatás alatt azt értik, hogy ahogyan növekszik egy klaszter, úgy növekszik az általa elfoglalt térrész, és úgy mozdul el a klaszter fokozatosan egy vagy több, eddig még nem klaszteresített egyed felé, ahelyett, hogy távolodna tőlük. Ez azt jelenti, hogy egy eddig nem klaszteresített egyed nagyobb valószínűséggel kerül egy már kialakult, nagy méretű klaszterhez, mint egy hozzá hasonlóbb, de kisebb méretű klaszterhez. Így az egyszerű lánc módszerek elég heterogén klasztereket eredményezhetnek. Ez a hatás vezet az előbbiekben leírt lánc jelenséghez és szerpentin-formához, ahol a lánc két végén lévő egyed lehet teljesen eltérő is.

Az egyszerű lánc módszer helyett a teljes lánc módszert is bírálva, az átlagos lánc módszert javasolják.

A teljes lánc módszerek „tér-tágító" hatására hívják fel a figyelmet. Ez azt jelenti, hogy a klaszterek növekedésük során egyre inkább eltávolodnak a nem klaszterezett egyedektől. Így azok nagyobb valószínűséggel formálnak új klasztermagot, és alakítanak új klasztert, minthogy hozzákapcsolódjanak valamelyik már kialakult klaszterhez.

Érdekes felfedezésük a „tér-tágító" módszerekkel kapcsolatosan, hogy mindig kialakítanak „hulladék" klasztereket, ahová az egymástól és más klasztertől is jelentősen elkülönülő, outlier egyedeket sorolják. Lance és Williams az átlagos lánc módszerek alkalmazását szorgalmazza. Szerintük ezek „tér-konzerváló" módszerek. A klaszterek növekedésük során sem közelebb, sem távolabb nem kerülnek a még nem klaszterezett egyedekhez.

Az angol cambridge-i klaszter-iskola legismertebb képviselői Jardine és Sibson. Ők az ausztrál iskolával ellentétben az egyszerű lánc módszereket részesítik előnyben. Jardine és Sibson a hierarchikus módszerek értékelésére két alapkritériumot fogalmazott meg:

- adatrendszertől a dendrogramig jól definiált legyen a transzformáció, azaz minden adatrendszerhez egyetlen, jól definiált eredményt kapjunk;
- a transzformáció folytonos legyen, vagyis az adatrendszerben bekövetkező kis változás kicsi változást eredményezzen a dendrogramban is.

Jardine és Sibson fenti kritériumainak csak az egyszerű lánc módszerek tesznek eleget.

Jardine és Sibson elismerik az egyszerű lánc módszerek hiányosságait, de szerintük a teljes lánc és átlagos lánc módszerek nem jelentik a problémák megoldását, mert feladják a folytonosságot.

Az angol iskola másik két képviselője Sokal és Sneath bár Jardine-nal és Sibson-nal sok tekintetben nem ért egyet, kiemeli az egyszerű lánc módszereknek azt a tulajdonságát, hogy az egyedek közötti kapcsolatokról átfogó képet adnak. Így gondos elemzéssel, a lánc problémák vizsgálatával értékes információkhoz juthatunk. Ezzel szemben viszont az átlagos lánc módszerek éppen „átlagos" jellegük miatt ezeket elkendőzik. Ezért az átlagos lánc módszerek szerintük sem oldják meg az egyszerű lánc módszerekkel kapcsolatos problémákat.

Sokal és Sneath véleménye a módszerek megválasztásáról: „a klaszterezési módszerek mindegyike érvényes és jó abban az esetben, ha konzisztensen alkalmazzuk."

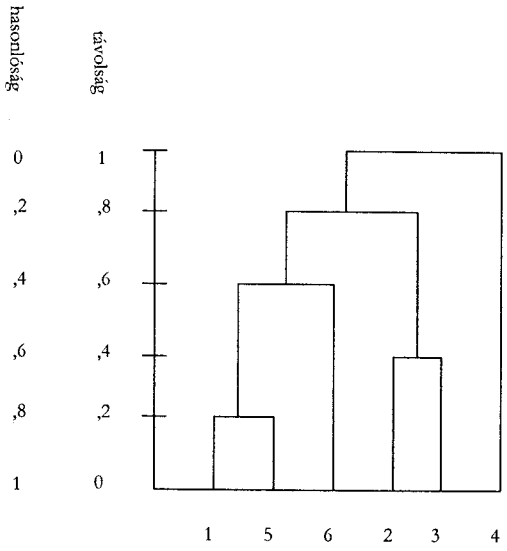
Az optimalizáló technikák alkalmazásának az adatok összes lehetséges particiója mellett kellene megnézni az optimalizálási kritériumot, és választani az optimálist. A gya-

korlatban az összes lehetőség megvizsgálása már közepes nagyságú adatrendszernél sem lehetséges, így ezek a módszerek lokális optimumot adnak. Különböző kezdő partíció-választással megismételve az eljárást, a legjobb kritérium értéknek megfelelő megoldást választhatjuk (természetesen ez még mindig lokális optimum).

Ha az adatrendszer jól strukturált, az eredményekben nem, vagy alig lesz eltérés.

9.7.2. A dendrogramról

A hierarchikus módszerek eredményeit szemléletesen prezentálhatjuk dendrogram, illetve fa-diagram segítésével. A dendrogram az egyedek klaszterbe épülésének és a klaszterek egymásba olvadásának adja jól következő geometriai képét a klaszterek különbözőségi szintjének megfelelően. A következő ábra 6 egyed klaszterezésének eredményét tartalmazza dendrogram formájában.



A függőleges tengely beosztása hasonlósági vagy távolságmértéknek felelhet meg.

A beosztásnál a klaszterek összevonásának szintjeit jelöljük α_j-val ($j = 1,2, \ldots, n$).

Az α_j szinteknek megfelelően jelöljük a klaszterek halmazát C_j-vel. A C_0 esetén $\alpha_0 = 0$, vagyis minden egyed külön klaszterbe tartozik.

Mivel $\alpha_0 \leq \alpha_1 \leq \alpha_2 \leq \ldots \leq \alpha_m$ a C_j klaszter a C_{j-1} klaszterek egyesítésével keletkezett. A C_j a klaszterek összevonásának csomópontjaihoz, az α_j pedig az egyesítés szintjéhez kapcsolható. Johnson írta le a dendrogramot ebben a formában. Két egyed, x_p és x_q között a távolságot a következőképpen definiálta $d(x_p, x_q) = \alpha_i$, ahol i az a legkisebb szám, amelyre igaz, hogy $x_p \in C_i$ és $x_q \in C_i$.

A fenti példa-dendrogramban pl. $d(x_2, x_3) = 0{,}4$, $d(x_1, x_3) = 0{,}8$. Ennek alapján a dendrogrammal ekvivalens távolságmátrixot szerkeszthetünk.

A példában ez a következő:

$$\mathbf{D} = \begin{bmatrix} 0 & 0{,}8 & 0{,}8 & 1 & 0{,}2 & 0{,}6 \\ 0{,}8 & 0{,}0 & 0{,}4 & 1 & 0{,}8 & 0{,}8 \\ 0{,}8 & 0{,}4 & 0 & 1 & 0{,}8 & 0{,}8 \\ 1 & 1 & 1 & 0 & 1 & 1 \\ 0{,}2 & 0{,}8 & 0{,}8 & 1 & 0 & 0{,}6 \\ 0{,}6 & 0{,}8 & 0{,}8 & 1 & 0{,}6 & 0 \end{bmatrix}$$

Johnson megmutatta, hogy a dendrogram kezdőpontjai kielégítik az ultrametrikus egyenlőtlenséget.

Legyen x, y és z tetszőleges három egyed. A köztük lévő távolság $d(x, y) = \alpha_j$, $d(y, z) = \alpha_k$. Ez azt jelenti, hogy x és y a C_j klaszterben, y és z pedig a C_k klaszterben van. Ha $i = \max(j, k)$, akkor C_i ugyanazon klaszterben található. Ez azt jelenti, hogy

$$d(x, y) \leq \alpha_i = \max(\alpha_j, \alpha_k)$$

$$d(y, z) \leq \alpha_i = \max(\alpha_j, \alpha_k)$$

és így

$$d(x, z) = \max[d(x, y),\ d(y, z)]$$

egyenlőtlenség miatt az ultrametrika ezen feltétele erősebb megszorítást jelent a metrika háromszög-egyenlőtlenségénél:

$$d(x, y) \leq d(x, y) + d(y, z)$$

Jardine és Sibson a dendrogramot a $[0, \infty)$ intervallumban ábrázolható függvényként értelmezi az egyedek közötti kapcsolatok leírására, amely incidencia leképezés az egyedek halmazai és az őket összekötő élek között valamilyen érintkezési szintnek megfelelően úgy, hogy

a. minden h' szinten adott klaszter h szerinti klaszterek egyesítése legyen, ahol
$$0 \leq h \leq h'$$
b. kielégítően nagy h esetén minden egyed egy klaszterbe tartozzon
c. ha adott h, akkor létezik olyan $\delta > 0$, hogy a klaszterezés h és $h + \delta$ szinten megegyezik.

Látható, hogy Jardine és Sibson feltételei hasonlóak Johnson definíciójához. A h szint Johnson α értékének felel meg. Jardine és Sibson nem feltételezi viszont, hogy $h = 0$ szinten minden egyed elkülönül.

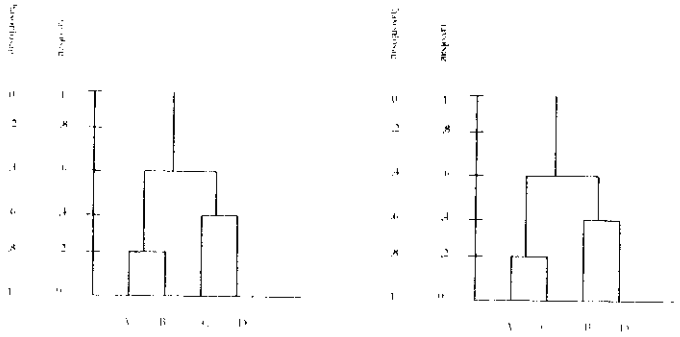
9.7.3. A dendrogramok összehasonlítása

Az egyedeket páronként összehasonlítva az eredeti adatrendszerből hasonlósági vagy távolságmátrixot nyerhetünk. Ezt a mátrixot különböző klaszterező módszerekkel dolgozhatjuk fel, amelyek eredményeként különböző dendrogramokat kaphatunk. A gyakorlati elemzéseknél a különböző módszerek nagy valószínűséggel eredményeznek különböző dendrogramokat, így ezeket a dendrogramokat össze kell hasonlítani.

Definiálni kell e dendrogramok eltérésének mérésére hasonlósági mértéket, amely nek a szokásos tulajdonságokkal kell rendelkeznie:

- legyen szimmetrikus
- ha a két dendrogram azonos, legyen eggyel egyenlő
- ha a két fa teljesen különböző, legyen 0

Például teljesen különböző az alábbi két dendrogram:



A fentiekben láttuk, hogyan lehet a dendrogramot hasonlósági- vagy távolságmátrixszal megfeleltetni.

Hartigan javasolt egy mértéket a hasonlósági mátrixok eltérésének mérésére.

Adott $\mathbf{s}_1$ és $\mathbf{s}_2$ hasonlósági mátrix távolsága:

$$d(\mathbf{s}_1, \mathbf{s}_2) = \sum_{i=1}^{n} \sum_{j=1}^{n} W(i, j)[s_1(i, j) - s_2(i, j)]^2/2,$$

ahol $W(i, j)$ a megfelelő $s(i, j)$ érték súlyfüggvénye.

Hartigan kidolgozott egy olyan hierarchikus klaszterelemző módszert, amely az eredeti egyedek közötti hasonlósági mátrix és a dendrogramból nyert hasonlósági mátrix közötti távolságot minimalizálja. Hartigan a $W(i, j)$ súlyokat a módszer részeként határozza meg.

Phipp, Williams és Clifford javasolta összeszámolni a dendrogram belső pontjainak a számát, amelyek az egyes induló pontpárok (egyedpárok) összekötéséhez szükségesek.

Például a fenti *a* és *b* dendrogramokban ez a következőképpen alakul:

egyed-párok	A, B	A, C	A, D	B, C	B, D	C, D
az a, dendrogram belső csúcsainak a száma	1	3	3	3	6	1
a b, dendrogram belső csúcsainak a száma	3	1	3	3	1	3

Az így nyert két adatsorból korrelációt számíthatunk, ami a két dendrogram hierarchikus kapcsolódásainak hasonlóságát méri.

Williams és Clifford a két adatsor megfelelő elemei különbsége abszolút értékének az összegét osztotta $\binom{n}{2}$-vel, és így disszimilaritási együtthatót kapott: D_d.

További mutatót közöl Dobson (1975). Vegyük a két dendrogramot és tegyük fel, hogy az i különböző szubfából j azonos mindkettőben. A $C_s = j/i$ mutató hasonlósági mérték lehet, amely nulla lesz akkor, amikor nincs közös szubfa, és egy, ha a teljes fa azonos.

9.8. Példa a klaszterelemzésre

(*Iskolai tantárgyak hierarchikus klaszterelemzése*)

Az alábbi mátrix 220 fiú tanuló hat tárgyból kapott eredményei közötti korrelációkat tartalmazza. (Lawley és Maxwell, 1971.)

A tantárgyak közötti kapcsolatok hasonlóságát a hierarchikus klaszterelemzéssel vizsgáljuk meg. A különböző hierarchikus klaszterelemző eljárások dendrogramját könnyebben megérthetjük, ha a korrelációs mátrixot először faktorelemezzük, majd a sokdimenziós skálázó eljárással helyezzük a latens kétdimenziós térben a hat tantárgyat.

Először a faktorelemzés főfaktor-eljárásának, majd a sokdimenziós skálázás legkisebb térelemző (ALSCAL) eljárásának eredményeit közöljük.

Factor Analysis

Correlation Matrix[a]

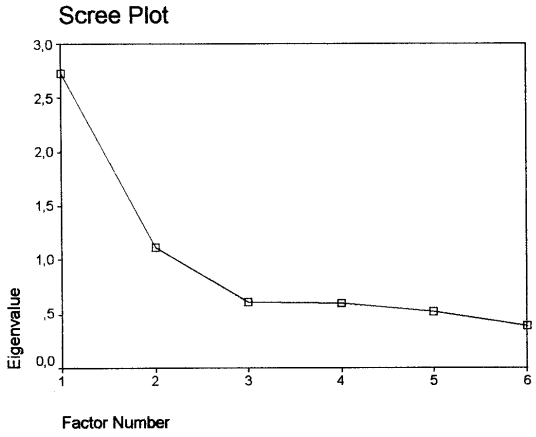
		FRANCIA	ANGOL	TORT	SZAMTAN	ALGEBRA	GEOMET
Correlation	FRANCIA	1,000	,439	,410	,288	,329	,248
	ANGOL	,439	1,000	,351	,354	,320	,329
	TORT	,410	,351	1,000	,164	,190	,181
	SZAMTAN	,288	,354	,164	1,000	,595	,470
	ALGEBRA	,329	,320	,190	,595	1,000	,464
	GEOMET	,248	,329	,181	,470	,464	1,000

a. Determinant = ,237

KMO and Bartlett's Test[a]

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.	,775

a. Neither a vector nor matrix of number of cases (N's) was supplied. Bartlett's Test of Sphericity cannot be computed.



Factor Matrix[a]

	Factor	
	1	2
FRANCIA	,586	,376
ANGOL	,594	,236
TORT	,432	,415
SZAMTAN	,712	-,336
ALGEBRA	,701	-,276
GEOMET	,584	-,183

Extraction Method: Principal Axis Factoring.

a. 2 factors extracted. 16 iterations required.

Communalities

	Extraction
FRANCIA	,485
ANGOL	,408
TORT	,358
SZAMTAN	,620
ALGEBRA	,567
GEOMET	,375

Extraction Method: Principal Axis Factoring.

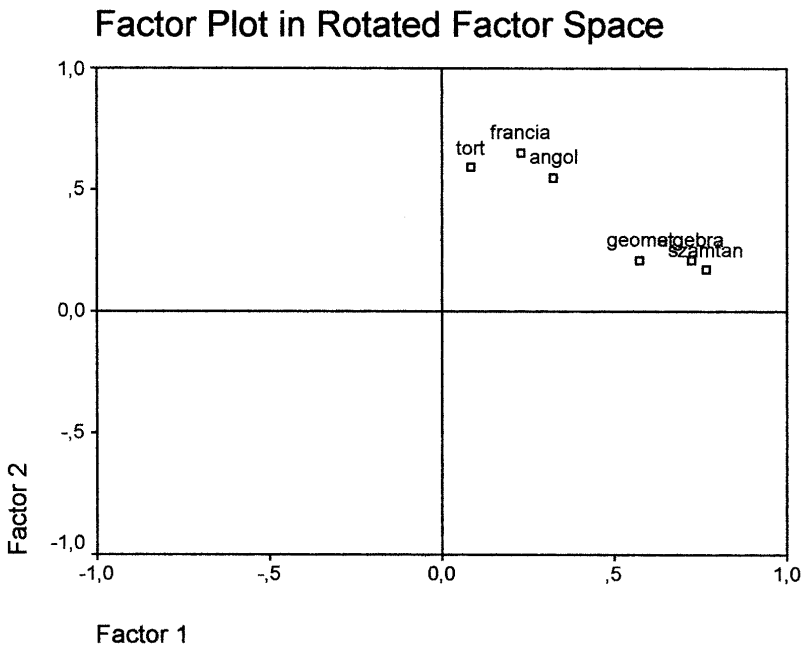
Total Variance Explained

	Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Factor	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,222	37,036	37,036	1,606	26,765	26,765
2	,592	9,864	46,900	1,208	20,136	46,900

Extraction Method: Principal Axis Factoring.

Rotated Factor Matrix[a]

a. Rotation converged in 3 iterations.



Sokdimenziós skálázás (Alscal)

Input mátrix:

Raw (unscaled) Data for Subject 1

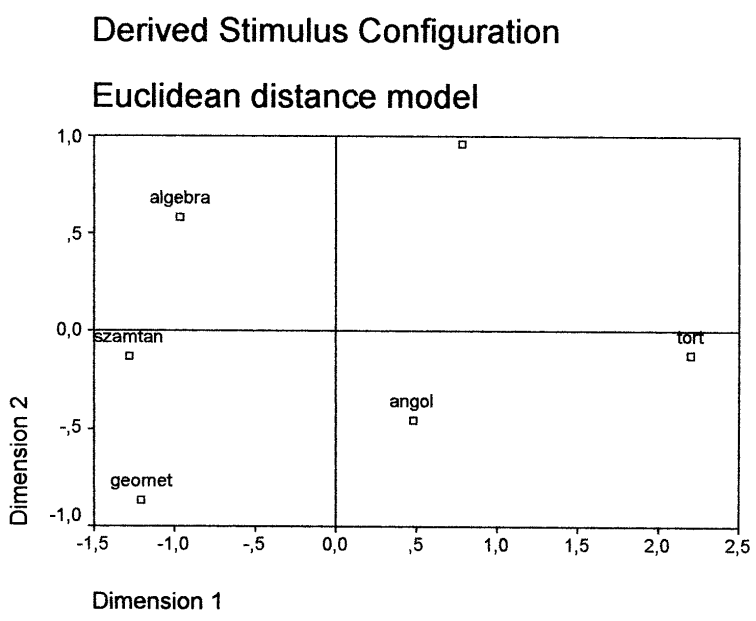
	1	2	3	4	5	6
1	,000					
2	,561	,000				
3	,590	,649	,000			
4	,712	,646	,836	,000		
5	,671	,680	,810	,405	,000	
6	,752	,671	,819	,530	,536	,000

For matrix
Stress = ,00655 RSQ = ,99971

Configuration derived in 2 dimensions

Stimulus Coordinates

Stimulus Number	Stimulus Name	Dimension 1	2
1	FRANCIA	,7858	,9693
2	ANGOL	,4770	-,4530
3	TORT	2,1992	-,1159
4	SZAMTAN	-1,2810	-,1235
5	ALGEBRA	-,9685	,5889
6	GEOMET	-1,2124	-,8657



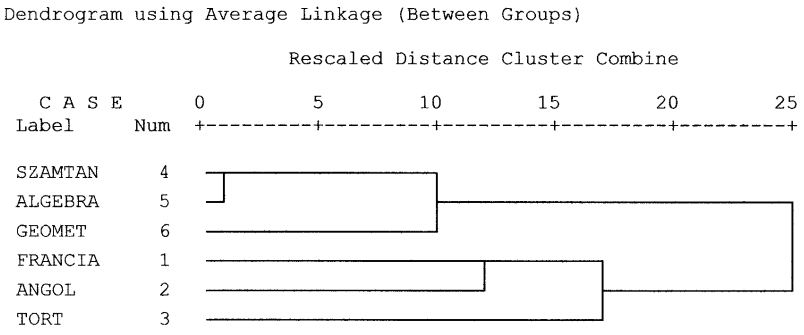
Hierarchikus agglomeratív klaszterelemzés

Proximity Matrix

Case	Matrix File Input					
	FRANCIA	ANGOL	TORT	SZAMTAN	ALGEBRA	GEOMET
FRANCIA		,561	,590	,712	,671	,752
ANGOL	,561		,649	,646	,680	,671
TORT	,590	,649		,836	,810	,819
SZAMTAN	,712	,646	,836		,405	,530
ALGEBRA	,671	,680	,810	,405		,536
GEOMET	,752	,671	,819	,530	,536	

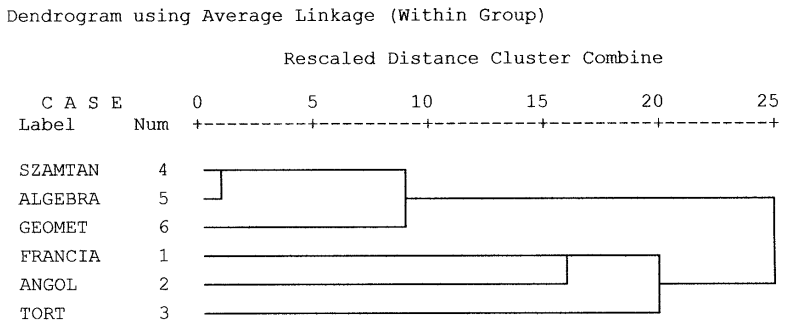
Average Linkage (Between Groups)

Dendrogram



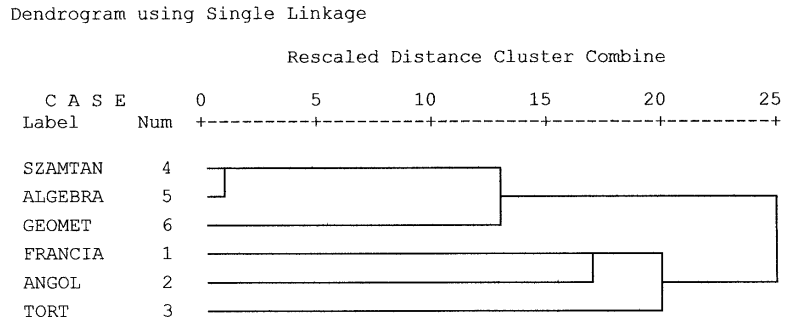
Average Linkage (Within Groups)

Dendrogram



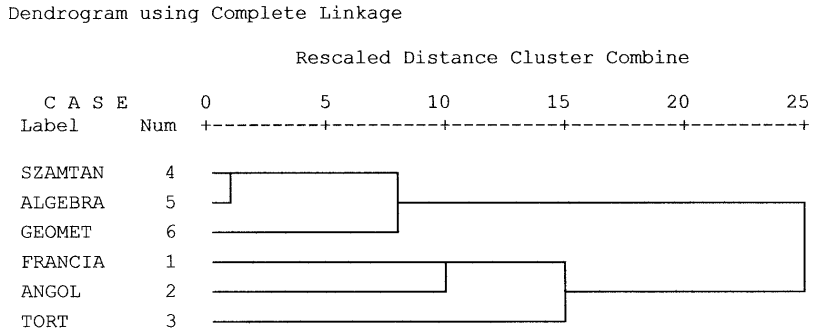
Single Linkage

Dendrogram



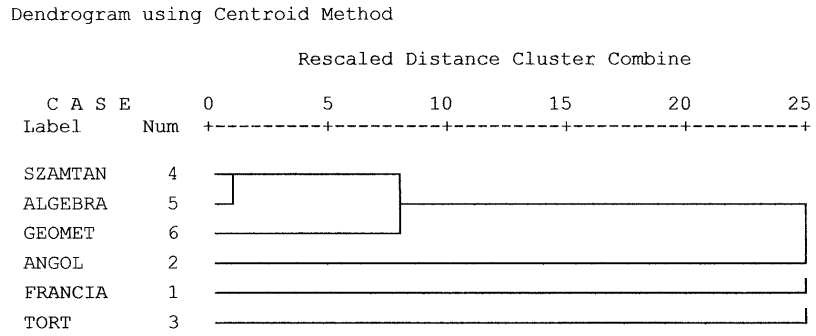
Complete Linkage

Dendrogram



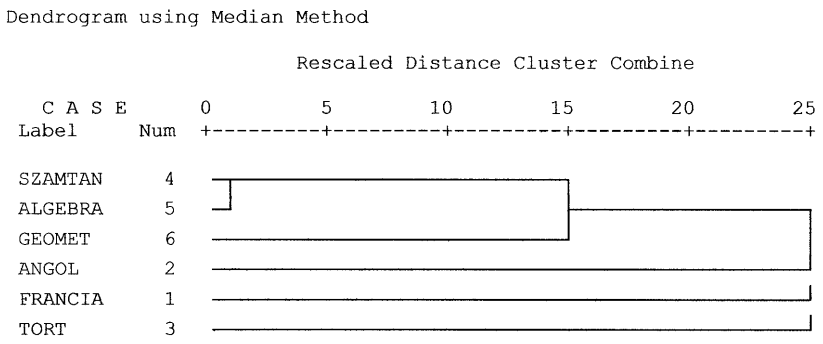
Centroid Linkage

Dendrogram



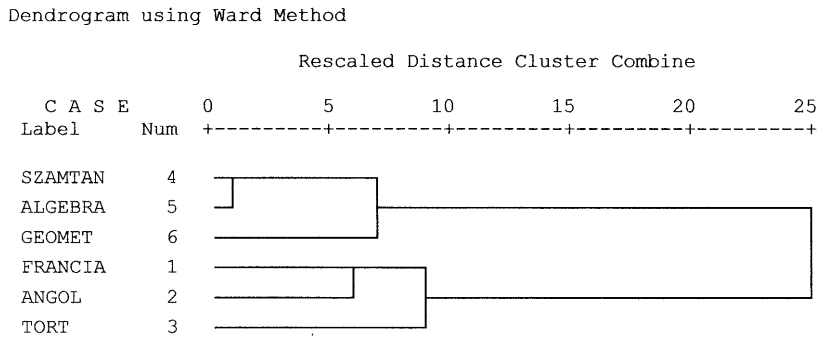
Median Linkage

Dendrogram



Ward Linkage

Dendrogram



III. rész

Latens változós modellek

10. fejezet

Általános latens változós modell

Két változóhalmazt különböztetünk meg alapvetően egymástól. Az egyik a közvetlenül megfigyelhető, mérhető, manifeszt változók halmaza: $\mathbf{x}' = [x_1, x_2, \ldots, x_m]$, $\mathbf{x} \in X$, a másik a közvetlenül nem mérhető, latens változók, faktorok halmaza: $\boldsymbol{\xi}' = [\xi_1, \xi_2, \ldots, \xi_r]$, $\boldsymbol{\xi} \in \Xi$.

A latens változók száma, r általában lényegesen kisebb, mint a megfigyelt változók száma, m $(r \ll m)$, mivel a latens változós modellek általában adatredukciós módszerek is, melyek a manifeszt változók nagy száma helyett kevés számú latens változóval próbálják a megfigyelt adatokat a lehető legpontosabban reprodukálni.

A megfigyelt változók mérési skáláját, mérési szintjét szokásosan négy kategóriába sorolják, úgymint nominális, ordinális, intervallum- és arányskála. Ettől eltérően, a következőkben a feltételezésünkhöz jobban illeszkedő kétosztályos klasszifikációt alkalmazzuk: eszerint a változók vagy metrikusak, vagy kategorikusak. A metrikus változó szinonimájaként használjuk a mennyiségi, kvantitatív változó kifejezéseket. A kategorikus változó szinonimájaként használjuk a minőségi, kvalitatív változó kifejezéseket.

A metrikus változók értékei valós számok (vagy a valós számoknak egy részhalmaza), valódi értékek, amelyek lehetnek folytonosak vagy diszkrétek.

A kategorikus változók a megfigyeléseket egymást kizáró, diszjunkt csoportokba, kategóriákba sorolják. A kategóriák lehetnek rendezettek is.

A latens változós módszereket a változók mérési skálája és közvetlen megfigyelhetősége szerint a következőképpen klasszifikálhatjuk:

		Manifeszt változók	
		Metrikus	Kategorikus
Latens változók	Metrikus	Faktorelemzés LISREL LVPLS	Latens tulajdonságelemzés Faktorelemzés kategorikus adatokkal
	Kategorikus	Latens profilelemzés	Latens osztályelemzés

Mind a megfigyelt, mind a latens változók valószínűségi változók, így a kapcsolataikat valószínűségeloszlással jellemezhetjük.

A latens változós modelleknél feltételezzük, hogy létezik a megfigyelt változóknak a latens változókra vonatkozó feltételes együttes sűrűségfüggvénye:

$$g(\mathbf{x} \mid \boldsymbol{\xi}).$$

Ha a manifeszt változók diszkrétek, akkor a g a feltételes valószínűségek halmaza. Ha a latens változók sűrűségfüggvénye $h(\boldsymbol{\xi})$, a megfigyelt változók (feltétel nélküli) együttes sűrűségfüggvényét a következőképpen írhatjuk:

$$f(\mathbf{x}) = \int h(\boldsymbol{\xi}) g(\mathbf{x} \mid \boldsymbol{\xi}) d\boldsymbol{\xi}.$$

A latens változós modellekben az érdeklődés a latens változók manifeszt változókra vonatkozó feltételes eloszlására irányul, vagyis arra, hogy mit tudunk a latens változókról a megfigyelt változók ismeretében:

$$h(\boldsymbol{\xi} \mid \mathbf{x}) = h(\boldsymbol{\xi}) g(\mathbf{x} \mid \boldsymbol{\xi}) / f(\mathbf{x}).$$

Ahhoz, hogy a $h(\xi \mid \mathbf{x})$ feltételes sűrűségfüggvényt megkapjuk, ismernünk kell a h, g és f sűrűségfüggvényeket, de csak az f sűrűségfüggvényt tudjuk becsülni, h és g nem határozható meg egyértelműen, ezért pótlólagos feltételezést kell tennünk. Ez a pótlólagos feltételezés a latens változós modelleknél a *megfigyelt változók latens változókra vonatkozó feltételes függetlensége*:

$$g(\mathbf{x} \mid \boldsymbol{\xi}) = \prod_{i}^{m} g_i(x_i \mid \boldsymbol{\xi}).$$

A manifeszt változók feltételes függetlensége azt fejezi ki, hogy a megfigyelt változók közötti kapcsolatok a manifeszt változóknak a latens változóktól való függőségéből származnak, vagyis a latens változók idézik elő a megfigyelt változók közötti kapcsolatokat. Így, ha a latens változók hatásait kontrolláljuk, akkor a megfigyelt változók függetlenekké válnak. A feltételes függetlenség-definícióját felhasználva a megfigyelt változók együttes sűrűségfüggvényét a következőképpen írhatjuk:

$$f(\mathbf{x}) = \int h(\boldsymbol{\xi}) \prod_{i}^{m} g_i(x_i \mid \boldsymbol{\xi}) d\boldsymbol{\xi}.$$

$f(x)$ függ a h és g_i függvények mellett a latens változók számától, r-től is. A gyakorlatban szeretnénk a lehető legkisebb r érték mellett megkeresni az egyenletet adekvátan reprezentáló h és g_i függvényeket.

Az előzőek bemutatására tekintsünk egy egyszerű számpéldát (forrás D. J. Bartholomew [1987]). Legyenek A és B dichotom változók, és legyen a 2×2-es gyakoriságtáblázat a következő:

	A	$\overline{A}$	
B	350	200	550
$\overline{B}$	150	300	450
	500	500	1000

Tételezzük fel, hogy az A és B változók között megfigyelt asszociáció egy harmadik dichotom változó (C) kontrollálásával megszűnik, vagyis a C változó eredményezi az A és B közötti kapcsolatot. Osszuk a C változó két kategóriája szerinti két almintára a megfigyeléseket, és nézzük meg az almintákban A és B együttes gyakoriságeloszlását:

C	A	$\overline{A}$	
B	320	80	400
$\overline{B}$	80	20	100
	400	100	500

$\overline{C}$	A	$\overline{A}$	
B	30	120	150
$\overline{B}$	70	280	350
	100	400	500

A két résztáblázatban az A és B változók közötti asszociáció megszűnt, így a teljes mintában megfigyelt kapcsolódás hamis asszociációnak tekinthető, amit a C változó idézett elő.

A következőkben röviden bemutatjuk a két legismertebb latens változós modellt.

10.1. Bináris manifeszt változók és egy bináris latens változó

Tegyük fel, hogy $x_1, x_2, \ldots, x_m$ a bináris megfigyelt változókat jelöli, ahol $x_i = 0$ vagy $x_i = 1$ minden i-re. Tételezzük fel, hogy a megfigyelt változók asszociációit egy ξ bináris változó magyarázza. A bináris latens ξ valószínűségi változó karakterisztikus eloszlása:

$$h(1) = P(\xi = 1) = \eta \quad \text{és} \quad h(0) = P(\xi = 0) = 1 - \eta.$$

A megfigyelt változók feltételes eloszlása:

$$g_i(x_i \mid \xi) = P(x_i \mid \xi) = \pi_{i\xi}^{x_i}(1 - \pi_{i\xi})^{1-x_i}, \quad (x_i, \xi \in \{0, 1\}).$$

Miután ebben az egyszerű esetben a h és g_i eloszlások adottak, csak a η, $\{\pi_{i0}\}$, és a $\{\pi_{i1}\}$ paramétereket kell becsülni.

Az $f(\mathbf{x})$ függvény:

$$f(\mathbf{x}) = \eta \prod_{i=1}^{m} \pi_{i1}^{x_i}(1 - \pi_{i1})^{1-x_i} + (1 - \eta) \prod_{i=1}^{m} \pi_{i0}^{x_i}(1 - \pi_{i0})^{1-x_i}.$$

A modell paramétereit becsülhetjük a maximum likelihood módszerrel (lásd részletesen később a megfelelő fejezetben), és a modell illeszkedését a megfigyelt együttes gyakoriságokhoz statisztikailag tesztelhetjük. Ha az illeszkedés nem elég jó, akkor próbálkozhatunk három vagy több latens osztállyal vagy folytonos latens változóval.

Elfogadhatóan illeszkedő modell esetén a megfigyelési egységeket a manifeszt változók mérési adatai alapján sorolhatjuk be az egyes latens osztályokba az *a posteriori* valószínűségek alapján. Az *a posteriori* valószínűségeket a feltételes sűrűségfüggvény

$h(\boldsymbol{\xi} \mid \mathbf{x})$ segítségével határozzuk meg. A konkrét esetben:

$$h(1 \mid \mathbf{x}) = P(\xi = 1 \mid \mathbf{x}) = \eta \prod_{i=1}^{m} \pi_{i1}^{x_i}(1-\pi_{i1})^{1-x_i} \Big/$$

$$\Big/ \left\{ \eta \prod_{i=1}^{m} \pi_{i1}^{x_i}(1-\pi_{i1})^{1-x_i} + (1-\eta) \prod_{i=1}^{m} \pi_{i0}^{x_i}(1-\pi_{i0})^{1-x_i} \right\} =$$

$$= 1 \Big/ \left[1 + \left(\frac{1-\eta}{\eta} \right) \exp \sum_{i=1}^{m} \left\{ x_i \log \frac{\pi_{i0}}{\pi_{i1}} + (1-x_i) \log \frac{1-\pi_{i0}}{1-\pi_{i1}} \right\} \right].$$

A besorolási szabály szerint az $\mathbf{x}$ mérési vektorral rendelkező megfigyelési egységet az 1 latens osztályba soroljuk, ha

$$h(1 \mid \mathbf{x}) > h(0 \mid \mathbf{x}),$$

vagy ha

$$X = \sum_{i=1}^{m} x_i \{\text{logit } \pi_{i0} - \text{logit } \pi_{i1}\} >$$

$$> \sum_{i=1}^{m} \log\{(1-\pi_{i1})/(1-\pi_{i0})\} - \text{logit } \eta,$$

ahol logit $u = \log\{u/(1-u)\}$.

A fentiekből látható, hogy a besorolási szabály az x_i-k lineáris függvénye.

10.2. Normális eloszlású változók

A másik általános eset, amikor mind a manifeszt, mind a latens változókról feltételezzük, hogy folytonosak és normális eloszlású valószínűségi változók. Tételezzük fel, hogy a megfigyelt változók ($\mathbf{x}$) együttes eloszlása többváltozós normális eloszlás $\boldsymbol{\mu}$ várható értékkel és $\boldsymbol{\Sigma}$ variancia-kovarianciamátrixszal:

$$\mathbf{x} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Tegyük fel, hogy a latens változók standard normális eloszlásúak:

$$\boldsymbol{\xi} \sim N_r(\mathbf{0}, \mathbf{I})$$

és

$$\mathbf{x} \mid \boldsymbol{\xi} \sim N_m(\boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\xi}, \boldsymbol{\Theta}),$$

ahol $\boldsymbol{\Lambda}$ a faktorsúlyok ($m \times r$) típusú mátrixa, $\boldsymbol{\Theta}$ a hibakomponensek varianciáinak diagonális mátrixa.

A fentiekből következik, hogy

$$\mathbf{x} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}).$$

Amennyiben $m = r$, $\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}$ pontosan egyenlő lesz $\boldsymbol{\Sigma}$-val, azonban $r < m$ esetén ez nem feltétlenül van így.

$\boldsymbol{\xi}$ *a posteriori* eloszlását a következőképpen határozhatjuk meg:

$$\boldsymbol{\xi} \mid \mathbf{x} \sim N_r\{\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}), (\boldsymbol{\Lambda}'\boldsymbol{\Theta}^{-1}\boldsymbol{\Lambda} + \mathbf{I})^{-1}\}.$$

10.3. A latens változó és a mérés

Minden empirikus társadalomtudományi kutatás a vizsgálat által körülhatárolt jelenségek megfigyelésén alapul. Az operacionalizálás transzformálja a megfigyeléseket változókká, és a változók kompozíciója definiálja a mérési skálát (itt ne a mérési skála – nominális, ordinális, intervallum, arány – típusaira gondoljunk), amely összefüggésbe hozható az elméleti koncepciókkal. A mérési skála minőségének kiértékelése meg kell hogy előzze a hipotézisről alkotott végső konklúziónkat, és a mérés minősége korlátozza is a konklúzió súlyát.

A mérés minőségének két legfontosabb eleme a mérés érvényessége (validity) és a mérés megbízhatósága (reliability). A két komponens nem független egymástól, mivel ha egyszer a mérés nem megbízható, nem lehet érvényes sem.

A megbízhatóság mérése a pszichometriai kutatásokban központi kérdés, ezért nem meglepő, hogy számtalan mutatót dolgoztak ki: split-half, Spearman–Brown-féle, Cronbach-féle, Kuder–Richardson-féle, Armor-féle, Heise- és Bohrnstedt-féle, Tucker–Lewis-féle stb. Keveset tudunk ezeknek a megbízhatósági mutatóknak a tulajdonságairól, alkalmazásaikról, hiszen leginkább a számítógépes program elérhetősége dönti el, hogy melyiket alkalmazzák.

A társadalomtudományokban a legritkább esetben lehetséges az elméleti fogalmakat közvetlenül mérni, így a mérhető indikátorokból kell a mérési skálát összeállítani, mint a manifeszt változók valamilyen függvényét. Keveset tudunk az így képzett mérési skála megbízhatóságáról. Be lehet látni, hogy általában a képzett változó megbízhatósága jobb, mint az itemek megbízhatósága, de számtalan kérdés merül fel ezzel kapcsolatban.

Tételezzük fel, hogy a megfigyelt változót (x) kifejezhetjük a valódi érték (t) és a mérési hiba (e) függvényeként: $x = f(t, e)$. A mérési skála (u) függvénye a megfigyelt változóknak: $u = g(x)$. Az érdekel bennünket, hogy a valódi érték és a mérési hiba hogyan befolyásolja a mérési skálát (u) a megfigyelt változókon (x) keresztül.

A mérés és a skálázás szkémáját a következő ábrával szemléltetjük:

A két átfedő rész mérési modellnek és mérési skálának nevezhető. Fontos különbséget tenni a kettő között, mivel különböző a mögöttes struktúrájuk. A mérési modellben feltételezzük, hogy a manifeszt változó két latens hatás eredője. Feltételezzük, hogy a latens valódi értékeknek létezik valamilyen struktúrája, és hogy befolyásolják a megfigyelt változókat. Ezt a kapcsolatot az $x = f(t, e)$ függvénykapcsolattal fejezhetjük ki. A mérési skála, a jobb oldali keret, a megfigyelt változók és a mérési skála függvénykapcsolatát fejezi ki: $u = g(x)$. A g függvényt a skála vagy az operacionalizálás definiálásával adhatjuk meg. Ha létezik a mérés különböző kritériuma (y), akkor a validitást a kritérium (y) és a mérési skála (u) közötti korrelációval mérhetjük.

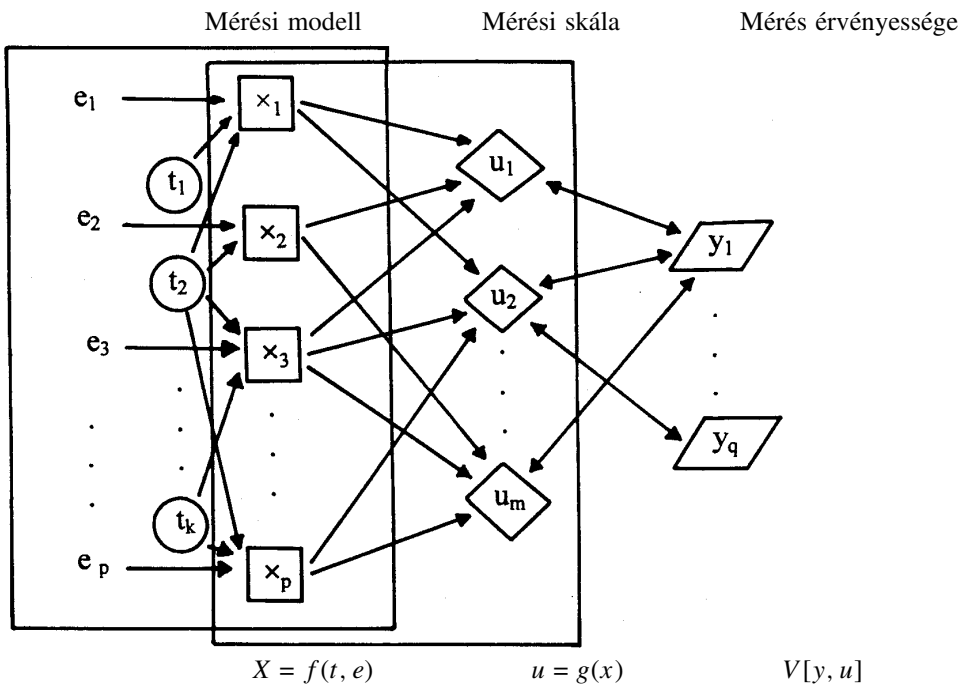
A congeneric (azonos eredetű) mérésnél azt tételezzük fel, hogy az egyes megfigyelt változók ugyanazt a latens tulajdonságot mérik, csupán a hibatagban különböznek, és bármelyik két változó valódi értéke (latens tulajdonságértéke, szisztematikus komponense) lineáris kapcsolatban van egymással. Jelölje x_i az i-edik változót. A klasszikus mérési modell a következő:

$$x_i = t_i + e_i \qquad i = 1, \ldots, m,$$

ahol t_i a valódi érték, szisztematikus latens komponens, e_i a hibaérték, hibakomponens.

A klasszikus mérési modellnél feltételezzük, hogy a valódi és a hibakomponens korrelálatlan

$$\operatorname{cov}(t_i, e_i) = 0,$$



és feltételezzük, hogy a különböző változók hibakomponensei is korrelálatlanok:

$$\operatorname{cov}(e_i, e_j) = 0.$$

Általában feltételezzük továbbá, hogy

$$E(e_i) = 0,$$

és minden valódi érték lineárisan függ egy latens valószínűségi változótól, τ-tól:

$$t_i = \mu_i + \beta_i \tau,$$

ahol feltételezzük, hogy $E(\tau) = 0$ és $\operatorname{var}(\tau) = 1$.

A congeneric mérési modell:

$$x_i = \mu_i + \beta_i \tau + e_i, \tag{10.1}$$

ahol $E(x_i) = \mu_i$.

Az i-edik változó megbízhatósági együtthatója (reliability):

$$\rho_i = \frac{\beta_i^2}{\beta_i^2 + \theta_i^2},$$

ahol $\operatorname{var}(t_i) = \beta_i^2$, $\operatorname{var}(e_i) = \theta_i^2$.

Ha m változóra írjuk fel a congeneric mérési modellt, akkor a (10.1) egyenlet a következőképpen írható:

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\beta}\,\tau + \mathbf{e}, \tag{10.2}$$

ahol $\mathbf{x}$, μ, $\mathbf{e}$ m elemű oszlopvektorok.

A változók variancia-kovarianciamátrixa:

$$\boldsymbol{\Sigma} = \boldsymbol{\beta}\,\boldsymbol{\beta}' + \boldsymbol{\Theta}, \tag{10.3}$$

ahol Θ a hibavarianciák diagonális mátrixa.

A (10.2) és (10.3) egyenletek a faktorelemzés alapegyenletei egy közös faktor esetére. A faktorelemzés eljárását alkalmazhatjuk a β és Θ paraméterek becslésére. A congeneric mérés speciális esetei a párhuzamos és a tau-ekvivalens mérések.

A párhuzamos mérésnél feltételezzük, hogy

$$\beta_1 = \beta_2 = \ldots = \beta_m$$

és

$$\theta_1 = \theta_2 = \ldots = \theta_m,$$

vagyis a párhuzamos mérésnél azonosak a szisztematikus komponensek varianciái és a hibakomponens varianciái is.

A tau-ekvivalens mérésnél csak a szisztematikus komponensek varianciáinak az egyenlőségét tételezzük fel, a hibatagok varianciái különbözhetnek:

$$\beta_1 = \beta_1 = \ldots = \beta_m.$$

A tau-ekvivalens mérésnél a változók kovarianciái megegyeznek, de a varianciák különbözhetnek.

Általánosan, ha s különböző congeneric tulajdonságú változóhalmazunk van, és $\mathbf{x}_g$ jelöli a g-edik csoport vektorváltozóját ($\mathbf{x}_g$ m_g elemű vektor), akkor $\mathbf{x}_g$ mérési modellje:

$$\mathbf{x}_g = \boldsymbol{\mu}_g + \boldsymbol{\beta}_g \boldsymbol{\tau}_g + \mathbf{e}_g.$$

Ha a latens szisztematikus komponensek korrelálnak egymással, akkor az s változóhalmazt együtt elemezve a következő modellt írhatjuk fel:

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\beta}\,\boldsymbol{\tau} + \mathbf{e}, \tag{10.4}$$

ahol $\mathbf{x}$ $p = (m_1 + \ldots + m_s)$ elemű vektor,

β $(p \times s)$ típusú kvázi-diagonális mátrix, ahol a diagonális „elemek” a β_g-k,

τ s elemű vektor, elemei a τ_s latens komponensek,

μ a változók várható értékeit tartalmazza.

Legyen Γ a τ variancia-kovarianciamátrixa. A megfigyelt változók variancia-kovarianciamátrixa:

$$\boldsymbol{\Sigma} = \boldsymbol{\beta}\,\boldsymbol{\Gamma}\,\boldsymbol{\beta}' + \boldsymbol{\Theta}_g, \tag{10.5}$$

ahol Θ_g a hibavarianciák p-edrendű mátrixa.

A (10.4) és (10.5) egyenletek formálisan megegyeznek a faktorelemzés modelljével, feltételezve q korrelált faktort, amelyek a diszjunkt változóhalmazokhoz tartoznak. A modellt a faktorelemzés becslési eljárásaival illesztjük az adatokhoz.

A szisztematikus latens komponensek feltételezésünk szerint korrelálnak egymással, így kifejezhetjük őket egy közös szisztematikus faktormodellel:

$$\boldsymbol{\tau} = \boldsymbol{\Lambda}\,\boldsymbol{\xi} + \mathbf{u},$$

ahol ξ a közös faktorok r-elemű vektora, $\mathbf{u}$ az egyedi faktorok vektora (q-elemű).

Legyen Φ a ξ variancia-kovarianciamátrixa, és Ψ az egyedi faktorok diagonális varianciamátrixa. Ekkor a τ szisztematikus latens komponensek variancia-kovarianciamátrixát a következőképpen írhatjuk:

$$\boldsymbol{\Gamma} = \boldsymbol{\Lambda}\,\boldsymbol{\Phi}\,\Lambda' + \boldsymbol{\Psi}. \tag{10.6}$$

A (10.6) egyenletet behelyettesítve a (10.5) egyenletbe a következőt kapjuk

$$\boldsymbol{\Sigma} = \boldsymbol{\beta}(\boldsymbol{\Lambda}\,\boldsymbol{\Phi}\,\Lambda' + \boldsymbol{\Psi})\boldsymbol{\beta}' + \boldsymbol{\Theta}_g. \tag{10.7}$$

Ez a másodrendű faktorelemzés alapegyenlete, ahol τ az elsőrendű faktorokat, ξ a másodrendű faktorokat tartalmazza.

11. fejezet

Latens struktúra-modell

A szociológiai kutatások – elsősorban a survey-adatokon végzett kutatások – nagyobbrészt kvalitatív (nominális vagy ordinális) változók közötti kapcsolatok elemzésére irányulnak. Gyakori az a feltételezés, hogy a változók közötti megfigyelt kapcsolatok magyarázhatók más kvalitatív változóval, az ún. tesztfaktorral, vagyis a megfigyelt változók közötti kapcsolat megszűnik, ha kontrolláljuk őket a tesztfaktorral. Ez azt jelenti, hogy a vizsgált mainfeszt változók feltételesen függetlenek egymástól, vagyis a tesztfaktor minden kategóriájához rendelt feltételes táblában a változók kölcsönösen függetlenek. Ha a tesztfaktor is megfigyelt és kvalitatív, akkor a hagyományos loglineáris, logit, probit modelleket alkalmazhatjuk az elemzésnél. Ha a tesztfaktor nem megfigyelt, latens kvalitatív változó, akkor a latens struktúraelemzés módszereit kell alkalmaznunk.

A latens struktúraelemzés hasonlít a faktorelemzéshez, azonban míg a faktorelemzés folytonos megfigyelt változókat magyaráz folytonos latens változókkal, a latens struktúraelemzésnél mind a manifeszt, mind a latens változók között lehetnek kvalitatív változók is. Ezenkívül a latens struktúraelemzésnél nem kell feltételeznünk sem a normális eloszlást, sem a mérési skála folytonosságát.

A latens struktúraelemzés kitüntetett esete, amikor a megfigyelt változók diszkrét, kategorikus változók (sokszor dichotómok), és egy vagy több kategorikus latens változónk van (T lehetséges kategóriával). Ezt az esetet hívjuk *latens osztályelemzés*nek. A modell alapfeltételezése, hogy a latens változók bármelyik kategóriájában a megfigyelt változók függetlenek egymástól, vagyis a manifeszt változók megfigyelt kapcsolatait az adatoknak a latens változó kategóriája szerinti klasszifikációja eredményezi.

A latens osztályelemzés célja, hogy jellemezze azt a kategorikus latens változót (változókat), amely magyarázza a megfigyelt kategorikus változók közötti asszociációkat:

- becsülje a latens változó relatív gyakoriság-eloszlását,
- becsülje a megfigyelt változók relatív gyakoriságait a latens változó kategóriáiban,
- a megfigyelt gyakoriság-eloszlások alapján következtessen a latens változó szubsztantív lényegére.

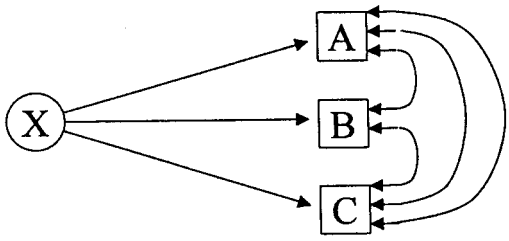
A latens struktúraelemzés speciális módszere a *latens tulajdonságelemzés* (latent trait analysis), amelyben a diszkrét megfigyelt változók asszociációit folytonos latens változókkal magyarázzuk. A harmadik módszer a *latens profilelemzés*, ahol a diszkrét latens változókkal magyarázzuk a folytonos megfigyelt változók kapcsolatait.

A latens struktúraelemzésről az első publikációk az 50-es években jelentek meg (lásd Green [1951], Anderson [1954] és Gibson [1959] munkáit). Az első alapos kifejtését Lazarsfeld és Henry (1968) végezte el. Az, hogy mégsem alkalmazták széles körben, elsősorban a számítási és statisztikai nehézségeknek köszönhető. A modell becslésére alkalmazott eljárások – (elsősorban az ún. determináns módszer (lásd Madansky 1960, Lazarsfeld és Henry, 1968, Lasy számítógépes program) – okozták a nehézséget, mivel gyakran adtak értelmezhetetlen eredményeket (pl. a valószínűségekre [0, 1] közé nem eső becslést, vagy egyszerűen azonos adathalmazra különböző eredményt). A hatékony, a maximum likelihood elv szerinti becslési eljárást Goodman (1974) és Haberman (1974) dolgozták ki. A paraméterek becslésére kidolgozott „iterative proportional scaling” eljárást alkalmazta Clogg (1977) is az MLLSA (Maximum Likelihood Latent Structure

Analysis) programban. Ez a program – helyes alkalmazás esetén – már elkerülte az előzőekben említett számítási problémákat. Az LCAG (Latent Class Models and Other Loglinear Models with Latent Variables) számítógépes program (Jacques Hagenaars munkája) is a Goodman (1974) által javasolt EM algoritmus alapján dolgozik, és a paraméterekre maximum likelihood becslést ad.

11.1. Latens osztály-modell

Tekintsünk egy háromdimenziós kereszttáblát, három megfigyelt változó, A, B, C együttes gyakoriság-eloszlását. Jelentse π_{ijk} annak elméleti valószínűségét, hogy egy megfigyelés az (i, j, k) cellába esik ($i = 1, \ldots, I; j = 1, \ldots, J; k = 1, \ldots, K$). Tegyük fel, hogy a manifeszt változók kapcsolatát egyetlen latens változó (X) magyarázza, amelynek T kategóriája van és jelölje π_{ijkt}^{ABCX} a várható valószínűségét annak, hogy egy megfigyelés az (i, j, k, t) cellába esik. Amíg a π_{ijk} közvetlenül (direkt) megfigyelhető, addig a π_{ijkt}^{ABCX} közvetetten (indirekt) figyelhető csak meg az ABC táblázatból) Ezt az egyszerű esetet mutatja a következő ábra.



11.1. ábra. Három megfigyelt és egy latens változó kapcsolata

Ennek alapján a következő kifejezések írhatók fel:

$$\pi_{ijk} = \sum_{t}^{T} \pi_{ijkt}^{ABCX}, \tag{11.1}$$

és

$$\pi_{ijk} = \sum_{t}^{T} \pi_{ijkt}^{ABCX} = \sum_{t}^{T} \pi_{t}^{X} \pi_{ijk}^{\overline{ABC}X}, \tag{11.2}$$

ahol a

$\pi_{ijk}^{\overline{ABC}X}$ feltételes valószínűsége annak, hogy egy megfigyelés az (i, j, k) cellába esik, feltéve, hogy a megfigyelés az X latens változó t-edik cellájában van,

π_{t}^{X} jelöli annak a valószínűségét, hogy a megfigyelés az X latens változó t-edik osztályába tartozik.

A (11.1) egyenlet azt a pótlólagos feltételezést is tartalmazza, hogy a megfigyelések egyetlen latens változó osztályaiba teljes egészében szétoszthatók (ez az egyedek latens osztályba sorolhatósága).

Ha feltételezzük, hogy az X latens változó minden kategóriájában (osztályában) az A, B, C manifeszt változók kölcsönösen függetlenek egymástól, akkor a $\pi_{ijkt}^{\overline{ABC}X}$ a következőképpen írható:

$$\pi_{ijkt}^{\overline{ABC}X} = \pi_{it}^{\overline{A}X}\pi_{jt}^{\overline{B}X}\pi_{kt}^{\overline{C}X}, \tag{11.3}$$

ahol $\pi_{it}^{\overline{A}X}$ annak a feltételes valószínűsége, hogy egy megfigyelés az A változó i-edik kategóriájába esik, feltéve hogy a megfigyelés az X latens változó t-edik osztályába tartozik.

A $\pi_{jt}^{\overline{B}X}$, $\pi_{kt}^{\overline{C}X}$ hasonlóan értelmezhető, mint a $\pi_{it}^{\overline{A}X}$.

A π_{ijkt}^{ABCX} a fentiek alapján:

$$\pi_{ijkt}^{ABCX} = \pi_t^X\pi_{it}^{\overline{A}X}\pi_{jt}^{\overline{B}X}\pi_{kt}^{\overline{C}X}. \tag{11.4}$$

Ennek alapján a (11.2) egyenletet a következőképpen írhatjuk:

$$\pi_{ijk} = \sum_t^T \pi_t^X\pi_{it}^{\overline{A}X}\pi_{jt}^{\overline{B}X}\pi_{kt}^{\overline{C}X}. \tag{11.5}$$

Ezt az egyenletet a latens osztályelemzés alapegyenletének nevezzük. Ez fejezi ki azt, hogy A, B, C megfigyelt változók *feltételesen függetlenek* az adott latens változó kategóriáiban, vagyis az X latens változó magyarázza az A, B, C változók között megfigyelt asszociációt. Eszerint az A, B, C változót a latens X változó indikátorainak tekintjük.

A latens osztály-modell paraméterei:

a) a latens osztály valószínűségei (π_t^X) a latens változó eloszlását írják le,

b) a feltételes valószínűségek ($\pi_{it}^{\overline{A}X}, \pi_{jt}^{\overline{B}X}, \pi_{kt}^{\overline{C}X}$), amelyek az egyes változók valószínűség-eloszlásait írják le a latens változó egyes osztályaiban.

Ezek a feltételes valószínűségek hasonlatosak a faktorelemzés faktorsúlyaihoz. A feltételes valószínűségek segítségével jellemezhetjük a latens osztályokat. Egy adott latens osztályon belül a feltételes valószínűségek mutatják, hogy az adott latens osztály a megfigyelt változók milyen jellemzőit tükrözi, mi az adott osztályok típusjegye.

Mivel a latens osztály-modell paraméterei valószínűségek, nem lehetnek negatívak, és ki kell hogy elégítsék a következő feltételezéseket:

$$\sum_t^T \pi_t^X = \sum_i^I \pi_{it}^{\overline{A}X} = \sum_j^J \pi_{jt}^{\overline{B}X} = \sum_k^K \pi_{kt}^{\overline{C}X} = 1. \tag{11.6}$$

A (11.4) egyenlet alapján így

$$\pi_t^X = \sum_{i,j,k} \pi_{ijkt}^{ABCX}, \tag{11.7}$$

és

$$\pi_t^X\pi_{it}^{\overline{A}X} = \sum_{j,k} \pi_{ijkt}^{ABCX}. \tag{11.8}$$

Jelölje $\pi_{ijkt}^{ABC\overline{X}}$ annak a feltételes valószínűségét, hogy egy megfigyelés a t-edik latens osztályba esik, feltéve, hogy az (i, j, k) cellában van.

A feltételes valószínűség definíciója szerint:

$$\pi_{ijkt}^{ABC\overline{X}} = \frac{\pi_{ijkt}^{ABCX}}{\pi_{ijk}}. \tag{11.9}$$

A (11.9) azonosságot felhasználva a (11.7) és (11.8) egyenleteket a következőképpen írhatjuk:

$$\pi_t^X = \sum_{i,j,k} \pi_{ijk} \pi_{ijkt}^{ABC\overline{X}}, \tag{11.10}$$

és a feltételes valószínűségek:

$$\pi_{it}^{\overline{A}X} = \frac{1}{\pi_t^X} \sum_{j,k} \pi_{ijk} \pi_{ijkt}^{ABC\overline{X}}, \tag{11.11}$$

$$\pi_{jt}^{\overline{B}X} = \frac{1}{\pi_t^X} \sum_{i,k} \pi_{ijk} \pi_{ijkt}^{ABC\overline{X}}, \tag{11.12}$$

$$\pi_{kt}^{\overline{C}X} = \frac{1}{\pi_t^X} \sum_{i,j} \pi_{ijk} \pi_{ijkt}^{ABC\overline{X}}, \tag{11.13}$$

A fenti öt egyenletet likelihood egyenletnek nevezzük, és ezek segítségével becsüljük a paraméterek vektorát:

$$\boldsymbol{\pi} = \left[\pi_t^X, \pi_{it}^{\overline{A}X}, \pi_{jt}^{\overline{B}X}, \pi_{kt}^{\overline{C}X} \right].$$

A (11.9), (11.11)–(11.13) egyenleteknél az általánosság megszorítása nélkül feltételezhető, hogy a $\pi_t^X > 0$ és $\pi_{ijk} > 0$.

11.1.1A paraméterek becslése

A latens osztályelemzésnél elsőként alkalmazott determináns becslési módszer (Anderson, 1954, Lazarsfeld és Henry [1968]) nem biztosított mindig megfelelő megoldást, mint ezt már korábban említettük.

McHugh (1954) javasolt egy efficiens becslési eljárást, majd Goodman (1974, 1979) közölt egy általánosabb és egyszerűbb megoldást, amely maximum likelihood becslést adott a latens osztályvalószínűségre és a feltételes valószínűségre. Ezt az eljárást alkalmazták az MLLSA (Maximum Likelihood Latent Structure Analysis, Clogg, 1977) számítógépes programban.

A becslést az „iteratív proportional scaling” eljárással végezzük. Jelölje a $\wedge$ szimbólum a maximum likelihood becslést.

Jelölje p_{ijk} a megfigyelt valószínűséget ($p_{ijk} = f_{ijk}/n$), a megfigyelt relatív gyakoriságot az (i, j, k) cellában.

Megmutatható, hogy a maximum likelihood becslés kielégíti a következő egyenleteket (lásd pl. Haberman, 1974, 1979):

$$\widehat{\pi}_t^X = \sum_{i,j,k} p_{ijk} \widehat{\pi}_{ijkt}^{ABC\overline{X}} \tag{11.14}$$

$$\widehat{\pi}_{it}^{\overline{A}X} = \frac{1}{\widehat{\pi}_t^X} \sum_{j,k} p_{ijk} \widehat{\pi}_{ijkt}^{ABC\overline{X}} \tag{11.15}$$

$$\widehat{\pi}_{jt}^{\overline{B}X} = \frac{1}{\widehat{\pi}_t^X} \sum_{i,k} p_{ijk} \widehat{\pi}_{ijkt}^{ABC\overline{X}} \tag{11.16}$$

$$\widehat{\pi}_{kt}^{\overline{C}X} = \frac{1}{\widehat{\pi}_t^X} \sum_{i,j} p_{ijk} \widehat{\pi}_{ijkt}^{ABC\overline{X}}. \tag{11.17}$$

Kiindulunk a paramétervektor egy kezdeti becsléséből, amely kielégíti a (11.6) egyenletet:

$$\widehat{\boldsymbol{\pi}}(0) = \left[\widehat{\pi}_t^X(0),\ \widehat{\pi}_{it}^{\overline{A}X}(0),\ \widehat{\pi}_{jt}^{\overline{B}X}(0), \widehat{\pi}_{kt}^{\overline{C}X}(0)\right].$$

A kezdeti becslés és a (11.4) egyenlet alapján meghatározzuk a π_{ijkt}^{ABCX} kezdeti értékét ($\widehat{\pi}_{ijkt}^{ABCX}(0)$.)

Ezt behelyettesítve a (11.1) egyenletbe jutunk a π_{ijk} kezdeti becsléshez ($\widehat{\pi}_{ijk}(0)$).

Ezután a (11.9) egyenlet alapján kapjuk a $\pi_{ijkt}^{ABC\overline{X}}$ kezdeti becslését ($\widehat{\pi}_{ijkt}^{ABC\overline{X}}(0)$).

Behelyettesítve ezeket a kezdeti becsléseket a (11.14)–(11.17) egyenletbe a paraméterek újabb becsléséhez jutunk $\widehat{\pi}(1)$. Miután kerekítési hibák miatt a (11.6) feltétel nem biztos, hogy teljesül, a $\widehat{\pi}(1)$ paramétervektor elemeit a (11.6) a feltételnek megfelelően újraskálázzuk. Ezután az eljárást addig folytatjuk, amíg *vagy* el nem érünk egy megadott számú iterációig, *vagy* π_{ijkt}^{ABCX} ((11.4) egyenlet) egymást követő becslései közötti különbség valamilyen küszöbértéknél (tolerancia-szint) kisebb nem lesz.

11.1.2. A modell identifikálása

Láttuk, hogy a paramétervektor $\boldsymbol{\pi}$ (elemei a latens osztályok valószínűségei és a feltételes valószínűségek) a (11.4) és az (11.1) transzformációval meghatározza a $\{\pi_{ijk}\}$ valószínűségek vektorát:

$$\pi_{ijk} = f_{ijk}(\boldsymbol{\pi}), \tag{11.18}$$

ami azt fejezi ki, hogy a megfigyelt változók várható valószínűségei a modell paramétereinek a függvényei. Amennyiben a paramétertér bármely két $\boldsymbol{\pi}_1 \neq \boldsymbol{\pi}_2$ vektorához különböző várható valószínűségek tartoznak, vagyis a várható valószínűségeket egy és csak egy paramétervektor generálja, akkor a modellt identifikálhatónak nevezzük. A paraméterek identifikálhatóságának vizsgálatára Mettugh (1956) korábbi eredményeinek általánosításával Goodman (1974) adott módszert. Az eljárás a transzformáció Jacobi mátrixának vizsgálatán alapul. Az (11.1) és (11.4) egyenletekkel definiált transzformáció akkor és csak akkor lesz nemszinguláris, ha az első deriváltak

$$\mathbf{H} = \left\{\frac{\partial\{\pi_{ijk}\}}{\partial\boldsymbol{\pi}}\right\} \tag{11.19}$$

mátrixának a rangja egyenlő a paramétervektor nem redundáns elemeinek a számával (az oszlopok számával). A $\mathbf{H}$ elemei az (11.1) és (11.4) egyenletek alapján közvetlenül meghatározhatók. Pl. az π_t^X szerinti deriváltakat tartalmazó oszlop elemei

$$\frac{\partial\pi_{ijk}}{\partial\pi_t^X} = \pi_{it}^{\overline{A}X}\pi_{jt}^{\overline{B}X}\pi_{kt}^{\overline{C}X} - \pi_{iT}^{\overline{A}X}\pi_{jT}^{\overline{B}X}\pi_{kT}^{\overline{C}X}. \tag{11.20}$$

A minta alapján becsülhetjük a parciális deriváltakat is, így juthatunk a $\widehat{\mathbf{H}}$ becsléshez, majd a $\widehat{\mathbf{H}}$ mátrix rangjának vizsgálatával határozhatjuk meg, hogy a modell identifikálható-e.

A megfigyelt gyakoriság-táblázat szabadságfoka

$$IJK - 1.$$

A becsült paraméterek száma

$$(T-1) + T(I-1) + T(J-1) + T(K-1) = (I+J+K-2)T - 1.$$

Ha a paraméterek száma nagyobb, mint a várható valószínűségek (π_{ijk}) száma, akkor a modell nem identifikálható.

A modell szabadságfoka egyenlő $IJK - 1$ mínusz a $\mathbf{H}$ rangja, vagyis

$$szf = IJK - 1 - [(I+J+K-2)T - 1],$$

vagy

$$szf = IJK - 1 - [(I+J+K-M+1)T - 1], \tag{11.21}$$

ahol M a megfigyelt változók száma.

Vagy másképpen:

szf = [a nem redundáns várható valószínűségek száma]

− [a nem redundáns paraméterek száma].

A modell identifikálhatóságának szükséges feltétele, hogy a modell szabadságfoka ne legyen negatív. Ha a modell nem identifikálható, akkor a paraméterekre tett pótlólagos feltételezésekkel azzá tehető. Goodman (1974) mutatott be eljárást arra, hogyan tehetünk egy nem identifikálható modellt identifikálhatóvá.

11.1.3A modell illesztése

A latens osztály-modell illesztésének jóságát mérhetjük úgy, hogy a megfigyelt gyakoriságokat összehasonlítjuk a modell által becsült gyakoriságokkal a khi-négyzet statisztika segítségével.

A Pearson-féle χ^2 statisztika:

$$\chi^2 = n \sum_{i,j,k} \frac{(p_{ijk} - \widehat{\pi}_{ijk})^2}{\widehat{\pi}_{ijk}}, \tag{11.22}$$

ahol $p_{ijk} = f_{ijk}/n$ és f_{ijk} a megfigyelt gyakoriság az (ijk) cellában,

n a megfigyelések száma,

$szf = IJK - 1 - [(I+J+K-M+1)T - 1]$ szabadságfokkal.

Egy másik, különösen nagy minták esetében előnyösebb mutató, a loglikelihood hányados (L^2) :

$$L^2 = 2n \sum p_{ijk} \ln(p_{ijk}/\widehat{\pi}_{ijk}), \tag{11.23}$$

ami szintén aszimptotikusan χ^2 eloszlást követ a fenti szabadságfokkal.

A log-likelihood hányados-statisztika előnye abban is megmutatkozik, hogy a latens változó adott számú (T) osztálya mellett a latens osztály valószínűségeire és a feltételes valószínűségekre vonatkozó hipotéziseket is tesztelhetjük vele, mivel az L^2 particionálható. Ez különösen fontos a konfirmatív modell illesztése esetében.

A két statisztikának azonos az eloszlása, ha a modell igaz és a minta elemszáma nagy, azonban különböznek kis minták, vagy az adatok nagyon egyenetlen cellák közötti eloszlása esetén. Kis minták esetén így előfordul, hogy a két statisztika különböző következtetésre vezet.

A modell illeszkedését vizsgálhatjuk az egyes cellákban külön is, a standardizált reziduálisok segítségével

$$e_{ijk}^{ABC} = \left(f_{ijk} - \widehat{F}_{ijk}\right) \Big/ \sqrt{\widehat{F}_{ijk}} \ ,$$

ahol $\widehat{F}_{ijk} = n\widehat{\pi}_{ijk}$ a modell által becsült gyakoriság.

Az egymást követő modelleket összevethetjük az alapmodellel (általában a kölcsönös függetlenséget feltételező H_0 modellel), és a változás mérésére a következő indexet használhatjuk:

$$R_i^2 = \frac{L^2(H_0) - L^2(H_i)}{L^2(H_0)},$$

ahol H_i az i-edik hipotézist jelöli, és $L^2(H_i)$ az ennek megfelelő log-likelihood hányados.

Az R_i^2 a modell javulását méri, értéke 0 és 1 közé esik.

A modell kiértékelésénél alkalmazhatjuk még az $F_i = L^2(H_i)/$szfok mértéket, amely megfelel a regresszióelemzés F statisztikájának (Haberman [1978]).

Az F_i statisztikák alapján az R_i^2-hez hasonlóan mérhetjük a modell javulását is $R_\omega^2 = (F_0 - F_1)/F_0$. A kettő között az a különbség, hogy az R_ω^2-knál figyelembe vesszük a szabadságfokok változását is.

11.1.4. A modell paramétereinek értelmezése

A latens osztályvalószínűségek

A latens osztályvalószínűségek (π_t^X) a minta latens osztályokba kerülésének arányait fejezik ki. Két alapvető kérdés merül fel: az egyik, hogy hány kategóriája, osztálya legyen a latens változónak, a másik, hogy a minta milyen arányban oszlik meg a latens osztályok között.

Az exploratív elemzés során, miután nincs elméleti feltevésünk az osztályok számára, a „lépésenkénti" eljárást alkalmazhatjuk. Kiindulhatunk a legegyszerűbb modellből, hogy ti. csak egy osztályunk van ($T = 1$), vagyis a változók kölcsönösen függetlenek. Ha ezt a modellt elfogadjuk, akkor nincs is szükség a latens változóra. Ha elvetjük, akkor először próbálkozhatunk a kétosztályos-modellel, s ha ezt is el kell vetnünk, akkor a háromosztályossal, és így tovább, amíg elfogadhatóan illeszkedő modellt nem kapunk. A latens osztályok száma az illeszkedő modellben azt fejezi ki, hogy a minta elemei hány típusba – osztályba – sorolhatók. Az osztályvalószínűségek pedig a típusok súlyát, arányát fejezik ki a mintán belül.

A latens feltételes valószínűségek

A feltételes valószínűségek $\pi_{it}^{\overline{A}X}$, $\pi_{jt}^{\overline{B}X}$, $\pi_{kt}^{\overline{C}X}$, a latens osztály (t) és a megfigyelt változók (A, B, C) (i, j, k) kategóriáinak kapcsolódásait fejezik ki, így hasonlítanak a

faktorsúlyokhoz. A feltételes valószínűségek jelzik annak a valószínűségét, hogy egy megfigyelés egy adott latens osztályban a megfigyelt változók mely kategóriáiba esne. Ennek alapján következtetni tudunk a latens osztály jellegére, típusjegyeket rendelhetünk hozzá, és értelmezhetjük, megnevezhetjük őket.

A megfigyelt és a latens változók kapcsolatát jellemezhetjük az esélyhányados logaritmusával (Goodman, 1974). Az asszociáció X latens és A megfigyelt változók között:

$$\phi_{it}^{A} = \ln\left[\pi_{it}^{\overline{A}X}\pi_{i+1,t+1}^{\overline{A}X} \Big/ \pi_{i,t+1}^{\overline{A}X}\pi_{i+1,t}^{\overline{A}X}\right].$$

Becker és Clogg (1986) megmutatta, hogy az esélyhányados logaritmusa kapcsolatba hozható a tetrakorikus korrelációval.

A latens osztály becslése

A latens osztályelemzés egyik célja lehet, hogy a megfigyelt változók számát redukáljuk, és meghatározzuk az X latens változó becslését ($\widehat{X}$) a megfigyelt változók alapján.

A $\pi_{ijkt}^{ABC\overline{X}}$ fejezi ki azt a feltételes valószínűséget, hogy egy megfigyelés a latens változó t-edik osztályába esik, feltéve, hogy az A, B, C változók (i, j, k) cellájában van. Lásd a (11.9) képletet $\left(\pi_{ijkt}^{ABC\overline{X}} = \pi_{ijkt}^{ABCX} \Big/ \pi_{ijk}^{ABC}\right)$.

Így egy adott (i, j, k) cellába tartozó megfigyelésekre kiszámíthatjuk a $\pi_{ijkt}^{ABC\overline{X}}$ feltételes valószínűségeket, és ha t' latens osztályhoz tartozik a legnagyobb valószínűség, akkor az (i, j, k) cellába tartozó megfigyeléseket a t'-edik latens osztályba soroljuk. Ezzel az osztályozási szabállyal a téves besorolások becsült valószínűsége $\varepsilon_{ijk} = 1 - \pi_{ijkt}^{ABC\overline{X}}$, és az adott (i, j, k) cellában a várható hibás osztályozás száma:

$$E_{ijk} = n\varepsilon_{ijk}\pi_{ijk}. \tag{11.24}$$

Az egész táblában a hibás osztályozások (klasszifikációk) várható száma:

$$E_2 = \sum_{ijk} E_{ijk}. \tag{11.25}$$

A hibás osztályozások becsült számát megkapjuk, ha a (11.24) és (11.25) képletekbe behelyettesítjük a megfelelő $\widehat{e}_{ijk}$ és $\widehat{\pi}_{ijk}$ becsléseket.

Tételezzük fel, hogy a megfigyelések latens osztályba sorolását úgy végezzük el, hogy eltekintünk attól, hogy az adott megfigyelés a manifeszt változók melyik kategóriáiba esik. Ha ismerjük a latens osztályok valószínűségeit (π_t^X), akkor minden megfigyelést X latens változó t^*-edik osztályába sorolunk, ha $\pi_{t^*}^X = \max \pi_t^X$.

A feltétel nélküli osztályba sorolás várható hibája:

$$\varepsilon_1 = 1 - \pi_{t^*}^X,$$

és a hibák várható száma:

$$E_1 = n\varepsilon_1. \tag{11.26}$$

A latens változó (X) és a megfigyelt változók együttese (ABC) közötti asszociációt mérhetjük a λ mutatóval (Goodman és Kruskal, 1954):

$$\lambda_{X \cdot ABC} = (E_1 - E_2)/E_1. \tag{11.27}$$

Ez a mutató hasznosan kiegészíti a modell jóságát mérő korábban ismertetett statisztikákat.

11.1.5A konfirmatív latens osztály-modell

Az eddig tárgyalt modellben nem volt előzetes hipotézisünk a modell paramétereire – kivéve a valószínűségek axiómáiból adódó korlátozó feltételezést –, ugyanúgy nem volt előzetes ismeretünk a latens változó osztályainak számáról sem (és a latens változók számáról sem), így az exploratív elemzés logikája szerint jártunk el. Kiindultunk egy adott modellből (általában egy egyszerű, vagy a legegyszerűbb modellből), és a modell illeszkedésének vizsgálata után, lépések sorozatával juthattunk el az elfogadhatóan illeszkedő modellhez, amelynek a paramétereit értelmezhettük. A konfirmatív elemzésnél már előzetes feltételezésünk van a modell egy vagy több paraméterére, és a modellt ezekkel a korlátozásokkal akarjuk illeszteni a megfigyelt adatokhoz.

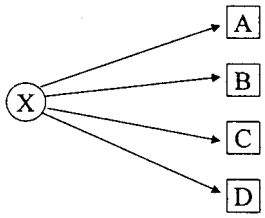
Általában a paraméterekre vonatkozó feltételezéseknek két fajtáját különböztetjük meg. Az egyikben azt írjuk elő, hogy a modell egy vagy több paramétere legyen egyenlő egymással, a másik feltételtípus, amikor *a priori* ismert egy vagy több paraméter értéke. A konfirmatív latens osztályelemzésnél a paraméterek lehetnek

a) rögzítettek egy adott értékhez,

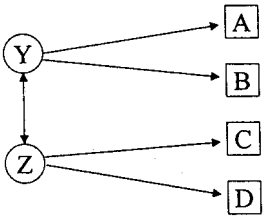
b) feltételes paraméterek, ismeretlenek, de egyenlők egy vagy több paraméterrel,

c) szabad, ismeretlen paraméterek.

Tekintsük az A, B, C, D megfigyelt változókat, és vizsgáljuk a következő, egy latens változós modellt:



Tegyük fel, hogy ez a modell nem illeszkedik elfogadhatóan az adatokhoz, így vizsgáljuk a következő ábrán látható modellt:



Ha Y és Z is dichotom latens változók, akkor a modellt kifejezhetjük úgy, mintha egy X latens változónk lenne négy kategóriával ($X \equiv YZ$).

Az X latens változó (1, 2, 3, 4) osztályai (Y, Z) együttes latens változó (1, 1), (1, 2), (2, 1) és (2, 2) kategóriáinak felelnek meg.

A modell paraméterei:

$$\pi_t^X = \pi_{rs}^{YZ}, \pi_{it}^{\overline{A}X} = \pi_{irs}^{\overline{A}YZ}, \pi_{jt}^{\overline{B}X} = \pi_{jrs}^{\overline{B}YZ}, \ldots \text{ stb.}$$

A modell feltételezi, hogy az A, B megfigyelt változók csak az Y, a C, D megfigyelt változók pedig csak a Z latens változótól függnek, valamint az A, B, C, D kölcsönösen függetlenek egymástól az adott Y, Z esetén, az Y és Z latens változók pedig kapcso-

lódnak egymáshoz. A következő pótlólagos feltételezések fejezik ki a fenti modellt:

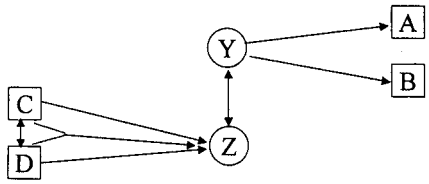
$$\pi^{\overline{A}YZ}_{i11} = \pi^{\overline{A}YZ}_{i12}, \qquad \pi^{\overline{A}YZ}_{i21} = \pi^{\overline{A}YZ}_{i22}$$

$$\pi^{\overline{B}YZ}_{j11} = \pi^{\overline{B}YZ}_{j21} \qquad \text{stb.}$$

A klasszikus példa az ilyen modellre Goodmantól (1974) származik: A és B dichotom változók a szavazás szándékát fejezik ki (1= republikánus jelölt, 2=más), C és D változók pedig a republikánus, illetve más jelöltről kialakított véleményt tudakolta (1=kedvező, 2=nem kedvező).

A latens szavazás szándékát Y, a latens véleményt Z változók fejezik ki, és az utolsó ábrának megfelelő modell szerint ezek magyarázzák az $ABCD$ asszociációt. A maximum likelihood arány $L^2 = 7{,}32$ (szabadságfok = 4), ami jó illeszkedést mutat, bár az Y és Z közötti becsült kapcsolat negatív és kicsi, elhanyagolható volt ($-0{,}08$ logaritmikus skálán), és ez nem magyarázható.

Clogg (1981) módosította az előző modellt a MIMC (Multiple-Indicator, Multiple-Cause) modellnek megfelelően:



A C és D változók nemcsak közvetlenül hatnak a latens változókra, hanem hat az interakciójuk is.

A CD együttes változó négy kategóriája (1, 2, 3, 4) a (C, D) változók (1, 1), (1, 2), (2, 1), (2, 2) kategóriáival függ össze. A $\pi^{\overline{A}YZ}_{irs}$ az A változó feltételes valószínűségét jelöli, azt a valószínűséget, hogy egy megfigyelés az A változó i-edik kategóriájába esik, feltéve, hogy az (Y, Z) változó (r, s) osztályba tartozik, és ehhez hasonlóan definiáljuk a többi feltételes valószínűségeket is.

A MIMC modell azt fejezi ki, hogy a $\pi^{\overline{A}YZ}_{irs}$ és a $\pi^{\overline{B}YZ}_{jrs}$ feltételes valószínűségek csak az Y latens változó r osztályától függnek (és nem függnek a Z latens változó s kategóriájától), a $\pi^{\overline{CD}YZ}_{n\ell rs}$ feltételes valószínűségek pedig csak a Z változó s osztályától (és nem függnek az Y változó r kategóriájától).

A modell fenti korlátozó feltételezéseit a következő azonosságok fejezik ki:

$$\pi^{\overline{A}X}_{11} = \pi^{\overline{A}X}_{12}, \quad \pi^{\overline{A}X}_{13} = \pi^{\overline{A}X}_{14} \tag{11.28}$$

$$\pi^{\overline{B}X}_{11} = \pi^{\overline{B}X}_{12}, \quad \pi^{\overline{B}X}_{13} = \pi^{\overline{B}X}_{14} \tag{11.29}$$

$$\pi^{\overline{CD}X}_{11} = \pi^{\overline{CD}X}_{13}, \quad \pi^{\overline{CD}X}_{21} = \pi^{\overline{CD}X}_{23} \tag{11.30}$$

$$\pi^{\overline{CD}X}_{31} = \pi^{\overline{CD}X}_{33}, \quad \pi^{\overline{CD}X}_{12} = \pi^{\overline{CD}X}_{14} \tag{11.31}$$

$$\pi^{\overline{CD}X}_{22} = \pi^{\overline{CD}X}_{24}, \quad \pi^{\overline{CD}X}_{32} = \pi^{\overline{CD}X}_{34}. \tag{11.32}$$

A fenti (11.28)–(11.32) azonosságokkal definiálhatjuk a MIMC-modellt a ($2\times2\times4$)-es $A \times B \times CD$ kereszttáblára. Az illeszkedés az említett példa esetén $L^2 = 7{,}04$ (szfok = = 2), és a Z hatása Y-ra +0,50, ami közel van az *a prioiri* elvárásunkhoz.

Kvázi latens osztályok

Amikor kérdőíves vizsgálatoknál az emberek véleményét, attitűdjét tudakolják, gyakran előfordul, hogy egy-egy kérdéskörnél az emberek egy csoportja minden kérdésre vagy a kérdések egy részére azonosan válaszol, pl. mindegyikkel egyetért, vagy mindegyikkel nem ért egyet. Bármi is legyen ennek a válaszállandóságnak az oka, figyelembe kell venni a modellben extra latens osztály szerepeltetésével. Tegyük fel, hogy az A, B, C, D megfigyelt dichotom változó (1,1,1,1) és (2,2,2,2) cellája alábecsült az eredeti kétosztályos modellben. Bővítsük a modellt két kvázi osztállyal. Ha egy megfigyelés történetesen a harmadik latens osztályba esik, akkor az (1,1,1,1) cellába kell tartoznia, így $\pi_{13}^{\overline{A}X} = \pi_{1}^{\overline{B}X}3 = \pi_{13}^{\overline{C}X} = \pi_{13}^{\overline{D}X} = 1$ és természetesen $\pi_{23}^{\overline{A}X} = \pi_{2}^{\overline{B}X}3 = \pi_{23}^{\overline{C}X} = \pi_{23}^{\overline{D}X} = 0$.

Ehhez hasonlóan adhatjuk meg a másik kvázi osztályhoz tartozás feltételét is. Ilyen modell illesztésére ad példát Hagenaars (1988).

11.1.6. Latens osztály-modell ordinális változókra

A szociológiai kutatásokban a nominális mérés mellett az ordinális mérési skálát használjuk a leggyakrabban. Általánosan elterjedt gyakorlat, hogy a sorrendi értékeket mint numerikus mennyiségeket tekintjük, és intervallummérést feltételező statisztikai eljárásokban szerepeltetjük az egyébként ordinális mérési szintű változókat. Ehelyett helyesebb, ha a latens osztály-modellt alkalmazzuk. A következőkben azt mutatjuk meg Clogg (1979) példája segítségével, hogyan lehet a latens osztály-modellt ordinális változókra alkalmazni.

Az elégedettség három indikátorát vizsgáljuk.

A: jelenti az elégedettséget a településsel (lakóhellyel),

B: jelenti az elégedettséget a hobbival (kedvenc időtöltés),

C: jelenti az elégedettséget a családdal.

Az adatok az 1975-ös General Social Survey-ből (USA) származnak.

Clogg az eredeti 7 fokozatú elégedettség-skálát átkódolta 3 kategóriába, ahol 1 jelentette a meglehetős (1=1, 2), 2 a közepes (2=3, 4), 3 az alacsony (3=5, 6, 7) elégedettségi szintet.

A megfigyelt háromdimenziós kereszttáblázatok:

B	A	$C = 1$	$C = 2$	$C = 3$
1	1	466	27	16
1	2	191	38	14
1	3	64	18	5
2	1	126	31	5
2	2	117	58	12
2	3	45	23	3
3	1	54	12	7
3	2	49	26	11
3	3	23	16	15

$n = 1472$

11.1. táblázat. Az elégedettség három trichotom indikátorának megfigyelt gyakorisága

Modell	Szabadságfok	Khi négyzet (χ^2)	Likelihood arány L^2	$R^2 = 1 - \frac{L^2(H)}{L^2(H_0)}$
H_0	20	339,43	259,17	
H_1	13	34,84	28,57	0,89
H_2	7*	2,32	2,36	0,99
H_3	14**	25,22	24,77	0,90
H_4	17	46,65	43,56	0,83

11.2. táblázat. Az illeszkedés jósága különböző latens osztály-modelleknél (a 11.1. táblázat adatai alapján)

Megjegyzés:

* A modellben $\pi_{23}^{\overline{B}X} = 0$, de ezt feltételként kezelve a szfok 6 helyett 7.

** A szabadságfok ténylegesen 12, de két becsült 0 paraméter miatt (azokat is feltételként kezelve) a szabadságfok 14.

Ahogy általában az exploratív elemzésnél, most is először a legegyszerűbb modellel próbálkozunk. H_0 a kölcsönös függetlenség modellje (vagyis azt tételezzük fel, hogy csak egy latens osztályunk van). A H_1 hipotézisben azt feltételezzük, hogy két latens osztályunk van. Ez a modell elfogadhatóan illeszkedik: $L^2 = 28{,}57$ (szfok = 13). A H_3 hipotézisben három latens osztályt tételezünk fel $L^2 = 2{,}36$ (szfok = 7). Ez a modell jól illeszkedik a megfigyelt adatokhoz, bár nem veszi figyelembe a változók rendezettségét. Ezért olyan feltételezéseket kell tennünk az illesztett modellre, hogy kielégítsük a következőket:

– az első latens osztály nem tartalmazhat alacsony elégedettségűeket (3. kategóriába esőket), de a közepes elégedettségűek, mint az 1. latens osztály hibája megengedett,
– a második latens osztályhoz tartozhatnak bármelyik elégedettségű kategóriákból,
– a harmadik latens osztályba nem tartozhatnak a meglehetősen elégedettek, a 2, és 3 kategóriák megengedettek, de a közepesen elégedettet válaszhibának tekintjük.

Ezzel a megkötöttséggel az X latens változó rendezett tulajdonságú lesz. Általánosabban azt mondhatjuk, hogy a t-edik latens osztályba a $(t-1,\ t,\ t+1)$ kategóriákba tartozó egyedek kerülhetnek.

A H_3 hipotézis ezeket a korlátozásokat tartalmazza:

$$\pi_{31}^{\overline{A}X} = \pi_{31}^{\overline{B}X} = \pi_{31}^{\overline{C}X} = 0, \tag{11.33}$$

$$\pi_{13}^{\overline{A}X} = \pi_{13}^{\overline{B}X} = \pi_{13}^{\overline{C}X}. \tag{11.34}$$

A likelihood-arány $h^2 = 24{,}77$ (szfok = 14) elfogadható illeszkedést mutat. A paraméterek becsléseit a 11.3. táblázat tartalmazza.

A latens osztályok valószínűségei

		H_2	H_3
1.	latens osztály	0,55	0,32
2.	latens osztály	0,41	0,66
3.	latens osztály	0,04	0,02

	A feltételes valószínűségek					
	1. latens osztály		2. latens osztály		3 latens osztály	
	H_2	H_3	H_2	H_3	H_2	H_3
$\pi_{1t}^{\overline{A}X}$	0,72	0,84	0,26	0,35	0,16	0,00*
$\pi_{2t}^{\overline{A}X}$	0,22	0,16	0,53	0,45	0,31	0,32
$\pi_{3t}^{\overline{A}X}$	0,06	0,00*	0,21	0,20	0,53	0,68
$\pi_{1t}^{\overline{B}X}$	0,79	0,92	0,32	0,42	0,15	0,00*
$\pi_{2t}^{\overline{B}X}$	0,14	0,08	0,50	0,39	0,00*	0,00**
$\pi_{3t}^{\overline{B}X}$	0,07	0,00*	0,18	0,19	0,85	1,00
$\pi_{1t}^{\overline{C}X}$	0,95	1,00	0,58	0,68	0,24	0,00*
$\pi_{2t}^{\overline{C}X}$	0,02	0,00**	0,36	0,25	0,28	0,39
$\pi_{3t}^{\overline{C}X}$	0,03	0,00*	0,06	0,07	0,48	0,61

11.3. táblázat. A háromosztályos modellek ($H_{2,3}$) paraméterei

$*$ előre megadott, feltételezett érték $**$ becsült érték

Clogg javasolt még egy négyosztályos modellt is, amelyben az első három latens osztály csak az elégedettség azonos szintjeit tartalmazhatta (t-edik osztályba csak a (t, t, t) cella kerülhetett), a negyedik osztályba pedig a nem konzisztens megfigyelések kerültek.

11.1.7. Mobilitásvizsgálat latens osztályelemzéssel

A következőkben Clogg (1981) vizsgálata alapján mutatjuk be a latens osztály-modellek alkalmazását mobilitástáblákra.

A hipotézis az, hogy a megfigyelt foglalkozási kategóriák közötti mobilitási folyamatot a mögöttes, különböző, latens osztályok befolyásolják. A latens osztály fogalom nem azonos a weberi osztályfogalommal, bár elég nagy a hasonlóság.

A weberi koncepció szerint a társadalmi osztályok az egyének olyan együttesei, amelyeknek közös a mobilitási esélyük. A latens osztály-modellben a latens osztályok jellemzője, hogy az osztályokon belüli statisztikusan független az eredeti és az elért (az apa és a fiú) státus.

Tekintsük az $I \times I$ típusú mobilitás táblát, ahol a sorokban az apa foglalkozása (F), az oszlopokban pedig a fiú foglalkozása található (S).

A várható valószínűség az (i, j) cellában $(i = 1, \ldots, I, \ \ j = 1, \ldots, J)$ π_{ij}.

Az apa foglalkozásának (eredeti státus) gyakoriság-eloszlása (peremeloszlás) $\pi_i^F = \sum_j \pi_{ij}$, és a mobilitási ráta $r_{ij} = \pi_{ij}/\pi_i^F$.

Az eredeti státus (apa foglalkozása) természetes módon megelőzi az elért, adott státust. Feltételezhetjük, hogy a kettő között létezik egy közvetítő, közbülső latens változó (X), amely magyarázza az F és S asszociációját.

Ha X megfigyelhető és diszkrét, akkor a loglineáris modellt illeszthetjük a mobilitástáblára ($F \times S|X$). Ha az F, S és X változók intervallummérési szintű manifeszt

változók, akkor a lineáris regresszió módszerét alkalmazhatjuk, és a modell feltétele szerint a parciális regressziós együttható $\beta_{FS\cdot X} = 0$. Ha F, S és X is diszkrét és X latens változó, akkor a latens osztály-modellt illeszthetjük a mobilitástáblára:

$$\pi_{ij} = \sum_{t=1}^{T} \pi_t^X \pi_{it}^{\overline{F}X} \pi_{jt}^{\overline{S}X}. \tag{11.35}$$

A $\pi_{it}^{F\overline{X}}$ jelöli annak a valószínűségét, hogy egy megfigyelés az i-edik eredeti státusból a t-edik latens osztályba kerül.

A $\pi_{it}^{F\overline{X}}$ feltételes valószínűség a következőképpen írható fel:

$$\pi_{it}^{F\overline{X}} = \pi_{it}^{FX}/\pi_i^F = \pi_{it}^{\overline{F}X}\pi_t^X/\pi_i^F. \tag{11.36}$$

A mobilitási ráta a feltételes valószínűségek függvényében:

$$r_{ij} = \sum_{i=1}^{T} \pi_{it}^{F\overline{X}} \pi_{jt}^{\overline{S}X}. \tag{11.37}$$

A (11.37) egyenlet azt fejezi ki, hogy az i státusból a j státusba kerülésnek a valószínűsége egyenlő az i státusból a t-latens osztályba és a t osztályból a j státusba kerülés valószínűségeinek szorzatösszegével (ahol feltételezzük, hogy a $(i \longrightarrow t)$ és $(t \longrightarrow j)$ függetlenek minden t-re $(t = 1, \ldots, T)$).

Tekintsünk most néhány speciális esetet.

A H_0 hipotézis szerint F és S függetlenek egymástól, $T = 1$ és így $\pi_{ij} = \pi_i^F \pi_j^S$. Ez a modell írja le a tökéletes vagy teljes mobilitást. A mobilitás ebben a modellben véletlenszerű, csak az eredeti és a tényleges peremeloszlástól függ.

Ha $T = 2$, akkor a két latens osztályon belül véletlenszerű a mobilitás, de a két osztály között létezik „osztálykorlát" (a mobilitás a $\pi_{it}^{\overline{F}X}$ és $\pi_{jt}^{\overline{S}X}$ feltételes valószínűségektől függ). Tekintsük Goodman (1969, 1972) kvázi-teljes mobilitási modelljét, amelyben $T = I + 1$ és ahol $\pi_{ii}^{\overline{F}X} = \pi_{ii}^{\overline{S}X} = 1$ $(i = 1, \ldots, I)$.

A (11.35) egyenlet ekkor a következő:

$$\pi_{ij} = \begin{cases} \pi_i^X + \pi_{I+1}^X \pi_{i,I+1}^{\overline{F}X} \pi_{j,I+1}^{\overline{S}X}, & \text{ha } i = j \\ \pi_{I+1}^X \pi_{i,I+1}^{\overline{F}X} \pi_{j,I+1}^{\overline{S}X}, & \text{ha } i \neq j. \end{cases} \tag{11.38}$$

A latens változó első I osztálya determinisztikus, mivel a feltétel szerint $\pi_{ii}^{\overline{F}X} = \pi_{ii}^{\overline{S}X} = 1$ és $\pi_{it}^{\overline{F}X} = \pi_{jt}^{\overline{S}X} = 0 \quad i \neq j$ esetén.

Vezessük be a következő azonosságokat:

$$\alpha_i = \pi_{I+1}^X \pi_{j,I+1}^{\overline{S}X},$$

$$\beta_j = \pi_{j,I+1}^{\overline{S}X},$$

és

$$\gamma_i = 1 + \pi_i^X/\alpha_i\beta_i \quad i = 1, \ldots, I.$$

Ekkor a (11.38) egyenletet a következőképpen fejezhetjük ki:

$$\pi_{ij} = \begin{cases} \alpha_i\beta_i\gamma_i, & \text{ha } i = j \\ \alpha_i\beta_j, & \text{ha } i \neq j. \end{cases} \tag{11.39}$$

Goodman (1972) a γ_i-t immobilitási indexnek nevezte. Ez a kvázi-teljes mobilitás modellje, amelyben azt tételezzük fel, hogy mindegyik státuskategóriában vannak olyanok, akik az adott státusban maradnak (az immobilitás determinisztikus ezekben a latens osztályokban), és van egy olyan latens osztály, ahol a mobilitás véletlenszerű, az apa és a fiú peremeloszlástól függ csak a feltételes táblában. A kvázi-teljes mobilitást Goodman (1974) módosította úgy, hogy az I determinisztikus latens státust (ahol a státus-örökség 1 valószínűségű esemény) két latens osztállyal bővítette a latens mobilok számára. Ezt a modellt Goodman kvázi-latens struktúrának nevezte, amelyben:

$$\pi_{ij} = \begin{cases} \pi_i^X + \sum_{t=I+1}^{I+2} \pi_t^X \pi_{it}^{\overline{F}X} \pi_{jt}^{\overline{S}X}, & \text{ha } i = j \\ \sum_{t=I+1}^{I+2} \pi_t^X \pi_{it}^{\overline{F}X} \pi_{jt}^{\overline{S}X}, & \text{ha } i \neq j. \end{cases} \tag{11.40}$$

Vezessük be a kvázi-mobilitás modelljénél alkalmazott jelöléseket:

$$\alpha_{i1} = \pi_{I+1}^X \pi_{i,I+1}^{\overline{F}X}$$
$$\alpha_{i2} = \pi_{I+2}^X \pi_{i,I+2}^{\overline{F}X}$$
$$\beta_{j1} = \pi_{j,I+1}^{\overline{S}X}$$
$$\beta_{j2} = \pi_{j,I+2}^{\overline{S}X}.$$

Ezek alapján:

$$\pi_{ij} = \begin{cases} \pi_i^X + \sum_{k=1}^{2} \alpha_{ik}\beta_{jk}, & \text{ha } i = j, \\ \sum_{k=1}^{2} \alpha_{ik}\beta_{jk}, & \text{ha } i \neq j. \end{cases} \tag{11.41}$$

Legyen $s_{ij} = \sum_k \alpha_{ik}\beta_{jk}$ és $\gamma_i^* = 1 + \pi_i^X / s_{ij}$.

A fentiek alapján:

$$\pi_{ij} = \begin{cases} s_{ij}\gamma_i^*, & \text{ha } i = j \\ s_{ij} & \text{ha } i \neq j. \end{cases} \tag{11.42}$$

A γ_i^* az i-edik státus-örökítésének a mértéke ($\gamma_i^* \leq 1$), az i-edik státusban maradók azon mértéke, amely nem magyarázható a modell által várható immobilitással (két latens mobil osztályt feltételezve).

A fenti modelleket Clogg (1981) illesztette angol és dán intergenerációs foglalkozási mobilitás táblázatokra. A dán adathalmaz forrása Svalastoga (1959), az angol tábláé Glass (1954). Mindkét táblázat 5×5-ös, ahol a foglalkozási kategóriák: (1) vezető értelmiség, (2) magasan kvalifikált irányító, (3) adminisztratív irányító, (4) szakmunkás, adminisztrátor, (5) betanított munkás.

Dánia ($n = 2391$)

Apa státusa*	Megkérdezett státusa				
	1	2	3	4	5
1	18	17	16	4	2
2	24	105	109	59	21
3	23	84	289	217	95
4	8	49	175	348	198
5	6	8	69	201	246

Anglia ($n = 3524$)

	1	2	3	4	5
1	50	45	8	18	8
2	28	174	84	154	55
3	11	78	110	223	96
4	14	150	185	741	447
5	0	42	72	320	411

11.4. táblázat. Intergenerációs mobilitási táblázat

* ahol a kategóriák a következő nyolc foglalkozási kategória összevonásával keletkeztek: (1) szellemi irányító, (2) vezető, (3) magasan kvalifikált irányító, (4) irányító, (5) adminisztrátor, (6) szakmunkás, (7) betanított munkás, (8) segédmunkás. Az összevonás (1), (2, 3), (4), (5, 6), (7, 8).

A 11.4. táblázatra először a H_0 kölcsönös függetlenség modelljét illesztjük, majd egy két latens osztályos modellt (H_1), és egy három latens osztályos modellt (H_2). A H_3 hipotézis a kvázi-teljes mobilitás hipotézisét fogalmazza meg.

A H_4 hipotézis a kvázi-teljes mobilitást fogalmazza meg arra az esetre, amikor az 1., 3. és 5. foglalkozási kategóriákra írjuk elő a státus-örökséget (a 2. és 4. kategóriákra nem), és két olyan latens osztályunk van, amelyekben a függetlenség, teljes mobilitás érvényes. A H_4 modell feltételei:

$$\pi_{11}^{\overline{F}X} = \pi_{11}^{\overline{S}X} = 1$$
$$\pi_{32}^{\overline{F}X} = \pi_{32}^{\overline{S}X} = 1$$
$$\pi_{53}^{\overline{F}X} = \pi_{53}^{\overline{S}X} = 1,$$

és a 4. és 5. latens osztályra nincs semmilyen korlátozó feltételünk.

A H_5 hipotézis a H_4 feltételeit kiegészíti a következőkkel:

$$\pi_{it}^{\overline{F}X} = \pi_{it}^{\overline{S}X} \quad \text{minden} \quad i = 1, \ldots, 5 \quad \text{és} \quad t = 4{,}5\text{-re}.$$

Ezek a feltételek azt írják elő, hogy a 4. és 5. latens osztályban az apa és a fiú foglalkozási peremeloszlásai azonosak, vagyis mindkét latens osztályban a foglalkozási kategóriák eloszlásai az idővel nem változtak.

Ha az előző hipotéziseknek megfelelő modellek illeszkedését vizsgáljuk, a 11.5. táblázatból azt láthatjuk, hogy a teljes mobilitás modellje nem illeszkedik az adatokhoz.

A H_1 modell illeszkedése is rossz, a három latens osztályos modell viszont már elég jól illeszkedik.

Berger (1982) az eredeti 8×8-as mátrixot vizsgálva arra az eredményre jutott, hogy a mobilitási tábla mögött egy háromosztályos társadalmi struktúra húzódik meg, mégpedig létezik egy felső osztály ehhez tartoznak az (1, 2, 3) kategóriák, egy középosztály

Modell	Anglia szfok	χ^2	L^2	Dánia &szfok	χ^2	L^2	
H_0 (függetlenség)	16	1225,59		821,89	16	754,10	654,10
H_1 (két latens osztály)	9	202,14		177,97	8	133,64	116,23
H_2 (három latens osztály)	2	50,55		50,85	2	33,28	32,67
H_3 (kvázi-teljes mobilitás $T = I + 1$)	11	327,3		249,50	11	270,30	248,70
H_3 (kétosztályos latens struktúra $T = I + 2$	6	9,35		9,76	6	8,06	8,23
H_4 (korlátozott kétosztályos latens struktúra $T = I + 2$	13	41,44		41,18	14	32,47	32,87

11.5. táblázat. Latens osztály és kvázi-latens stuktúramodellek illesztése a mobilitási táblára

(4, 5) kategóriákkal, és egy alsóosztály (6, 7, 8) kategóriákkal. Láttuk, hogy az összevont 5×5-ös táblázat, amelyre a bemutatott modelleket C. Clogg illesztette, a foglalkozási kategóriákat ettől eltérően vonta össze. Ez a két modell nem tartalmazott külön feltételeket a paraméterekre, így minden latens osztályon belül a teljes mobilitás hipotézise érvényesül.

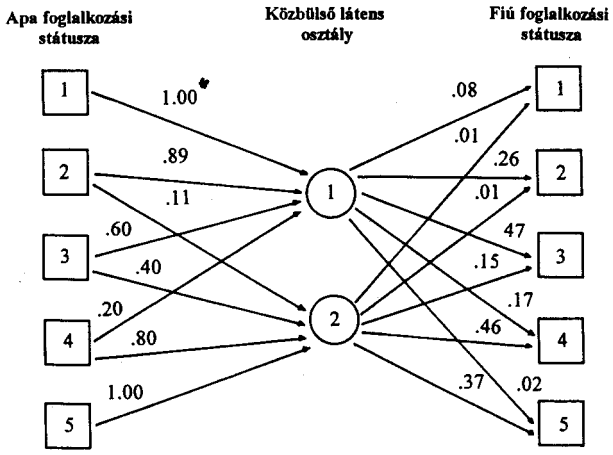
A H_3 modellnél a kvázi-latens hipotézisek alapján feltételeztük, hogy az 1., 3. és 5. státusokban létezik a státus-átörökítés, és két olyan latens osztályt definiáltunk, amelyekben független az apa és a fiú foglakozási státusa. Ez a modell, ahogy a 11.5. táblázatban látható, mindkét országban nagyon jól illeszkedik a mobilitási táblához.

A H_3 modell módosítja a H_4-et úgy, hogy a negyedik és ötödik latens osztályban az apa és a fiú foglalkozáseloszlása azonos marad.

Latens osztályok valószínűségei	$\widehat{\pi}_1^X$	$\widehat{\pi}_2^X$	$\widehat{\pi}_3^X$	$\widehat{\pi}_4^X$	$\widehat{\pi}_5^X$
Anglia	0,012	0,013	0,061	0,212	0,701
Dánia	0,006	0,058	0,059	0,335	0,541
Feltételes valószínűségek a 4. latens osztályban					
	1	2	3	4	5
Anglia	0,095	0,479	0,173	0,226	
Dánia	0,067	0,363	0,340	0,179	
Feltételes valószínűségek a 5. latens osztályban					
	1	2	3	4	5
Anglia	0,0000	0,056	0,128	0,532	0,285
Dánia	0,0000	0,000	0,209	0,510	0,281

11.6. táblázat. A H_5 hipotézis paramétereinek becslése

A (11.37) egyenletben a mobilitási ráta a feltételes valószínűségek szorzatainak összege. A kétosztályos modell dániai eredményeit láthatjuk a következő ábrán:



Megjegyzés: Az 1,00 együttható a modell identifikálhatósága érdekében tett feltételezés (*a priori* érték).

11.2. ábra. Kétosztályos modell

11.2. Latens tulajdonság-modell

A latens tulajdonság-modellben (latent trait model) a kvalitatív (diszkrét értékekkel rendelkező) megfigyelt változók asszociációit kvantitatív, folytonos latens változóval magyarázzuk. A latens változót latens tulajdonságnak, latens jellemzőnek nevezzük. A trait a pszichológiában használatos kifejezés, olyan általános tulajdonságot, jellemző vonást jelent, amelynek segítségével a személyiségeket meg tudjuk különböztetni. Statisztikai értelemben a latens tulajsonság (trait) megfelel a közös faktor terminológiának.

A latens változónak (X) $T = n$ „osztálya" van (ahol n a minta elemszáma). Jelölje μ_i a latens változó értékét a t-edik megfigyelés esetén ($i = 1, \ldots, n$).

Tételezzük fel, hogy a megfigyelt változók dichotom változók, jelölje őket Y_j ($Y_j = 1$ vagy $Y_j = 0$). Jelölje Y_{ij} az i-edik megfigyelés válaszát a j-edik változóra, y_{ij} pedig a megfigyelt értéket.

Jelölje a feltételes valószínűségeket $\pi_{j|i} = p(Y_{ij} = 1|X = \mu_i)$.

Attól függően, hogy a megfigyelt és a latens változók közötti kapcsolatot milyen függvénytípussal írjuk le, különbözőképpen fejezzük ki a feltételes valószínűségeket. A logisztikus függvényt feltételezve:

$$\pi_{j|i} = P(Y_{ij} = 1 \mid X = \mu_i) = \frac{1}{1 + \exp(-z_{ij})},$$

ahol $z_{ij} = \alpha_j \mu_i + c_j$, ahol c_j a j-edik változó „nehézségi fokát" jelzi (azt, hogy milyen arányban választották a mintából), α_j a faktorsúlyokhoz hasonló értelmű, a változó diszkriminatív erejét fejezi ki.

Normális eloszlást feltételezve:

$$\pi_{j|i} = P(Y_{ij} = 1 \mid X = \pi_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_{ij}} \exp\left(-\frac{1}{2}z^2\right) dz.$$

A Rasch-féle modell a feltételes valószínűségeket a következőképpen határozza meg:

$$\pi_{j|i} = (e^{\mu_i - c_j}) \Big/ \left(1 + e^{\mu_i - c_j}\right), \tag{11.43}$$

ahol a nevező $\Delta_{ij} = (1 + e^{\mu_i - c_j})$ az ún. normalizáló konstans.

Vegyük észre, hogy a Rasch-modell speciális faktorelemző modell, ahol a megfigyelt és a latens változók közötti kapcsolat nemlineáris, és amelyben a faktorsúlyok azonosak (11.1).

A Rasch-modell likelihood függvénye (Andersen, 1980)

$$L(\boldsymbol{\mu}, \mathbf{c}|\mathbf{y}) = D^{-1} \exp\left(\sum_i \mu_i y_{i+} - \sum_j c_j y_{tj}\right), \tag{11.44}$$

ahol $D^{-1} = \prod_i \prod_j \Delta_{ij}$.

Az y_{i+} az i-edik megfigyelés „1” válaszainak a száma.

Összesen n y_{i+} értéket figyeltünk meg a J számú változóra, amelyek $J+1$ különböző értéket vehetnek fel. Legyen n_r, $(r = 0, 1, \ldots, J)$ azoknak a száma, akikre fennáll, hogy $y_{i+} = r$ és $\sum_r n_r = n$.

Jelölje R az így csoportosított változót. Duncan (1984) megmutatta, hogy a Rasch-féle modell feltételes formája felírható a következő módon:

$$\ln(F_u) = \lambda + \lambda_{R(r)} + \sum_j \lambda_{j(kj)}, \tag{11.45}$$

ahol u a gyakoriságtáblázat egy cellája $\{k_1, k_2 \ldots, k_J\}$, $r = \sum_j k_j$, $\lambda_{R(r)}$ latens változó fő-hatása, $\lambda_{j(kj)}$ a j-edik változó fő-hatása.

A modell ebben a formában túlparametrizált, ezért pótlólagos feltételezést kell tennünk (ez a szokásos feltételezés):

$$\sum_r \lambda_{R(r)} = 0,$$

$$\lambda_{j(1)} + \lambda_{(j2)} = 0,$$

és a redundancia elkerülése érdekében:

$$\lambda_{R(J)} = 0.$$

A (11.45) loglineáris modell azt fejezi ki, hogy a megfigyelt változók függetlenek az adott R latens változó esetén.

Ez a modell megfelel a latens osztály-modellnek, ahol $X = R$, $\pi_n^X = n_r/n$ és $T = J + 1$.

11.3. Latens profil-modell

A latens profil-modellben a folytonos megfigyelt változók kapcsolatait diszkrét latens változókkal magyarázzuk.

A latens profil-modellben feltételezzük, hogy létezik a megfigyelt változóknak a latens változókra vonatkozó feltételes együttes sűrűségfüggvénye:

$$g(\mathbf{x} \mid \mathbf{y}). \tag{11.46}$$

Jelölje a latens változók sűrűségfüggvényét $h(\mathbf{y})$.

Az $\mathbf{x}$ együttes sűrűségfüggvénye kifejezhető a következőképpen:

$$f(\mathbf{x}) = \int_{R_\mathbf{y}} h(\mathbf{y}) g(\mathbf{x}|\mathbf{y}) d\mathbf{y}, \tag{11.47}$$

ahol $R_\mathbf{y}$ az $\mathbf{y}$ változó értéktere.

Miután $\mathbf{y}$ latens, nem megfigyelt változókat tartalmaz, ezért az $\mathbf{y}$ változókról az $\mathbf{x}$ manifeszt változók ismeretében tudunk csak valamit mondani. Ezt a feltételes sűrűségfüggvénnyel fejezzük ki:

$$h(\mathbf{y}|\mathbf{x}) = h(\mathbf{y}) g(\mathbf{x}|\mathbf{y}) / f(\mathbf{x}). \tag{11.48}$$

Ahhoz azonban, hogy meghatározzuk $h(\mathbf{y}|\mathbf{x})$ feltételes sűrűségfüggvényt, ismernünk kell h és g sűrűségfüggvényeket, de csak az f-et tudjuk becsülni. Ebből látszik, hogy (11.47) és (11.48) alapján nem tudjuk egyértelműen definiálni a $h(\mathbf{y}|\mathbf{x})$-et, ezért további feltételeket kell tennünk h és g függvényekre.

A modellben csupán azt tételezzük fel, hogy az $\mathbf{x}$ és $\mathbf{y}$ változók függnek egymástól. Ha az x változók függnek az y változóktól, akkor az x változók korrelálnak egymással is. Ha az x_1 és x_2 függ az y_1 változótól, akkor az y_1 változó idézi elő az x_1 és x_2 közötti korrelációt. Ha x_1 és x_2 korrelálatlanok, akkor nem tételezhetjük fel, hogy van valami közös bennük. Ha az x_1 és x_2 korrelálatlanok akkor, ha az y_1 változót konstansként tartjuk (ha a hatását kiszűrjük), nincs szükség további y változókra. Általában ha az x változók közötti korrelációkat az y változók idézik elő, akkor az y változók hatását kiszűrve, az x változók korrelálatlanok lesznek. Ez a feltételes függetlenség axiómája. Eszerint

$$g(\mathbf{x}|\mathbf{y}) = \prod_i g_i(x_i \mathbf{y}) \tag{11.49}$$

adott r számú latens változó esetén. Feltételezzük tehát, hogy az r számú latens változó teljesen megmagyarázza az x változók kapcsolatait. Az $f(\mathbf{x})$ függvényt a következőképpen írhatjuk

$$f(\mathbf{x}) = \int_R h(\mathbf{y}) \prod_i g_i(x_i|\mathbf{y}) d\mathbf{y}. \tag{11.50}$$

Könnyű belátni, hogy ha a megfigyelt változók binárisak, és a latens változónak K kategóriája van, akkor

$$g_i(x_i|y_j) = P(x_i|y_j) = \pi_{ij}^{x_i} (1 - \pi_{ij})^{1-x_i},$$

és

$$g(\mathbf{x}|\mathbf{y}) = \prod_i \pi_{ij}^{x_i} (1 - \pi_{ij})^{1-x_i}. \tag{11.51}$$

Ha annak a valószínűsége, hogy egy megfigyelés a j-edik latens kategóriába esik η_j $(\sum_j \eta_j = 1)$, akkor

$$f(\mathbf{x}) = \sum_j \eta_j \prod_i \pi_{ij}^{x_i}(1-\pi_{ij})^{1-x_i}, \tag{11.52}$$

és annak a valószínűsége, hogy az $\mathbf{x}$ megfigyelés a j-edik latens kategóriába esik:

$$h(j|\mathbf{x}) = \eta_j \prod_i \pi_{ij}^{x_i}(1-\pi_{ij})^{1-x_i}/f(\mathbf{x}). \tag{11.53}$$

Vegyük észre, hogy ez megegyezik a latens osztály-modellel bináris megfigyelt változók esetén.

Térjünk most vissza az eredeti problémára, arra, hogy a metrikus manifeszt változókat kategorikus latens változókkal akarjuk kifejezni. Ekkor a megfigyelt változók együttes sűrűségfüggvénye a (11.50) egyenlet szerint:

$$f(\mathbf{x}) = \sum_j^K \eta_j \prod_i^m g_i(x_i|j), \tag{11.54}$$

ahol $g_i(x_i|j)$ az x_i változó feltételes eloszlása a j-edik osztályban.

A (11.54) modell becslését D. J. Bartholomew (1987) maximum likelihood becslési eljárása alapján mutatjuk be.

11.3.1. Maximum likelihood becslés

Ez a becslési eljárás lényegét tekintve megegyezik a kategorikus manifeszt változókra kidolgozott eljárással, a különbség $g_i(x_i|j)$ eltérő megválasztásából adódik.

A likelihood függvény:

$$L = \sum_{k=1}^n \ln f(\mathbf{x}_k) = \sum_{k=1}^n \ln \left\{ \sum_j^K \eta_j \prod_i^m g_i(x_i|j) \right\}. \tag{11.55}$$

Ezt a függvényt kell maximalizálni a $\sum \eta_j = 1$ feltétel mellett:

$$\phi = L + \lambda \sum_j \eta_j.$$

A parciális deriválásokat elvégezve:

$$\frac{\partial \phi}{\partial \eta_j} = \sum_k \{g(\mathbf{x}_k|j)/f(\mathbf{x}_k)\} + \lambda. \tag{11.56}$$

A Bayes-tétel szerint:

$$h(j|\mathbf{x}) = \eta_j g(\mathbf{x}_k|j)/f(\mathbf{x}_k). \tag{11.57}$$

Ezt behelyettesítve a (11.56) egyenletbe:

$$\sum_k h(j|\mathbf{x}_k) = \lambda \eta_j.$$

Mindkét oldalt j szerint összegezve a $\lambda = n$ azonossághoz jutunk, így az első becslő egyenletünk:

$$\widehat{\eta}_j = \sum_k^n h(j|\mathbf{x}_k)/n. \tag{11.58}$$

Tételezzük fel, hogy

$$g_i(x_i|j) \equiv g(x_i|\theta_{ij}).$$

Ekkor

$$\frac{\partial \phi}{\partial \theta_{ij}} = \sum_k \eta_j \frac{\partial g}{\partial \theta_{ij}} \Big/ g(x_{ik}|\theta_{ij}). \tag{11.59}$$

Tételezzük fel továbbá, hogy a $g(x_i|\theta_{ij})$ θ_{ij} várható értékű és egységnyi varianciájú normális eloszlást követ. Így:

$$\frac{\partial g}{\partial \theta_{ij}} = (x_{ik} - \theta_{ij}) g(x_i|\theta_{ij}).$$

A (11.59) egyenletből ezek alapján:

$$\sum_k (h_j|\mathbf{x})(x_{ik} - \theta_{ij}) = 0,$$

amiből a második becslő egyenletet kaphatjuk:

$$\widehat{\theta}_{ij} = \sum_k x_{ik} h(j|\mathbf{x}_k) \Big/ \sum_k h(j|\mathbf{x}_k), \tag{11.60}$$

vagy

$$\widehat{\theta}_{ij} = \sum_k x_{ik} h(j|\mathbf{x}_k) \Big/ n\widehat{\eta}_j.$$

A $h(j|\mathbf{x}_k)$ a $\{\eta_j\}$ és $\{\theta_{ij}\}$ paramétereknek a függvénye.

A becslés a (11.58) és (11.60) egyenletek alapján végezzük:

a) kiindulunk az *a posteriori* valószínűségek kezdeti becsléseiből:

$$\{h(j|\mathbf{x}_k)\},$$

b) a (11.58) és (11.60) egyenletek alapján megkapjuk a $\{\eta_j\}$ és $\{\widehat{\theta}_{ij}\}$ első becsléseit,

c) ezeket behelyettesítve a (11.57) egyenletbe a $\{h(j|\mathbf{x}_k)\}$ újabb becsléseihez jutunk,

d) visszatérve a b) lépéshez, addig folytatjuk a számításokat, amíg az eljárás nem konvergál.

A latens osztály tagságának becslése (Latens osztályba sorolás)

A megfigyelési egységeket a latens osztálytagság *a posteriori* valószínűségei alapján soroljuk be a latens osztályokba.

Ha több mint két osztályunk van, két osztály (j és k) relatív *a posteriori* valószínűsége:

$$h(j|\mathbf{x})/h(k|\mathbf{x}).$$

Mivel $h(j|\mathbf{x}) = \eta_j f(\mathbf{x}|j)/f(\mathbf{x})$, az $\mathbf{x}$ megfigyelés inkább a j-edik osztályba tartozik, ha

$$\frac{h(j|\mathbf{x})}{h(k|\mathbf{x})} > 1.$$

Ha az x_i változók együttes eloszlása normális minden osztályban:

$$f(\mathbf{x}|j) = (2\pi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\sum_{i}^{m}(x_i - \mu_i(j))^2\right\}.$$

Így

$$f(\mathbf{x}|j)/f(\mathbf{x}|k) = \exp\left\{\sum_{i}^{m} x_i\mu_i(j) - \frac{1}{2}\sum_{i}^{m}\mu_i^2(j) - \sum_{i}^{m} x_i\mu_i(k) + \frac{1}{2}\sum_{i}^{m}\mu_i^2(k)\right\}.$$

Az osztályba sorolást a

$$\sum_{i}^{m} x_i\mu_i(j) - \frac{1}{2}\sum_{i}^{m}\mu_i^2(j) + \log\eta_j \tag{11.61}$$

értékek alapján végezzük. A legvalószínűbb latens osztály az, amely esetén (11.61) maximális értéket ad. Láthatjuk, hogy a normális eloszlás feltételezésével végzett osztályba sorolás az adatok lineáris függvénye.

12. fejezet

Az exploratív faktorelemzés módszerei

12.1. Főkomponens-elemzés

Ha vizsgálni akarjuk m darab változó egymás közötti kapcsolatát, akkor a paraméterek száma, amit becsülnünk és értelmeznünk kell, meglehetősen nagy lesz.

	m várható érték
	m szórás
	$\frac{(m^2-m)}{2}$ kovariancia
Összesen:	$\frac{(m^2-m)}{2} + 2m$ paraméter

Milyen problémákat okozhat ez:

a) Ha két változónk van, akkor csak öt paramétert, ha $m = 10$, akkor 65 paramétert, és ha $m = 30$, akkor 495 paramétert kell becsülni és értelmezni.

A vizsgált változók számának növelésével a paraméterek számának ilyen arányú növekedése nagymértékben megnehezíti a többletinformáció felhasználását az elemzésben.

b) A többváltozós regressziós modellnél láttuk, milyen problémát jelent, ha a magyarázó változók egymással páronként korreláltak. Ilyen esetben a többváltozós becslő függvény együtthatói (b) nem határozhatók meg egyértelműen. Mivel a gazdasági-társadalmi életben található változókra inkább az a jellemző, hogy egymással összefüggnek, a b együtthatók értelmezhetetlenségének veszélye a változók számának növelésével (a megfigyelések számának a változatlansága mellett) fokozódik.

Ha a változók páronként korrelálatlanok, akkor mindössze 2 m paramétert kell becsülni és értelmezni, továbbá a b regressziós súlyok egyértelműen meghatározhatók, értelmezhetők valamint a többszörös korreláció négyzet ekkor lesz maximális, és egyszerűen a függő és a magyarázó változók korrelációi négyzetének az összegére bontható. Ez így nagyon elegáns és főleg egyszerűsíti a problémát. A gyakorlatban azonban a korrelálatlanság feltevése ritkán teljesül.

Ha az eredeti változók felhasználásával elő tudnánk állítani olyan változókat, amelyek lényeges információvesztés nélkül írnák le (jellemeznék) az eredeti változókat, és ugyanakkor teljesülne rájuk a korrelálatlanság feltevése, akkor célszerűbb lenne ezeket használni, ezekkel leírni a vizsgált eseményeket.

A főkomponens-elemzés segítségével tudunk előállítani ilyen, az eredeti változókból lineáris transzformációval nyert latens (nem megfigyelt) változókat.

Főkomponensnek nevezzük azokat a latens változókat, amelyek

a) A megfigyelt változók lineáris kombinációi.

b) Korrelálatlanok, és mindegyik komponensre az együtthatók négyzetösszege 1 (normalizálási feltétel).

c) Az első komponens a maximális szórású, a második szórása maximális a maradékok között stb.

Az m változó n-szeri megfigyeléseit a $\mathbf{Z}$ mátrix tartalmazza ($n \times m$). Tegyük fel, hogy az m változó mindegyike standardizált, így $\mathbf{Z}'\mathbf{Z}/n = \mathbf{R}$ egy ($m \times m$) típusú korrelációmátrix. A standardizált változók olyan lineáris kombinációit keressük, amelyekre teljesül, hogy maximális szórásúak, korrelálatlanok, és a lineáris kombinációk együtthatóinak négyzetösszege 1.

Egy főkomponenst a következő alakban írhatunk fel:

$$\mathbf{y} = \mathbf{Z}\mathbf{q},$$

ahol $\mathbf{Z}$ ($n \times m$)-es megfigyelési mátrix (standardizált változókkal), $\mathbf{q}$ az együtthatók m elemű vektora, $\mathbf{y}$ a főkomponens n elemű vektora.

Az y főkomponens várható értéke:

$$\frac{\mathbf{1}'\mathbf{y}}{n} = \frac{\mathbf{1}'\mathbf{Z}\mathbf{q}}{n} = 0, \quad \text{mivel}$$

$$\frac{\mathbf{1}'\mathbf{Z}}{n} = \mathbf{0}', \text{ és szórásnégyzete}$$

$$\frac{\mathbf{y}'\mathbf{y}}{n} = \frac{(\mathbf{Z}\mathbf{q})'(\mathbf{Z}\mathbf{q})}{n} = \frac{\mathbf{q}'\mathbf{Z}'\mathbf{Z}\,\mathbf{q}}{n} = \mathbf{q}'\mathbf{R}\mathbf{q}.$$

Az első főkomponenst ($\mathbf{y}_1$) megkapjuk, ha megkeressük azokat az együtthatókat ($\mathbf{q}_1$), amelyek esetén $\mathbf{y}_1$ szórásnégyzete maximális lesz, és kielégíti a normalizálási feltételt $\mathbf{q}_1'\mathbf{q}_1 = 1$. Tehát maximalizálni akarjuk a $\mathbf{q}_1'\mathbf{R}\mathbf{q}_1$-et a $\mathbf{q}_1'\mathbf{q}_1 = 1$ feltétel mellett. A feladat megoldására a Lagrange-féle multiplikátor-módszert használhatjuk.

A

$$\phi_1 = \mathbf{q}_1'\mathbf{R}\mathbf{q}_1 - \alpha_1(\mathbf{q}_1'\mathbf{q}_1 - 1) \tag{12.1}$$

a Lagrange-függvény, ahol α_1 a Lagrange-féle multiplikátor.

A függvény maximuma ott lehet, ahol a derivált egyenlő 0-val.

$$\frac{\partial \phi_1}{\partial \mathbf{q}_1'} = 2\mathbf{R}\,\mathbf{q_1} - 2\alpha_1\mathbf{q}_1 = \mathbf{0}$$

osztva 2-vel nyerhetjük a következő alakot

$$(\mathbf{R} - \alpha_1\mathbf{I})\mathbf{q}_1 = \mathbf{0}. \tag{12.2}$$

Ennek az egyenletrendszernek ($\mathbf{q}_1$ ismeretlen) csak akkor van a triviálistól különböző megoldása, ha

$$|\mathbf{R} - \alpha_1\mathbf{I}| = 0, \tag{12.3}$$

azaz a determináns értéke egyenlő 0-val. Eszerint α_1 az $\mathbf{R}$ mátrix sajátértéke, a $\mathbf{q}_1$ pedig az $\mathbf{R}$ sajátvektora. Mivel a determináns kifejtésénél m-ed fokú polinomhoz jutunk, az egyenletnek m gyöke lesz. Ezek közül a legnagyobb sajátértéket választjuk, ezt jelöljük α_1-gyel, amit ha behelyettesítünk a (12.2)-be, megkapjuk az ehhez tartozó sajátvektort. Ha a (12.2) egyenletet átrendezzük, kapjuk

$$\mathbf{R}\mathbf{q}_1 - \mathbf{q}_1\alpha_1 = \mathbf{0},$$

$$\mathbf{R}\mathbf{q}_1 = \mathbf{q}_1\alpha_1.$$

Beszorozzuk balról $\mathbf{q}_1'$-gyel

$$\mathbf{q}_1'\mathbf{R}\mathbf{q}_1 = \mathbf{q}_1'\mathbf{q}_1\alpha_1 = \alpha_1,$$

ekkor azt kapjuk, hogy a legnagyobb sajátérték adja a $\mathbf{z}$ változók normalizált lineáris kombinációjának a maximális szórásnégyzetét.

Az első komponens megkeresése után olyan főkomponenst keresünk

$$\mathbf{y}_2 = \mathbf{Z}\mathbf{q}_2,$$

amely a maradékok közül maximális varianciájú és korrelálatlan $\mathbf{y}_1$-gyel, azaz:

$$\frac{\mathbf{y}_2'\mathbf{y}_1}{n} = \frac{(\mathbf{Z}\mathbf{q}_2)'(\mathbf{Z}\mathbf{q}_1)}{n} = \frac{\mathbf{q}_2'\mathbf{Z}'\mathbf{Z}\mathbf{q}_1}{n} = \mathbf{q}_2'\mathbf{R}\mathbf{q}_1 = \mathbf{q}_2'\mathbf{q}_1\alpha_1 = 0.$$

Ez csak akkor teljesül, ha $\mathbf{q}_2'\mathbf{q}_1 = 0$, ami azt jelenti, hogy két sajátvektor ortogonális (merőleges).

A Lagrange-féle módszert követve maximalizálni akarjuk:

$$\phi_2 = \mathbf{q}_2'\mathbf{R}\mathbf{q}_2 - \alpha_2(\mathbf{q}_2'\mathbf{q}_2 - 1) - \nu_1\mathbf{q}_2'\mathbf{R}\mathbf{q}_1,$$

$$\frac{\partial \phi_2}{\partial \mathbf{q}_2'} = 2\mathbf{R}\mathbf{q}_2 - 2\alpha_2\mathbf{q}_2 - \nu_1\mathbf{R}\mathbf{q}_1,$$

egyenlővé téve az utolsó egyenletet $\mathbf{0}$-val, osztva 2-vel és balról szorozva $\mathbf{q}_1'$-gyel mindkét oldalát, kapjuk:

$$\mathbf{q}_1'\mathbf{R}\mathbf{q}_2 - \alpha_2\mathbf{q}_1'\mathbf{q}_2 - \frac{\nu_1}{2}\mathbf{q}_1'\mathbf{R}\mathbf{q}_1 = \mathbf{0},$$

amelyet redukálunk:

$-\nu_1 \mathbf{q}_1' \mathbf{R} \mathbf{q}_1 = \mathbf{0}$, amelyből ν_1-nek 0-val kell egyenlőnek lennie, és a α_2 az $\mathbf{R}$ második sajátértéke, valamint $\mathbf{q}_2$ az ezzel összefüggő sajátvektor.

Ezt az eljárást követve meghatározhatjuk az m darab maximális α_j varianciájú komponenst:

$$\mathbf{y}_j = \mathbf{Z}\mathbf{q}_j \quad (j = 1, \ldots, m).$$

A α_j sajátértékekre fennáll[1] :

$$\sum_{j=1}^{m} \alpha_j = \text{trace}(\mathbf{R}) = m,$$

valamint

$$\prod_{j=1}^{m} \alpha_j = |\mathbf{R}|.$$

Ha a sajátértékeket egy $\mathbf{A}$ diagonális mátrix megfelelő diagonális elemeibe helyezzük, akkor

$$\mathbf{R}\,\mathbf{Q} = \mathbf{Q}\,\mathbf{A},$$

ahol

$$\mathbf{Q}'\mathbf{Q} = \mathbf{Q}\,\mathbf{Q}' = \mathbf{I}, \quad \text{és így}$$
$$\mathbf{Q}'\mathbf{R}\,\mathbf{Q} = \mathbf{A} = \mathbf{Y}'\mathbf{Y}/n,$$

és ez az $\mathbf{y}$ komponensek korrelációmátrixa, amelyből látszik, hogy teljesültek a komponensekre kikötött feltételek, hogy maximális szórásúak és páronként korrelálatlanok legyenek.

A főkomponens-elemzésnek egyik legérdekesebb felhasználási lehetősége, hogy míg az m komponens együtt pontosan reprodukálja a korrelációmátrixot:

$$\mathbf{R} = \mathbf{Q}\,\mathbf{A}\,\mathbf{Q}',$$

az első r komponens ($r < m$) a $\mathbf{z}$ változók varianciájának magasabb részét magyarázza, mint bármely más r ortogonális faktor.

Ez azért lényeges, mivel ha redukálni akarjuk a változóink számát m korrelált változóról ($\mathbf{z}$) $r < m$ korrelálatlan változóra ($\mathbf{y}$), akkor az első r főkomponens fogja a lehető legnagyobb részét magyarázni a $\mathbf{z}$ változók szórásnégyzetének.

Formailag:

$$\mathbf{R} = \mathbf{Q}\,\mathbf{A}\,\mathbf{Q}' = \alpha_1 \mathbf{q}_1 \mathbf{q}_1' + \alpha_2 \mathbf{q}_2 \mathbf{q}_2' + \ldots + \alpha_m \mathbf{q}_m \mathbf{q}_m' = \mathbf{R}_1 + \mathbf{R}_2 + \ldots + \mathbf{R}_m,$$

vagyis az m főkomponens teljesen megmagyarázza a korrelációmátrixot.

Amikor az első r komponenst választjuk ki, akkor a korrelációs mátrixot két részre bontjuk

$$\mathbf{R} = \widehat{\mathbf{R}} + \widetilde{\mathbf{R}},$$

ahol

$$\widehat{\mathbf{R}} = \mathbf{R}_1 + \mathbf{R}_2 + \ldots + \mathbf{R}_r$$

ún. *reprodukált korrelációmátrix* (az elnevezés onnan ered, hogy ha az első r főkomponens teljesen megmagyarázza x változók szórását, így akkor visszakapjuk $\mathbf{R}$-et ($\mathbf{R}_i$) jelenti az i-edik komponens hozzájárulását a reprodukált mátrixhoz), és

$$\widetilde{\mathbf{R}} = \mathbf{R} - \widehat{\mathbf{R}} = \mathbf{R}_{r+1} + \mathbf{R}_{r+2} + \ldots + \mathbf{R}_m$$

[1] A trace ($\mathbf{R}$) az $\mathbf{R}$ mátrix nyomát jelenti. Egy kvadratikus mátrix nyomán a mátrix diagonális elemeinek összegét értjük.

reziduális korrelációmátrix.

A főkomponens-elemzés igénye leggyakrabban abban az esetben merül fel, ha megfigyelt változók között erős korrelációt észlelünk. Túlzott óvatosságnak tűnik, mégis érdemes elvégezni az $\mathbf{R}$ mátrix elemeinek szignifikancia-vizsgálatát. *A Bartlett-féle gömbölyűség-próbát* használhatjuk a szignifikancia ellenőrzésére.

A nullhipotézis az, hogy a megfigyelt változók korrelációmátrixa egységmátrix $\mathbf{R} = \mathbf{I}$ (azaz a változók páronként korrelálatlanok), ami ekvivalens azzal az állítással, hogy $|\mathbf{R}| = 1$.

A próba elnevezése onnan ered, hogy a standardizált korrelálatlan változók a térben kör, illetve gömb alakot öltenek.

A χ^2 eloszlást követő valószínűségi változó definíciója (a próba kritériuma):

$$\chi^2_{0,5(m^2-m)} = -\left[(n-1)\frac{1}{6}(2m+5)\right]\ln|\mathbf{R}|$$

Ha a hipotézist elvetjük, tehát χ^2 abszolút értéke nagyobb mint az elméleti χ^2_r, akkor az eredeti változóinkat korreláltaknak tekintjük.

A próba erejét Knapp és Swoyer (1967) vizsgálta meg, és elég nagynak találták. Ha $n = 20$, $m = 10$ és $p = 0{,}05$, gyakorlatilag visszautasíthatjuk a H_0 hipotézist, amikor z_i változók közötti páronkénti korreláció 0,36 vagy több, és $n = 200$, $m = 10$ és $p = 0{,}05$ esetén a H_0 hipotézist akkor vethetjük el, ha z_i változók közötti korreláció 0,09 vagy több.

A Bartlett-féle gömbölyűség-próbát használjuk abban az esetben is, amikor $r < m$, és így $\mathbf{R} = \widehat{\mathbf{R}} + \widetilde{\mathbf{R}}$.

A nullhipotézis az, hogy a reziduális korrelációmátrix ($\widetilde{\mathbf{R}}$) $\mathbf{0}$-val egyenlő.

$$\chi^2_{0,05(m-r)(m-r-1)} = -\left[(n-1)\frac{1}{6}(2m+5) - \frac{2}{3}r\right]\ln\chi_{m-r},$$

ahol

$$\chi_{m-r} = \frac{|\mathbf{R}|}{\prod_{j=1}^{r}\alpha\left[\left(m-\sum_{j=1}^{r}\alpha_j\right)/(m-r)\right]^{m-r}}.$$

A főkomponens-modell illeszkedésének jóságát mérhetjük a Kaiser–Meyer–Olkin-féle KMO-mutatóval (Kaiser, 1974), amely a megfigyelt változók korrelációs együtthatóinak nagyságát veti össze a parciális korrelációs együtthatók nagyságával:

$$\mathrm{KMO} = \frac{\sum_{\substack{i\\ i\neq j}}\sum_{j} r_{ij}^2}{\sum_{\substack{i\\ i\neq j}}\sum_{j} r_{ij}^2 + \sum_{\substack{i\\ i\neq j}}\sum_{j} a_{ij}^2},$$

ahol r_{ij} az i-edik és j-edik megfigyelt változó közötti korrelációs együttható, a_{ij} az i-edik és j-edik megfigyelt változók közötti parciális korrelációs együttható (a parciális korrelációt a megfigyelt változók főkomponensekkel nem magyarázott részei közötti korrelációval becsüljük).

A KMO-mutató értéktartománya 0 és 1 között van. Közelebb esik 1-hez akkor, amikor a főkomponens modell illeszkedik az adatokhoz, mivel a lineáris hatás kiszűrésével a parciális korrelációk kicsik lesznek.

Kaiser kategorizálta a KMO mértéket:

$0{,}9 \leq$ KMO	nagyon jó,
$0{,}8 \leq$ KMO $< 0{,}9$	jó,
$0{,}7 \leq$ KMO $< 0{,}8$	közepes,
$0{,}6 \leq$ KMO $< 0{,}7$	gyenge,
$0{,}5 \leq$ KMO $< 0{,}6$	rossz,
KMO $< 0{,}5$	elfogadhatatlan.

A KMO-minta alkalmassági mutatóját számíthatjuk változónként is.

Az i-edik változó minta alkalmassági mutatója (measure of sampling adequacy, MSA):

$$MSA_i = \frac{\sum\limits_{j \neq i} r_{ij}^2}{\sum\limits_{j \neq i} r_{ij}^2 + \sum\limits_{j \neq i} a_{ij}^2}.$$

Faktorok mint latens változók

A főkomponenseket a $\mathbf{z}$ változók transzformációjával (lineáris függvényével) állítottuk elő:

$$\mathbf{Y} = \mathbf{Z}\,\mathbf{Q},$$

amelynek elemei (az egyes főkomponensek) 0 várható értékűek és $\sqrt{\alpha_j}$ szórásúak. Ha standardizáljuk az $\mathbf{y_j}$ komponenseket, vagyis a $\mathbf{z}$ standardizált változók standardizált lineáris transzformációját vesszük, akkor a faktorokhoz ($\mathbf{f}_j$) jutunk:

$$\mathbf{f}_j = \mathbf{y}_j \,/\sqrt{\alpha_j}$$

és az összes faktort tartalmazó mátrix:

$$\mathbf{F} = \mathbf{Y}\,\mathbf{A}^{-\frac{1}{2}} = \mathbf{Z}\,\mathbf{Q}\,\mathbf{A}^{-\frac{1}{2}} = \mathbf{Z}\,\mathbf{B}.$$

Az előzőekben láttuk, hogy a főkomponens-elemzés módszert ad a változók számának csökkentésére, vagyis a főkomponens-elemzés olyan általánosan használt eljárás, amely a független változók minimális számának meghatározására használható úgy, hogy az eredeti változók szórásnégyzetének legnagyobb részét megmagyarázzák a főkomponensek.

Ezeket az $\mathbf{f}_j$ faktorokat nem megfigyelt, latens változóknak is hívjuk, mivel értékeit nem közvetlenül figyeltük meg, hanem a megfigyelt változók lineáris kombinációival állítottuk elő, ahol az együtthatókat a $\mathbf{B}$ faktor-együtthatómátrix tartalmazza.

Talán a legérdekesebb számunkra a faktorok értelmezése érdekében az eredeti változók ($\mathbf{z}$) és a faktorok közötti korreláció. A faktorok és változók közötti korrelációt tartalmazó mátrixot *faktorstruktúrának* nevezzük. A faktorstruktúrában a λ_{jk} adja a j-edik változó és a k-adik faktor korrelációját. A struktúra k-adik oszlopa segít értelmezni, elnevezni a k-adik faktort, míg a j-edik sora a faktorok hozzájárulását adja az j-edik változó szórásnégyzetéhez.

Ha a faktorstruktúrát a $\mathbf{S}$ mátrix jelöli, akkor

$$\mathbf{S} = \frac{\mathbf{Z}\,\mathbf{F}}{n} = \frac{\mathbf{Z}'\mathbf{Z}\,\mathbf{Q}\,\mathbf{A}^{-\frac{1}{2}}}{n} = \mathbf{R}\,\mathbf{Q}\,\mathbf{A}^{-\frac{1}{2}},$$

és mivel $\mathbf{R}\,\mathbf{Q} = \mathbf{Q}\,\mathbf{A}$

$$\mathbf{S} = \mathbf{Q}\,\mathbf{A}\,\mathbf{A}^{-\frac{1}{2}} = \mathbf{Q}\,\mathbf{A}^{\frac{1}{2}}.$$

Egy másik értelmezésre ad lehetőséget, ha az eredeti változókat fejezzük ki a faktorok lineáris kombinációiként.

Induljunk ki az

$$\mathbf{F} = \mathbf{Z}\,\mathbf{B} = \mathbf{Z}\,\mathbf{Q}\,\mathbf{A}^{-\frac{1}{2}}$$

egyenletből.

Szorozzuk mindkét oldalt jobbról $\mathbf{A}$-val:

$$\mathbf{F}\,\mathbf{A} = \mathbf{Z}\,\mathbf{Q}\,\mathbf{A}^{\frac{1}{2}} = \mathbf{Z}\,\mathbf{S},$$

mivel

$$\mathbf{S}'\mathbf{S} = \left(\mathbf{Q}\,\mathbf{A}^{\frac{1}{2}}\right)'\left(\mathbf{Q}\,\mathbf{A}^{\frac{1}{2}}\right) = \mathbf{A}^{\frac{1}{2}}\mathbf{Q}'\mathbf{Q}\,\mathbf{A}^{\frac{1}{2}} = \mathbf{A},$$

így

$$\mathbf{F}\,\mathbf{A} = \mathbf{F}\,\mathbf{S}'\mathbf{S} = \mathbf{Z}\,\mathbf{S},$$

amiből

$$(\mathbf{Z} - \mathbf{F}\,\mathbf{S}')\mathbf{S} = \mathbf{0},$$

ami csak úgy állhat fenn, ha

$$\mathbf{Z} = \mathbf{F}\,\mathbf{S}'.$$

Az $\mathbf{Z}$ mátrix i-edik oszlopvektora az i-edik változó megfigyelt és standardizált értékeit reprezentálja:

$$\mathbf{z}_i = s_{i1}\mathbf{f}_1 + s_{i2}\mathbf{f}_2 + \ldots + s_{im}\mathbf{f}_m,$$

a $\mathbf{z}_i$ változó szórásnégyzete:

$$\frac{\mathbf{z}_i'\mathbf{z}_i}{n} = \frac{s_{i1}^2\mathbf{f}_1'\mathbf{f}_1}{n} + \frac{s_{i2}^2\mathbf{f}_2'\mathbf{f}_2}{n} + \ldots + \frac{s_{im}^2\mathbf{f}_m'\mathbf{f}_m}{n}.$$

Mivel a faktorok páronként függetlenek és szórásuk 1:

$$1 = s_{i1}^2 + s_{i2}^2 + \ldots + s_{im}^2,$$

amely azt mutatja, hogy az $\mathbf{S}$ i-edik sorában lévő faktorsúlyok négyzeteinek összege a z_i változó varianciáját (szórásnégyzetét) adja. Ha az $\mathbf{S}$ mátrix elemeinek a négyzetét rendre összeadjuk, akkor megkapjuk a faktorok teljes hozzájárulását a $\mathbf{z}$ változók teljes szórásnégyzetéhez (varianciájához), m-hez.

Vegyük a megfigyelt változók korrelációmátrixát:

$$\mathbf{R} = \frac{\mathbf{Z}'\mathbf{Z}}{n} = \frac{(\mathbf{F}\,\mathbf{S}')'\mathbf{F}\,\mathbf{S}'}{n} = \frac{\mathbf{S}\,\mathbf{F}'\mathbf{F}\,\mathbf{S}'}{n} = \mathbf{S}\,\mathbf{S}'.$$

Az $\mathbf{R}$ felbontható m mátrix összegére:

$$\mathbf{R} = \mathbf{s}_1\mathbf{s}_1' + \mathbf{s}_2\mathbf{s}_2' + \ldots + \mathbf{s}_m\mathbf{s}_m'.$$

Az $\mathbf{s}_i\mathbf{s}_i'$ mátrix adja az i-edik faktor hozzájárulását az $\mathbf{R}$ korrelációmátrixhoz, vagyis az $\mathbf{s}_i\mathbf{s}_i'$ a korrelációmátrix i-edik faktor által magyarázott részét jelenti.

Az α_i-t gyakran kifejezik a változók teljes szórásnégyzetének (m) százalékában. Ez a százalék a teljes varianciának azt az arányát adja, amelyet az i-edik faktor magyaráz. Ezt szokták nevezni a *relatív faktor-hozzájárulásnak*.

Ha $\mathbf{S}$ teljes faktormátrix, akkor az m komponens mindegyikének nem 0-a a saját értéke (szórásnégyzete), így $\mathbf{S}$ sorelemei négyzetének összege kiadja az 1-et.

Azonban, ha $\mathbf{S}$-ből hiányoznak faktorok ($m - r$), amelyeket elhanyagoltunk, akkor $\mathbf{S}_r$ soraiban az elemek négyzetösszege kevesebb lesz mint egy. Ezt az összeget hívjuk a változók *kommunalitásának*:

$$h_i^2 = \sum_{j=1}^{r} s_{ij}^2.$$

A h_i^2 tehát az i-edik változó szórásnégyzetének a faktorok által együttesen magyarázott része.

A faktorelemzésnél az a cél, hogy maximalizáljuk a kommunalitásokat, miközben minimalizáljuk a faktorok számát, más szóval a faktorok minimális számát akarjuk meghatározni, amellyel az eredeti változók szórásnégyzetének maximális hányadát tudjuk magyarázni.

Láthattuk, hogy annyi faktort találhatunk, amennyi az $\mathbf{R}$ nullától különböző sajátértékeinek a száma, és azok mindegyike a lehető legjobban járul hozzá a teljes szórásnégyzethez. A cél az, hogy minél kevesebb legyen a faktorok száma. Mivel a sajátértékek csökkenő sorozatot alkotnak, egy r sorszámtól kezdve viszonylag kicsi sajátértékeket kapunk. Ezen faktorok hozzájárulása a teljes varianciához már kicsi, így ezeket elhanyagolhatjuk.

Mit jelent azonban az, hogy kicsi? Használhatunk statisztikai próbát annak eldöntésére, hogy egy faktort elhagyhatunk-e vagy sem, vagy alkalmazhatunk néhány szemléletesebb „szabályt" (ami természetesen nem pótolja a statisztikai próbát).

A sajátértékeket sorrendben rajzoljuk be egy koordináta-rendszerbe. Így egy csökkenő görbéhez jutunk. Ha a sajátérték kisebb mint egy, akkor elhanyagoljuk (mivel a faktor kevesebbet magyaráz, mint egy megfigyelt változó). Ha a görbe hirtelen esést mutat néhány magas sajátérték után, és az eséstől jobbra már csak nagyon lassan süllyed, akkor ezeket a sajátértékeket elhanyagolhatjuk.

A faktorok számának (r) meghatározásához számításba vehetjük a következő négy kritériumot:

1. Az r értékét úgy határozzuk meg, hogy ezzel a faktorszámmal elérjünk egy előre meghatározott küszöbértéket, mondjuk 80%-ot, amilyen százalékos arányban magyarázni tudjuk az összvarianciát.

2. Az r értéke egyenlő az átlagos sajátértéknél nagyobb sajátértékek számával.

3. A sajátértékek diagramjának tanulmányozása alapján: ha a diagram meredek esését egy enyhébb lejtő követi, akkor az enyhe lejtőhöz közelítő vonal előtti sajátértékek száma adja a faktorok számát.

4. Annak a hipotézisnek a tesztelésével, hogy r a faktorok megfelelő száma:

$$H_0 : \mathbf{\Sigma} = \mathbf{\Lambda}\,\Lambda' + \mathbf{\Theta}.$$

A tesztstatisztika:

$$\left(n - \frac{2m + 4mr + 11}{6}\right) \ln\left(\frac{\widehat{\mathbf{\Lambda}}\,\widehat{\mathbf{\Lambda}}' + \Theta}{|\mathbf{R}|}\right),$$

közelítően χ^2 eloszlású, ha H_0 hipotézis nem vethető el (a szabadságfok: $\nu = \frac{1}{2}[(m - - r)^2 - m - r]$).

Határozzuk meg a faktorok számát a következő két példában (Alvin C. Rencher, 1995)

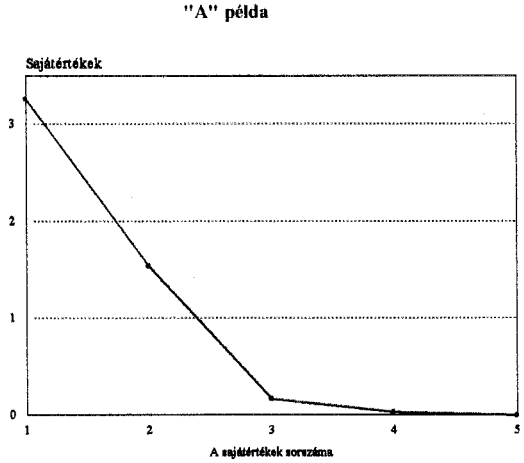
– Az „A" példában az $\mathbf{R}$ mátrix szinguláris.

– a „B" példában $r = 4$ esetén a tesztstatisztika értéke:

$$\chi^2 = 9{,}039$$

Az elméleti érték (a szabadságfok = 11):

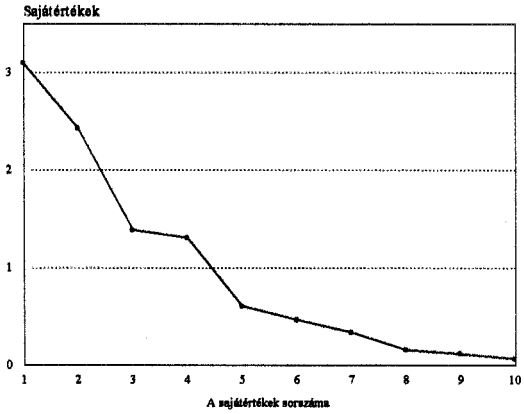
$$\chi^2_{0{,}05,11} = 19{,}68$$



12.1. ábra. A sajátértékek diagramja

„A” példa

Sajátértékek:	3,263	1,538	0,168	0,031	0.
Arány:	0,6526	0,3076	0,0336	0,0062	0.
Kumulatív:	0,6526	0,9602	0,9938	1,0	



12.2. ábra. A sajátérték diagramja

„B” példa

Sajátértékek:	3,10	2,43	1,39	1,31	0,61	0,47	0,34	0,16	0,12	0,07
Arány:	0,31	0,243	0,139	0,131	0,061	0,047	0,034	0,016	0,012	0,007
Kumulatív:	0,31	0,553	0,692	0,823	0,884	0,931	0,965	0,981	0,993	1,0

Az „A” példában az első kritérium szerint *r* értéke (a faktorok száma) 2, mivel az első sajátérték a korrelációmátrix nyomának 65%-át reprodukálja, míg az első kettő

együttesen 96%-át magyarázza az összvarianciának. A második kritérium szerint a faktorok száma szintén 2, mivel a második sajátérték 1,54, és a harmadik 0,17. Az „A” példa sajátértékeinek ábráját vizsgálva a harmadik kritériumnak megfelelő faktorszám szintén 2. A negyedik kritérium nem alkalmazható ebben az esetben, mivel a korrelációmátrix szinguláris. Három kritérium szerint a faktorok száma: $r = 2$.

A „B” példában az első négy sajátérték együttesen 82%-át reprodukálja a $tr(\mathbf{R})$-nek. A második kritérium szerint $r = 4$, mivel $\alpha_4 = 1{,}31$ és $\alpha_5 = 0{,}61$. A harmadik kritérium szerint a „B” példa sajátérték diagramját tanulmányozva $r = 4$, mivel az ötödik sajátértéktől lanyhul a diagram esése. A negyedik kritérium szerint a khi négyzet értéke $r = 4$ esetén:

$$\chi^2 = 9{,}039$$

A szabadságfok: $v = 1/2[(m - r)^2 - m - r] = 1/2[(10 - 4)^2 - 10 - 4)] = 11$.

Mivel az elméleti érték (a szabadságfok = 11):

$$\chi^2_{0{,}05,\ 11} = 19{,}68$$

az empirikus érték kisebb az elméleti értéknél, a nullhipotézist, miszerint a négy faktoros modell illeszkedik az adatokhoz, nem utasíthatjuk el. A „B” példában a négy kritérium mindegyike az $r = 4$ megoldást javasolta.

A faktorok értelmezése

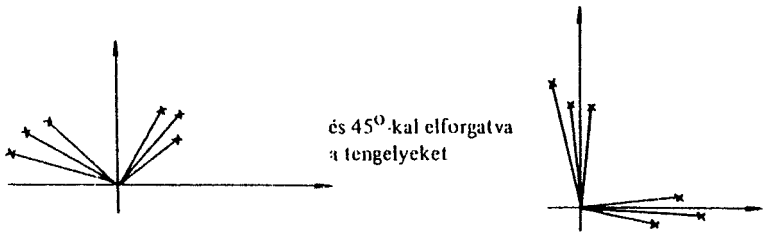
A változók – mint láttuk – felírhatók a faktorok függvényében:

$$\mathbf{Z} = \mathbf{F}\mathbf{S}'$$

A faktorokat a hozzájuk tartozó együtthatók, a faktorsúlyok ($\mathbf{S}'$ elemei), azaz a faktorstruktúra alapján értelmezhetjük.

Az értelmezés akkor könnyű, amikor a faktorok alapján elkülöníthető csoportokra tudjuk osztani a változókat. Ha ez nincs így, akkor nagyon nehezen, vagy nem értelmezhetők a faktorok. Ilyenkor a faktorsúlyok ortogonális transzformációjával próbálkozhatunk.

Ez geometriailag azt jelenti, hogy új koordinátatengelyekre térünk át, vagyis elforgatjuk bizonyos szöggel a tengelyeket. Pl.



Ha $\mathbf{T}$ az ortogonális transzformáció mátrixa, azaz fennáll $\mathbf{T}\mathbf{T}' = \mathbf{I}$, akkor az új faktorstruktúra

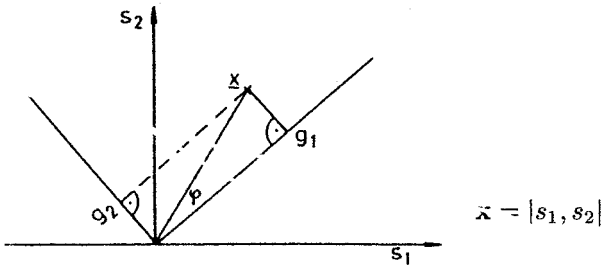
$$\mathbf{G} = \mathbf{S}\mathbf{T}.$$

A transzformáció nem változtatja meg a kommunalitásokat:

$$\mathbf{G}\mathbf{G}' = (\mathbf{S}\mathbf{T})(\mathbf{S}\mathbf{T})' = \mathbf{S}\mathbf{T}\mathbf{T}'\mathbf{S}' = \mathbf{S}\mathbf{S}'$$

A transzformáció elvégzésére több módszert dolgoztak ki. Ezeknek egy része geometriailag, más része analitikus úton határozza meg a transzformációs mátrixot.

Mivel a faktorok akkor értelmezhetők jól, ha a faktorsúlyok között csak viszonylag nagy és kicsi abszolút értékek találhatók, a tengelyeket úgy kell elforgatni, hogy a transzformált faktorsúlyokra ez teljesüljön. Két faktor esetén ezt a következő ábra szemlélteti.



$$\mathbf{x} = |s_1, s_2|$$

Az elforgatásnál tehát azt akarjuk elérni, hogy az **x** vektornak egyik tengelyre maximális, a másikra minimális vetülete essék.

A két vetületet (egyben **x** vektor új koordinátáit) a következőképpen kaphatjuk:

$$g_1 = h \cos \varphi \tag{12.4}$$

$$g_2 = h \sin \varphi$$

ahol h a vektor hossza (a $h^2 = s_1^2 + s_2^2$, azaz a kommunalitással egyenlő).

A (12.4) egyenletből

$$g_1 g_2 = h^2 \cos \varphi \sin \varphi = \frac{1}{2} h^2 \sin 2\varphi.$$

Ha φ kicsi, akkor $\sin 2\varphi$ is kicsi lesz, és mivel h^2 konstans, a $g_1 g_2$ értéke is kicsi lesz.

A középiskolában tanult Pitagorasz-tétel szerint

$$g_1^2 + g_2^2 = h^2$$

így

$$h^4 = (g_1^2 + g_2^2)^2 = g_1^4 + g_2^4 + 2g_1^2 g_2^2 = g_1^4 + g_2^4 + \frac{1}{2} h^4 \sin^2 2\varphi.$$

Mivel h^4 konstans, ha $\sin 2\varphi$ csökken, akkor $(g_1^4 + g_2^4)$ értékének növekedni kell.

Így megtaláltuk a transzformáció egy lehetséges kritériumát. Keressük azt a **T** transzformációs mátrixot, amely mellett **G** (**G** = **S T**) mátrix elemei negyedik hatványának összege maximális. Ezt a kritériumot nevezik quartimax-módszernek:

$$\sum_j^p \sum_i^n g_{ij}^4 \longrightarrow \max .$$

Egy másik analitikus megközelítés, amelyet a leggyakrabban használnak, a *varimax-megoldás.*

A varimax-kritérium a g^2 szórásának maximalizálása, vagyis

$$\sigma^2 = \sum_j^p \sigma_j^2 = \frac{1}{m} \sum_j^p \sum_i^m g_{ij}^4 - \frac{1}{m} \sum_j^p \left(\sum_i^m g_{ij}^2 \right)^2$$

szórásnégyzet összeg maximalizálása.

A faktorok transzformációja nem változtatja meg a faktorok korrelálatlanságát, valamint kommunalitását.

A faktorok becslése

Az $\mathbf{F}$ mátrixot úgy vezettük be, mint a faktorok mátrixát. A faktorokat a megfigyelt változókkal állítottuk elő. Ha a faktorok száma m, akkor a faktorok egyértelműen előállíthatók a következő formulával:

$$\mathbf{F} = \mathbf{Z}\,\mathbf{Q}\,\mathbf{A}^{-\frac{1}{2}}.$$

Azonban, ha az $\mathbf{Z} = \mathbf{F}\,\mathbf{S}'$ modell helyett az $\mathbf{Z} = \mathbf{F}\,\mathbf{S}' + \mathbf{U}$ modellt alkalmazzuk, ahol $\mathbf{U}$ $(n \times m)$-es mátrix, a véletlen komponenseket tartalmazza, azaz a faktorok számát redukáltuk ($r < m$), a faktorokat az egyértelmű előállítás helyett csak becsülhetjük. Így tulajdonképpen egy többváltozós regressziós problémához jutottunk. Ha ismerjük a változók és a faktorok közötti korrelációkat, a regressziós együtthatókat a következő formula adja:

$$\mathbf{W} = \mathbf{R}^{-1}\mathbf{S}.$$

Ezt felhasználva a faktorok becslése:

$$\mathbf{F} = \mathbf{Z}\,\mathbf{R}^{-1}\mathbf{S}$$

és a faktorok szórásnégyzete:

$$\mathbf{F}'\mathbf{F} = \frac{(\mathbf{Z}\,\mathbf{R}^{-1}\mathbf{S})'(\mathbf{Z}\,\mathbf{R}^{-1}\mathbf{S})}{n} = \frac{\mathbf{S}'\mathbf{R}^{-1}\mathbf{Z}'\mathbf{Z}\,\mathbf{R}^{-1}\mathbf{S}}{n} = \mathbf{S}'\mathbf{R}^{-1}\mathbf{S}.$$

Amennyiben $r < m$, a faktorok száma kisebb, mint a változók száma, $\mathbf{R}$ mátrix helyett a reprodukált korrelációmátrixot használhatjuk a faktorpontok számításához.

Az $\mathbf{Z} = \mathbf{F}\,\mathbf{S}'$ alapegyenletből kiindulva, két lépéssel juthatnak a faktorpontokhoz:

$$\mathbf{Z}\,\mathbf{S} = \mathbf{F}\,\mathbf{S}'\mathbf{S}$$
$$\mathbf{F} = \mathbf{Z}\,\mathbf{S}(\mathbf{S}'\mathbf{S})^{-1}$$

12.2. Főfaktorok módszere

A faktorelemzés abból a klasszikus méréselméleti feltételezésből indul ki, hogy a megfigyelt változók kifejezhetők két komponens, a szisztematikus vagy közös- és a hibakomponens lineáris függvényével. Szimbolikusan:

$$\mathbf{Z} = \mathbf{C} + \mathbf{E}, \tag{12.5}$$

ahol $\mathbf{Z}$ a megfigyelt változók standardizált mátrixa ($(n \times m)$ típusú, ahol n a megfigyelési egységek, m a változók száma), $\mathbf{C}$ a közös komponensek mátrixa ($n \times m$ típusú), $\mathbf{E}$ a hibakomponensek mátrixa ($n \times m$ típusú).

Feltételezzük, hogy a két komponens korrelálatlan egymással ($\mathbf{C}'\mathbf{E} = \mathbf{E}'\mathbf{C} = \mathbf{0}$), valamint hogy a hibakomponensek függetlenek. Ez utóbbi feltételezésből következik, hogy a hibák variancia-kovarianciamátrixa diagonális ($\mathbf{E}'\mathbf{E}/n = \mathbf{U}^2$).

A közös komponensek $\mathbf{C}$ mátrixát a faktorok lineáris kombinációjával fejezzük ki:

$$\mathbf{C} = \mathbf{F}\,\mathbf{\Lambda}', \tag{12.6}$$

ahol $\mathbf{F}$ a faktorértékek mátrixa ($n \times r$ típusú, ahol n a megfigyelések száma, r a közös faktorok száma), Λ a faktorsúlyok mátrixa ($m \times r$ típusú).

A harmadik feltételként általában előírjuk, hogy a faktorok legyenek függetlenek ($\mathbf{F}'\mathbf{F}/n = \mathbf{I}$). Ez a feltétel azonban nem tartozik a modell alapfeltételeihez, így sokszor megengedjük, hogy a faktorok korreláljanak egymással. Ekkor $\mathbf{F}'\mathbf{F}/n = \Phi$. A továbbiakban azonban az $\mathbf{F}'\mathbf{F}/n = \mathbf{I}$ feltételt tartjuk érvényesnek. A faktormodell három feltétele alapján a korrelációs mátrixot a következőképpen bonthatjuk fel:

$$\begin{aligned} \mathbf{R} &= \mathbf{Z}'\mathbf{Z}/n \\ &= \mathbf{C}'\mathbf{C}/n + \mathbf{U}^2 \\ &= (\mathbf{F}\,\Lambda')'(\mathbf{F}\,\Lambda')/n + \mathbf{U}^2 \\ &= \Lambda\,\Lambda' + \mathbf{U}^2, \end{aligned} \tag{12.7}$$

ahol $\mathbf{R}$ a megfigyelt változók páronkénti korrelációs együtthatóit tartalmazó mátrix ($m \times m$ típusú).

Thurstone (1947) a (12.7) egyenletet a faktorelemzés alapegyenletének nevezte.

Láthatjuk, hogy a megfigyelt változók közötti korrelációkat reprodukálni tudjuk a faktorsúlyokkal, és a változók varianciáinak (a diagonális elemeknek) a faktorokkal nem magyarázott része pedig a hibavarianciákkal egyenlő. Az egyes változók varianciájának a közös komponensekkel (faktorokkal) megmagyarázott részét kommunalitásnak nevezik. Rendezzük a kommunalitásokat a $\mathbf{H}^2$ diagonális mátrix megfelelő elemeibe, ekkor:

$$\mathbf{H}^2 = \mathbf{I} - \mathbf{U}^2, \tag{12.8}$$

vagy

$$\mathbf{U}^2 = \mathbf{I} - \mathbf{H}^2.$$

A faktorelemzésnél a korrelációmátrixokból indulunk ki, és keressük a faktorsúlyok mátrixát. Ha az $\mathbf{U}^2$ mátrix ismert, akkor a Λ mátrixot az $\mathbf{R} - \mathbf{U}^2 = \Lambda\,\Lambda'$ egyenletből kiindulva határozzuk meg. Matematikailag a feladat egy sajátérték-sajátvektor problémára vezethető vissza (Rummel, 1970). Az egyenlet:

$$(\mathbf{R} - \mathbf{U}^2)\mathbf{Q} = \mathbf{A}\,\mathbf{Q}. \tag{12.9}$$

A $\mathbf{Q}$ jelöli $\mathbf{R} - \mathbf{U}^2$ sajátvektorát, és $\mathbf{A}$ a megfelelő sajátértékeket, ezek alapján a faktorsúlyok mátrixa: $\Lambda = \mathbf{Q}\,\mathbf{A}^{1/2}$. A faktorsúlyokat mint a megfigyelt változók ($\mathbf{Z}$) és a nem megfigyelt faktorok ($\mathbf{F}$) közötti korrelációs együtthatókat értelmezhetjük, ha feltesszük a faktorok korrelálatlanságát, továbbá igaz, hogy $\Lambda\,\Lambda'$ egyenlő a sajátértékek diagonális mátrixával. Ez azt jelenti, hogy egy adott faktor súlyainak a négyzetösszege (a sajátérték) az adott faktor hozzájárulását adja az összvarianciához (a sajátértékek összege egyenlő a megfigyelt változók korrelációmátrixának a nyomával – trace ($\mathbf{R}$) –, és ez egyenlő a változók számával). A faktorok fontosságát szokás ezért a sajátértékek relatív, az összvarianciához viszonyított mértékével is kifejezni.

A (12.8) azonosságra visszatérve láthatjuk, hogy

$$\mathbf{R} - \mathbf{U}^2 = \mathbf{R} - \mathbf{I} + \mathbf{H}^2 = \Lambda\,\Lambda',$$

vagyis a faktorsúlyok mátrixának és saját transzponáltjának szorzata egyenlő a korrelációmátrixszal, ha a diagonális elemeket kicseréljük a kommunalitásokkal.

Az előzőekben feltételeztük, hogy az $\mathbf{U}^2$ mátrix ismert. A gyakorlatban a hibavarianciákat nem ismerjük, hanem becsüljük őket. Attól függően, hogy a kommunalitásokat milyen módon becsüljük, különböző faktorelemző eljárásokat különböztetünk meg.

Főkomponens-elemzés

Főkomponens-elemzésnél elkerüljük a kommunalitások problémáját, és $\mathbf{R} - \mathbf{U}^2$ dekompozíciója helyett az $\mathbf{R}$ mátrixot magyarázzuk a faktorsúlyokkal. A legfőbb jellemzője a főkomponenseknek, hogy mindegyik komponens a lehető legnagyobb mértékben járul hozzá a megfigyelt változók varianciájához. A faktorsúlyokat a következő homogén egyenlet megoldása alapján kapjuk:

$$(\mathbf{R} - \alpha\mathbf{I})\mathbf{q} = \mathbf{0}, \tag{12.10}$$

mátrix sajátvektorát pedig $\mathbf{q}$ jelöli.

Az első sajátértékhez tartozó faktor magyarázza a legnagyobb részét a megfigyelt változók varianciájának, a sorrendben következő faktorok csökkenő mértékben járulnak hozzá az összvarianciához.

A faktorsúlyokat a sajátértékek és a sajátvektorok alapján számítjuk:

$$\mathbf{\Lambda} = \mathbf{Q}\,\mathbf{A}^{1/2}. \tag{12.11}$$

A változók számával megegyező számú főkomponens pontosan reprodukálja a korrelációmátrixot (a diagonális elemeket is). A gyakorlatban azonban az 1-nél kisebb sajátértékekhez tartozó főkomponenseket elhagyjuk, így a reprodukált korrelációmátrix általában csak közelíti a megfigyelt korrelációmátrixot, de nem lesz egyenlő vele. Az r számú főkomponenshez ($r < m$) az első r legnagyobb sajátérték tartozik, így az első r számú faktor a lehető legnagyobb, de csökkenő mértékben járul hozzá a változók kommunalitásához.

A főkomponens-elemzés alkalmazása akkor indokolt, ha a megfigyelt változók száma nagy, és az első főkomponensekhez tartozó sajátértékek kiugróan magasak.

Főfaktorok módszere

A faktorelemzésnél széles körben használatos az a megoldás, amelyet a főfaktorok módszerének neveznek. Ez abban különbözik a klasszikus főkomponens-elemzéstől, hogy a korrelációmátrix diagonális elemeit kicseréljük a kommunalitások becsléseivel (ezek legtöbbször az egyes változók többi változóra vonatkozó többszörös korrelációs együtthatók közül a maximálisaknak az abszolút értékei), és az így redukált korrelációmátrixra alkalmazzuk a főkomponens-elemzés módszerét. Ennek a módszernek van iteratív változata, amikor a kommunalitásokat az iteráció egyes lépéseiben kicseréljük az adott lépésbeli becslésekkel:

$$\mathbf{U}^2 = \text{diag}\,(\mathbf{R} - \mathbf{\Lambda}\,\mathbf{\Lambda}'),$$

és addig folytatjuk az eljárást, amíg a kommunalitások becslései az egyes lépések során már csak elhanyagolható mértékben változnak. Könnyen megmutatható, hogy ez az eljárás ugyanahhoz a faktorsúly-mátrixhoz vezet, amelyet az $(\mathbf{R} - \mathbf{\Lambda}\,\mathbf{\Lambda}')$ reziduális mátrix diagonálison kívüli elemei négyzetösszegének minimalizálásával kapnánk.

12.3. Image-elemzés

Tekintsünk m megfigyelt változót, és jelölje z_j ($j = 1, 2, \ldots, m$) a j-edik megfigyelt változót. Tételezzük fel, hogy a megfigyelt változók standardizáltak, vagyis mindegyik várható értéke 0 és szórása 1.

Becsüljük a megfigyelt változók mindegyikét a többi $(m-1)$ változóval (a legkisebb négyzetek módszerével):

$$p_j = \sum_{k=1}^{m} w_{jk} z_k, \tag{12.12}$$

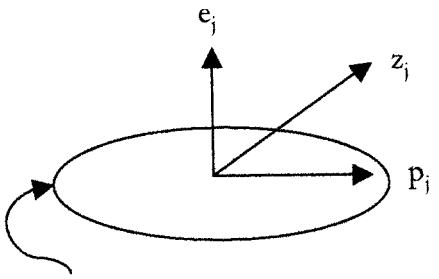
ahol a w_{jk} standardizált regressziós együttható a többváltozós regressziós egyenletben a k-adik változó hatását mutatja a j-edik változóra. Definíció szerint $w_{jj} = 0$, mivel egy változó önmagára vonatkozó regressziójától eltekintünk.

A p_j valószínűségi változót a j-edik megfigyelt változó image-ének (képének) nevezzük. Geometriailag ez a j-edik változónak a többi $(m-1)$ változó terére vonatkozó vetülete. A j-edik változó anti-image-e:

$$e_j = z_j - p_j, \tag{12.13}$$

ami a j-edik változónak azon része, amit a többi $(m-1)$ változó nem magyaráz. Az anti-image a többi változó terére ortogonális, vagyis az anti-image független a többi változótól.

A fentieket szemlélteti a következő ábra:



A z_k változók által kifeszített tér ($k = 1, 2, \ldots, m,\ k \neq j$)
Az image és anti-image geometriai reprezentációja

Általánosságban a (12.13) egyenletet a következőképpen írhatjuk:

$$\mathbf{z} = \mathbf{p} + \mathbf{e}, \tag{12.14}$$

ahol $\mathbf{z}$ a megfigyelt változók m-elemű oszlopvektora, $\mathbf{p}$ a megfelelő image-ek vektora, $\mathbf{e}$ pedig az anti-image-eket tartalmazza. A (12.14) egyenletet nevezzük az image-elemzés kiindulási feltételének.

Az m számú megfigyelt változó image-einek vektorát a (12.12) egyenlet alapján a következőképpen határozzuk meg:

$$\mathbf{p} = \mathbf{W}\,\mathbf{z}, \tag{12.15}$$

ahol $\mathbf{W}$ a többváltozós regressziós együtthatók mátrixa, amelynek minden sora a megfelelő változó regressziós becsléseinek együtthatóit tartalmazza. A regresszióelmélet szerint az együtthatók mátrixát a következő képlet alapján számítjuk:

$$\mathbf{W} = \mathbf{I} - \mathbf{S}^2\mathbf{R}^{-1}, \tag{12.16}$$

ahol $\mathbf{R}$ a megfigyelt változók korrelációmátrixa, és

$$\mathbf{S}^2 = [\text{diag } \mathbf{R}^{-1}]^{-1} \tag{12.17}$$

a becslések hibájának varianciamátrixa (diagonális mátrix), vagy másképpen az anti-image varianciák diagonális mátrixa.

A fentiekben feltételeztük, hogy a megfigyelt változók korrelációs mátrixa ($\mathbf{R}$) nem szinguláris.

A megfigyelt változók image-einek variancia-kovarianciamátrixát a (12.15) és (12.16) egyenletek alapján a következőképpen írhatjuk fel:

$$\begin{aligned}\mathbf{G} = E(\mathbf{p}\,\mathbf{p}') &= (\mathbf{I} - \mathbf{S}^2\mathbf{R}^{-1})\mathbf{R}(\mathbf{I} - \mathbf{R}^{-1}\mathbf{S}^2) \\ &= \mathbf{R} + \mathbf{S}^2\mathbf{R}^{-1}\mathbf{S}^2 - 2\mathbf{S}^2.\end{aligned} \tag{12.18}$$

Hasonlóan határozhatjuk meg az anti-image-ek variancia-kovarianciamátrixát ($\mathbf{Q}$) a fenti egyenletek alapján:

$$\mathbf{Q} = E(\mathbf{e}\,\mathbf{e}') = \mathbf{S}^2\mathbf{R}^{-1}\mathbf{S}^2. \tag{12.19}$$

Vegyük észre a (12.17) egyenlet alapján, hogy

$$\text{diag } \mathbf{Q} = \mathbf{S}^2, \tag{12.20}$$

vagyis az anti-image varianciák az $\mathbf{S}^2$ diagonális mátrix diagonálelemei. A (12.18) és (12.19) egyenleteket kombinálva jutunk az image-elemzés alapegyenletéhez:

$$\mathbf{R} = \mathbf{G} - \mathbf{Q} + 2\mathbf{S}^2, \tag{12.21}$$

ami azt fejezi ki, hogy a megfigyelt változók korrelációmátrixát felbonthatjuk az image és anti-image kovarianciáival összefüggő részekre.

Guttman (1956) a (12.12) és (12.13) egyenleteket parciális image-nek és parciális anti-image-nek nevezte, megkülönböztetve azon elméleti hipotetikus változóktól, amelyek a változók általános terében léteznek, ha $m \longrightarrow \infty$.

Így a j-edik változó általános image-e:

$$\pi_j = \lim_{m \longrightarrow \infty} \sum_{k=1}^{m} w_{jk} z_k \qquad (w_{jj} = 0), \tag{12.22}$$

ahol w_{jk}-k az általános regressziós együtthatók.

Hasonlóan az általános anti-image definíciója:

$$\varepsilon_j = z_j - \pi_j. \tag{12.23}$$

Guttman (1956) bebizonyította, hogy ha

$$\lim_{m \longrightarrow \infty} \frac{r}{m} = 0, \tag{12.24}$$

akkor az image-elemzés $m \longrightarrow \infty$ esetén (az általános image-elemzés) és a faktorelemzés megegyeznek.

Részletezve:

$$p_j \longrightarrow \pi_j = c_j, \tag{12.25}$$

ahol c_j a z_j változó közös faktorok által megmagyarázott része.

Hasonlóan az anti-image faktoranalitikus megfelelője:

$$e_j \longrightarrow \varepsilon_j = y_j, \tag{12.26}$$

ahogy y_j a z_j változó egyedi része, az a rész, amit a közös faktorok nem magyaráznak ($z_j = c_j + y_j$).

A faktorelemzés feltételezi, hogy a megfigyelt változók közös faktorai által nem magyarázott egyedi részek közötti, valamint a közös faktorok és az egyedi részek közötti kovarianciák nullák:

$$g_{jk} = \mathrm{Cov}(e_j, e_k) \longrightarrow \mathrm{Cov}(\varepsilon_j, \varepsilon_k) = \mathrm{Cov}(y_j, y_k) = 0 \qquad (j \neq k), \tag{12.27}$$

$$\mathrm{Cov}(p_j, e_k) \longrightarrow \mathrm{Cov}(\pi_j, \varepsilon_k) = \mathrm{Cov}(c_j, y_k) = 0. \tag{12.28}$$

A faktorelemzés modellje szerint érvényes még további három azonosság:

$$g_{jj} = \mathrm{Var}(p_j) \longrightarrow \mathrm{Var}(\pi_j) = \mathrm{Var}(c_j) = h_j^2, \tag{12.29}$$

$$q_{jj} = \mathrm{Var}(e_j) \longrightarrow \mathrm{Var}(\varepsilon_j) = \mathrm{Var}(y_j) = u_j^2, \tag{12.30}$$

$$g_{jk} = \mathrm{Cov}(p_j, p_k) \longrightarrow \mathrm{Cov}(\pi_j, \pi_k) = \mathrm{Cov}(c_j, c_k) = r_{jk} \quad (j \neq k), \tag{12.31}$$

ahol h_j^2 a j-edik változó kommunalitása, u_j^2 a j-edik változó egyedisége ($h_j^2 + u_j^2 = 1$) és r_{jk} a j-edik és a k-adik változó közötti korrelációs együttható.

Általánosságban azt mondhatjuk a (12.23) azonosság feltételezésével, hogy

$$\mathbf{G} \longrightarrow \mathbf{R} - \mathbf{U}^2, \tag{12.32}$$

valamint

$$\mathrm{diag}\ \mathbf{Q} = \mathbf{S}^2 \longrightarrow \mathbf{U}^2 \tag{12.33}$$

$$\mathbf{Q} \longrightarrow \mathbf{U}^2, \tag{12.34}$$

ahol $\mathbf{U}^2$ az egyediségek diagonális mátrixa.

A fenti egyenletek alapján azt mondhatjuk, hogy az image kovarianciamátrix jól közelíti a redukált korrelációmátrixot ($\mathbf{R} - \mathbf{U}^2$). A redukált korrelációmátrix diagonális elemei a megfigyelt változók kommunalitásai, vagyis a változók varianciáinak a közös faktorok által magyarázott része. Az image kovarianciamátrix diagonális elmei az egyes változóknak a többi változóra vonatkozó többszörös korrelációs együtthatójának a négyzetei (ezeket gyakran tekinthetjük a kommunalitások legjobb becsléseinek), a diagonálison kívüli elemek az image-kovarianciák, amelyek a megfelelő korrelációktól csak 0-hoz közelítő mértékben különböznek, véges n esetén a különbség egyenlő az anti-image kovarianciákkal.

A közelítést akkor mondhatjuk jónak, amikor az anti-image kovarianciamátrix közel diagonális, a diagonális elemein kívüli elemek a 0-hoz közeliek.

A faktormátrixot Rao (1955) kanonikus faktorelemzése alapján határozzuk meg. Rao eljárása, amely azonos megoldást ad Lawley (1940) maximum likelihood módszerével, a következő sajátérték-sajátvektor feladatra vezethető vissza:

$$(\mathbf{U}^{-1}\mathbf{R}\,\mathbf{U}^{-1} - \beta\mathbf{I})\boldsymbol{\xi} = \mathbf{0}. \tag{12.35}$$

Legyen az $\mathbf{U}^2$ mátrix kezdeti becslése $\mathbf{S}^2$, ekkor a (12.35) egyenlet a következőképpen írható:

$$[\mathbf{S}^{-1}\mathbf{R}\,\mathbf{S}^{-1} - b\,\mathbf{I}]\mathbf{x} = \mathbf{0}. \tag{12.36}$$

Jelölje $\mathbf{B}$ az $\mathbf{R}^* = \mathbf{S}^{-1}\mathbf{R}\,\mathbf{S}^{-1}$ mátrix sajátértékeinek diagonális mátrixát, akkor $\mathbf{B}$ diagonális elemei (b) a β becslési, $\mathbf{X}$ pedig az $\mathbf{R}^*$ egységnyi hosszúságú sajátvektorait tartalmazza (oszlopvektorai a ξ becslései).

A főkomponens-elemzés eredményeit felhasználva a (12.35) egyenletből az $\mathbf{R}^*$ mátrix faktormátrixa $\mathbf{X}\mathbf{B}^{1/2}$. Ennek alapján magának az $\mathbf{R}$ mátrixnak a faktormátrixa:

$$\boldsymbol{\Lambda}_r = \mathbf{S}\,\mathbf{X}\,\mathbf{B}^{1/2}. \tag{12.37}$$

Az image kovarianciamátrix $\mathbf{G}$ faktormátrixát a $\mathbf{G}^* = \mathbf{S}^{-1}\mathbf{G}\,\mathbf{S}^{-1}$ mátrix sajátértékei és sajátvektorai alapján számítjuk. Harris (1962) mutatta meg a (12.19) és a (12.21) egyenletek alapján, hogy $\mathbf{G}^*$ sajátvektorai egyenlők $\mathbf{R}^*$ sajátvektoraival, de a sajátértékek különböznek, $\mathbf{B}$ helyett $(\mathbf{B} - \mathbf{I})^2\mathbf{B}^{-1}$-vel egyenlők. Így a $\mathbf{G}$ mátrix faktormátrixa:

$$\mathbf{\Lambda}_g = \mathbf{S}\mathbf{X}[(\mathbf{B} - \mathbf{I})^2\mathbf{B}^{-1}]^{1/2}. \tag{12.38}$$

A két faktormátrixot összehasonlítva látható, hogy közöttük az oszlopvektoraik skálaterjedelmében van különbség, mégpedig $\mathbf{\Lambda}_g = \mathbf{\Lambda}_r\mathbf{D}$, ahol

$$\mathbf{D} = \mathbf{I} - \mathbf{B}^{-1}. \tag{12.39}$$

Harris mutatott rá arra is, hogy a két faktormátrix megfelelő oszlopai közötti korrelációk megegyeznek. Ezenkívül belátható, hogy az image-faktorok ($\mathbf{\Lambda}_g$), hasonlóan a kanonikus faktorokhoz, invariánsak az eredeti mértékegységek megváltoztatásával szemben. Fontos rámutatni a $\mathbf{G}^*$ és az $\mathbf{R}^*$ mátrixok sajátértékei közötti eltérésre. A $\mathbf{G}^*$ mátrix $(b-1)^2/b$ sajátértéke és az $\mathbf{R}^*$ mátrix b sajátértéke közötti függvény monoton növekvő $b > 1$ esetén, azonban ha $b < 1$, akkor a függvény pontosan ellentétesen mozog, monoton csökkenővé válik ($b = 1$ esetén a függvény értéke 0). A sajátértékek ezen ellentétes mozgása miatt az 1-nél kisebb sajátértékű faktorokat ki kell hagynunk az értelmezésből.

12.4. Rao-féle kanonikus faktorelemzés

A Rao-féle kanonikus faktorelemzés során keressük azokat a faktorokat, amelyek maximálisan korrelálnak a megfigyelt változókkal, és korrelálatlanok egymással.

Induljunk ki az alapmodellből, abból, hogy a megfigyelt változókat kifejezhetjük a szisztematikus vagy közös komponensek, és a hiba vagy egyedi komponensek lineáris függvényeként:

$$\mathbf{Z} = \mathbf{C} + \mathbf{E}, \tag{12.40}$$

ahol $\mathbf{Z}$ a standardizált megfigyelt változók mátrixa ($n \times m$ típusú, ahol n a megfigyelési egységek száma, m a változók száma),

$\mathbf{C}$ a közös komponensek mátrixa (szintén $n \times m$ típusú),

$\mathbf{E}$ a hibakomponensek mátrixa (szintén $n \times m$ típusú).

Feltételezzük, hogy a közös és az egyedi komponensek korrelálatlanok, és hogy az egyedi komponensek függetlenek egymástól.

A közös komponenseket a következőképpen fejezzük ki:

$$\mathbf{C} = \mathbf{F}\,\mathbf{\Lambda}', \tag{12.41}$$

ahol $\mathbf{F}$ a nem standardizált faktorértékek mátrixa ($n \times r$ típusú, ahol n a megfigyelési egységek száma, r a faktorok száma), $\mathbf{\Lambda}$ a faktorsúlyok mátrixa ($m \times r$ típusú, ahol m a megfigyelt változók száma, r a faktorok száma).

A kanonikus faktor-modellt elsősorban Harris (1956, 1962, 1967, 1968) anyagai alapján tárgyaljuk.

Tartalmazza a $\mathbf{Z}$ mátrix a megfigyelt változók standardizált értékeit súlyozva oly módon, hogy $\mathbf{Z}\,\mathbf{Z}'$ a megfigyelt változók közötti páronkénti korrelációs együtthatókat adja! Ha a megfigyelt z változókat és a nem megfigyelt $\mathbf{F}$ faktorokat egy hipermátrixba

foglaljuk, akkor a

$$\begin{pmatrix} \mathbf{Z}' \\ \mathbf{F}' \end{pmatrix} [\mathbf{Z}, \mathbf{F}] = \begin{pmatrix} \mathbf{R} & \mathbf{\Lambda} \\ \mathbf{\Lambda}' & \mathbf{I} \end{pmatrix} \tag{12.42}$$

úgynevezett szupermátrixhoz jutunk.

Anderson (1958) adott kanonikus elemzést az ilyen mátrixra. A fenti mátrix esetén a kanonikus korrelációk négyzeteit a következő egyenlet α_i paraméterei adják:

$$|\mathbf{\Lambda}\,\mathbf{\Lambda}' - \alpha\mathbf{R}| = 0. \tag{12.43}$$

Az egyenlet α_i gyökei a megfigyelt változóhalmaz és a nem megfigyelt faktorok halmaza közötti kanonikus korrelációk négyzetei.

Ha $r = m$, vagyis a faktorok száma megegyezik a változók számával, akkor $\mathbf{\Lambda}\,\mathbf{\Lambda}' = \mathbf{R}$, és a determináns minden gyöke egyenlő 1-gyel, ha $\mathbf{R}$ nem szinguláris. (Ha $\mathbf{R}$ szinguláris, a rangja kisebb a rendjénél, akkor egy vagy több α_i egyenlő lesz 0-val.)

Ha $r < m$, akkor két lehetőség adódhat. Az első lehetőség az $\mathbf{R}$ mátrix hiányos faktorelemzése, ebben az esetben a determináns gyökei közül r számú 1-gyel, $(m - 1)$ számú pedig 0-val egyenlő. A másik esetet kommunalitás típusú megoldásnak nevezik az irodalomban. Ekkor $\mathbf{\Lambda}\,\mathbf{\Lambda}'$ reprodukálja (közelítőleg) $\mathbf{R}$ diagonálison kívül elemeit (a korrelációkat), de a diagonális elemeket nem, vagyis $\mathbf{\Lambda}\,\mathbf{\Lambda}' = \mathbf{R} - \mathbf{U}^2$, ahol $\mathbf{U}^2$ az egyedi varianciák becsléseinek diagonális mátrixa. Ebben az esetben módosíthatjuk a determináns egyenletet:

$$|\mathbf{R} - \mathbf{U}^2 - \alpha\mathbf{R}| = 0,$$

vagy

$$|\mathbf{R} - \beta\mathbf{U}^2| = 0, \tag{12.44}$$

ahol $\beta_i = 1/(1 - \alpha_i)$, és $\alpha_i = (\beta_i - 1)/\beta_i$.

Az α_i értékekről feltételezzük, hogy nemnegatív valós számok (kanonikus korrelációk négyzetei), így a β_i-k nagyobbak 1-nél.

Számítási megfontolások miatt Rao transzformálta ezt az egyenletet a következő egyenletté, amelynek gyökei – ahogyan ezt Anderson (1958) megmutatta – egyenlők a transzformált egyenlet gyökeivel:

$$|\mathbf{U}^{-1}\mathbf{R}\,\mathbf{U} - \mathbf{I} - \beta\mathbf{I}| = 0. \tag{12.45}$$

A fenti egyenletben az $\mathbf{R}$ mátrixról feltételezzük, hogy pozitív definit, $\mathbf{U}^2$ pedig, mivel diagonális mátrix, melynek elemei 0 és +1 közé esnek, szintén pozitív definit.

Jelölje $\mathbf{Q}$ az $\mathbf{U}^{-1}\mathbf{R}\,\mathbf{U}^{-1}$ mátrix normalizált sajátvektorait. Ekkor:

$$\mathbf{U}^{-1}\mathbf{R}\,\mathbf{U}^{-1} = \mathbf{Q}\,\langle\beta_i\rangle\,\mathbf{Q}', \tag{12.46}$$

amiből:

$$\mathbf{R} - \mathbf{U}^2 = \mathbf{U}\,\mathbf{Q}\,\langle\beta_i - 1\rangle\,\mathbf{Q}'\mathbf{U}. \tag{12.47}$$

Olyan $\mathbf{U}^2$ mátrixot kell választanunk, hogy $\mathbf{R} - \mathbf{U}^2$ Gram-féle mátrix legyen, akkor:

$$\mathbf{\Lambda} = \mathbf{U}\,\mathbf{Q}\,\langle\beta_i - 1\rangle^{1/2}. \tag{12.48}$$

Az $\mathbf{U}^2$ mátrix diagonális elemeinek kezdeti becslésére Harris (1963) ad eljárást, amely szerint:

$$u_j^2 < 1/r^{jj}, \tag{12.49}$$

ahol r^{jj} az $\mathbf{R}^{-1}$ mátrix megfelelő diagonális eleme.

Guttman javaslata, hogy u_j^2 kezdeti értéke $s_j^2 = 1/r^{jj}$ legyen (az egyes változóknak a többi változóra vonatkozó többszörös korreláció négyzete: $1 - 1/r^{jj}$).

Harris (1967) javasolta a kezdeti becslésnek a következő módosítását:

$$[v_j^2]' = [r^{jj}]'[(r^{jk})^2]^{-1}, \tag{12.50}$$

ahol $[(r^{jk})^2]^{-1}$ kvadratikus mátrix, melynek elemei $\mathbf{R}^{-1}$ megfelelő elemeinek a négyzetei,

$[r^{jj}]'$ az $\mathbf{R}^{-1}$ mátrix diagonális elemeiből álló sorvektor.

Megmutatható, hogy az egyedi varianciák és fenti kezdeti becsléseik között a következő egyenlőtlenség írható fel:

$$0 < u_j^2 < v_j^2 < s_j^2 < 1,$$

vagyis v_j^2 jobb kezdeti becslése az egyedi varianciának, mint a hagyományosan alkalmazott s_j^2.

Harris (1963) idézett tanulmányában ismerteti Rao eljárását a kezdeti értékek módosítására és így a kanonikus faktorok becslésére.

A kezdeti értékeket a (12.45) egyenletbe helyettesítjük, majd megoldjuk a sajátérték-sajátvektor feladatot. A kezdeti értéket a sajátértékek és sajátvektorok felhasználásával a következőképpen módosítjuk:

$$g_{jj} = \sum_{i=1}^{r} ({}_1\beta_i - 1)_1 q_{ji}^2 + 1. \tag{12.51}$$

A (12.51) egyenlettel módosított kezdeti becsléseket visszahelyettesítjük a (12.45) egyenletbe, majd megoldva az egyediségek újabb becsléséhez jutunk. Addig folytatjuk az iteratív eljárást rögzített m érték mellett, ameddig a megoldás nem konvergál (0,005 tolerancia szinten). Ezután számíthatjuk a (12.48) egyenlet alapján a faktormátrixot.

12.5. Alfa-faktorelemzés

Az alfa-faktorelemzés a pszichometrikus faktorelemzés kategóriába tartozik abban az értelemben, ahogyan Kaiser és Caffrey (1965) különbséget tett a faktorelemzés statisztikus és pszichometrikus alkalmazása között. (A statisztikus megközelítés a populáció jellemzőire következtet a minta megfigyelési egységei alapján, míg a pszichometrikus megközelítés a változóknak egy mintájából a változók univerzumának jellemzőire következtet.)

Az alfa-faktorelemzés során feltételezzük, hogy a változóknak létezik egy általános tere, és a vizsgálatba bevont változóhalmaz annak csupán egy véletlen jellegű mintája. Keressük a megfigyelt változóknak azokat a közös faktorait, amelyek a legmagasabb korrelációt mutatják az általános tér megfelelő faktoraival. A megfigyelt változók faktorainak megbízhatóságát (reliability) a Cronbach-féle alfa együtthatóval mérjük, ami a Kuder–Richardson megbízhatósági együttható általánosítása.

A Cronbach-féle alfa együttható a megbízhatósági együtthatók ún. felező (split-half) típusához sorolható. Ennél az eljárásnál a vizsgált problémára összegyűjtött változókat két részre bontjuk, így két becsléshez juthatunk. Ezután a Spearman–Brown-formulát alkalmazva megbízhatósági együtthatót számítunk, ami hasonlít a két mérés között számított korrelációhoz. Cronbach (1951) megmutatta, hogy az alfa együttható egyenlő a felező együtthatók várható értékével.

Bentler (1968), valamint Kaiser és Caffrey (1965) is definiált egy alfa együtthatót, amelyek különböző eredményekre vezettek. Ezeket a későbbiekben ismertetjük.

Az eljárás bemutatásakor most is a faktorelemzés alapmodelljéből, a $\mathbf{Z} = \mathbf{C} + \mathbf{E}$ egyenletből indulunk ki. A klasszikus méréselmélet a véletlenszerű hiba mellett megkülönböztet szisztematikus hibát is, így a megfigyelt adatokat három hatás eredőjeként értelmezzük: $\mathbf{Z} = \mathbf{C} + \mathbf{E}_s + \mathbf{E}_e$. Ennek megfelelően a hibavarianciát két részre bontjuk: $\mathbf{U}^2 = \mathbf{E}_s^2 + \mathbf{U}_e^2 = \mathbf{I} - \mathbf{H}^2$, ahol $\mathbf{H}^2$ a közös komponensek varianciáit, vagy más néven a kommunalitásokat tartalmazza.

A megbízhatósági együttható a nem véletlenszerű komponensek varianciáinak és a teljes varianciának az aránya, vagy másképpen fogalmazva, a szisztematikus komponensek (a valódi – elméleti – értékek) varianciájának és a megfigyelt értékek varianciájának az aránya:

$$r_{zz} = \frac{\mathbf{1}'(\mathbf{R} - \mathbf{U}_e^2)\mathbf{1}}{\mathbf{1}'\mathbf{R}\,\mathbf{1}}, \tag{12.52}$$

ahol $\mathbf{1}$ az összegzővektort jelöli.

Ezt az együtthatót nevezik Kuder–Richardson-féle együtthatónak. Thomson (1940) és Peel (1948) oldották meg először azt a problémát, hogy a változókhoz úgy rendeljenek súlyokat, hogy a megbízhatósági együttható maximális legyen. A megoldásra Bentler (1968) adott korszerű leírást. Kereste azt a $\mathbf{w}$ súlyvektort, amely mellett a következő függvény felveszi maximumát:

$$r_M = \frac{\mathbf{w}'(\mathbf{R} - \mathbf{U}_e^2)\mathbf{w}}{\mathbf{w}'\mathbf{R}\,\mathbf{w}}. \tag{12.53}$$

A függvénynek $\mathbf{w}$ szerinti parciális deriváltját véve, azt egyenlővé téve nullával, és némi algebrai átalakítás után (lásd Bentler [1968], 336–337) a következő sajátérték-sajátvektor egyenlethez jutunk:

$$[\mathbf{U}_e^{-1}(\mathbf{R} - \mathbf{U}_e^2)\mathbf{E}_e^{-1} - \beta^2\mathbf{I}]\mathbf{v} = \mathbf{0}, \tag{12.54}$$

ahol $\mathbf{v} = \mathbf{U}_e\mathbf{w}$,

$\beta = r_M/(1 - r_M)$.

A fenti egyenlet megoldása adja azokat a súlyokat, amelyek maximálják a megbízhatósági együtthatót, és ezen egyenlet, amit a főkomponens- és faktorelemzésnél jól ismerünk, megoldásainak sajátértékei és sajátvektorai alapján számítjuk a faktorelemzés mátrixait.

Cronbach definiált a megbízhatóság mérésére egy másik együtthatót (Cronbach, 1951):

$$\alpha = \frac{m}{m-1}\left(1 - \frac{\Sigma V_i}{V_t}\right), \tag{12.55}$$

ahol m a változók száma,

V_t a varianciák és kovarianciák összege,

V_i az i-edik változó varianciája.

Bentler bebizonyította, hogy ha érvényes a belső konzisztenciának $\mathbf{U}_e^2$-re vonatkozó feltétele: $\mathbf{1}'\mathbf{U}_e^2\mathbf{1} = m(1 - r_{ij}^a)$, ahol r_{ij}^a a változók közötti átlagos korrelációt jelöli, akkor az α együttható megegyezik a (12.52) egyenletben definiált megbízhatósági együtthatóval (a bizonyítást lásd Bentler (1968), 337–338):

$$r_{zz} = \alpha = \frac{m}{m-1}\left(1 - \frac{\mathbf{1}'\mathbf{I}\,\mathbf{1}}{\mathbf{1}'\mathbf{R}\,\mathbf{1}}\right). \tag{12.56}$$

Maximalizáljuk a Cronbach-féle α-t a (12.56) egyenletből kiindulva.

$$\alpha = \frac{m}{m-1}\left(1 - \frac{\mathbf{p}'\mathbf{I}\,\mathbf{p}}{\mathbf{p}'\mathbf{R}\,\mathbf{p}}\right). \tag{12.57}$$

Az α együtthatót maximalizálhatjuk

$$\gamma^2 = \frac{\mathbf{p}'\mathbf{R}\,\mathbf{p}}{\mathbf{p}'\mathbf{I}\,\mathbf{p}} \tag{12.58}$$

maximalizálásával. A szélsőérték-számítást elvégezve a következő egyenlethez jutunk:

$$(\mathbf{R} - \gamma^2\mathbf{I})\mathbf{p} = \mathbf{0}. \tag{12.59}$$

Láthatjuk, hogy a Cronbach-féle alfa maximalizálásával a főkomponens-elemzéssel megegyező eredményre jutunk.

A főkomponensek számának meghatározásához helyettesítsük be a (12.59) egyenletet (12.57)-be:

$$\alpha = \frac{m}{m-1}\left(1 - \frac{1}{\gamma^2}\right). \tag{12.60}$$

Ebből láthatjuk, hogy a megbízhatósági együttható akkor lesz 0-nál nagyobb, ha $\gamma^2 > 1$ (a sajátérték nagyobb 1-nél), és ez pedig jól ismert kritérium a főkomponensek számának meghatározásához.

Bentler mutatott rá, hogy a faktorelemzésnél az $(\mathbf{R} - \mathbf{U}_e^2)$ mátrix helyett az $(\mathbf{R} - \mathbf{U}^2)$ mátrixot vizsgáljuk, és mivel $\mathbf{R} - \mathbf{U}^2 = (\mathbf{R} - \mathbf{U}_e^2) - \mathbf{U}_s^2$, a faktorelemzésnél a szisztematikus komponenshez nem vesszük hozzá a specifikus részt, így a közös varianciából kihagyjuk a specifikus varianciát. Ennek megfelelően Bentler javasolt egy mutatót, amely a közös komponens varianciájának és a teljes varianciának az arányát méri:

$$\alpha_0 = \frac{\mathbf{1}'(\mathbf{R} - \mathbf{U}^2)\mathbf{1}}{\mathbf{1}'\mathbf{R}\,\mathbf{1}}. \tag{12.61}$$

Ezt a mutatót belső konzisztencia-együtthatónak nevezte, és kereste azt a $\mathbf{w}$ súlyvektort, amely az

$$\alpha_0 = \frac{\mathbf{w}'(\mathbf{R} - \mathbf{U}^2)\mathbf{w}}{\mathbf{w}'\mathbf{R}\,\mathbf{w}} \tag{12.62}$$

együtthatót maximalizálja. A maximalizálási feladatot az előzőekhez hasonlóan elvégezve a következő sajátérték-sajátvektor egyenlethez juthatunk:

$$[\mathbf{U}^{-1}(\mathbf{R} - \mathbf{U}^2)\mathbf{U}^{-1} - \mu^2\mathbf{I}]\mathbf{v} = \mathbf{0}, \tag{12.63}$$

ahol $\mathbf{v} = \mathbf{U}\,\mathbf{w}$, és

$$\mu^2 = \frac{\alpha_0}{1 - \alpha_0}. \tag{12.64}$$

A fenti egyenlet alapján a belső konzisztencia-együtthatót kifejezhetjük az i-edik sajátérték segítségével:

$$\alpha_{0i} = \frac{\mu_i^2}{\mu_i^2 + 1}. \tag{12.65}$$

A (12.65) azonosságból láthatjuk, hogy nem találjuk a változóknak olyan súlyozását, amellyel a közös komponenseket pontosan elő tudnánk állítani, vagyis hogy a megbízhatóság tökéletes legyen. A fentiek alapján azt mondhatjuk, hogy annál nagyobb lesz a faktorelemzés megbízhatósága, minél nagyobb lesz az első sajátértékeknek a nagysága, amihez szükséges feltétel, hogy a változók száma, valamint a változók számának és az $(\mathbf{U}^{-1}(\mathbf{R} - \mathbf{U}^2)\mathbf{U}^{-1})$ sajátérték-mátrix rangjának (r) aránya nagy legyen.

A faktorsúlyok mátrixát a sajátértékek és sajátvektorok alapján számíthatjuk most is ki.

Kaiser és Caffrey (1965) definiált a Cronbach-féle α-hoz hasonló, de csak a közös komponensekre koncentráló mutatót:

$$\alpha_c = \frac{m}{m-1}\left(1 - \frac{\mathbf{w}'\mathbf{H}^2\mathbf{w}}{\mathbf{w}'(\mathbf{R}-\mathbf{U}^2)\mathbf{w}}\right). \tag{12.66}$$

Kaiser és Caffrey kereste a változóknak azt a súlyozását, amelyik az α_c mutatót maximalizálja. A maximumfeladatot megoldva a következő egyenlethez jutottak:

$$[\mathbf{H}^{-1}(\mathbf{R}-\mathbf{U}^2)\mathbf{H}^{-1} - \delta^2\mathbf{I}]\mathbf{q} = \mathbf{0}, \tag{12.67}$$

ahol $\mathbf{q} = \mathbf{H}\,\mathbf{w}$.

A δ_i sajátérték és az α_{ci} közötti összefüggés:

$$\alpha_{ci} = \frac{m}{m-1}\left(1 - \frac{1}{\delta_i^2}\right). \tag{12.68}$$

A megbízhatóság a fenti egyenlet alapján kisebb vagy egyenlő lesz 0-val, ha a sajátérték kisebb vagy egyenlő 1-gyel. Ezért az alfa-faktorelemzésnél csak azokat a faktorokat vesszük figyelembe, amelyeknél a sajátértékek nagyobbak 1-nél.

Kaiser és Caffrey alfa-faktorelemző eljárása – mint említettük – a megfigyelt változók közös komponenseire koncentrál. Könnyen megmutatható, hogy tulajdonképpen a közös komponensek korrelációmátrixának faktorelemzését végzi el.

Tegyük fel a jelölések egyszerűsége miatt, hogy a standardizált változók $\mathbf{Z}$ mátrixát a megfigyelési egységek száma négyzetgyökének reciprokával súlyoztuk, így két változó között a korrelációt $\mathbf{z}_i\mathbf{z}_j/n$ helyett egyszerűen $\mathbf{z}_i\mathbf{z}_j$ formulával számíthatjuk.

Bontsuk fel két megfigyelt változó korrelációs együtthatóját a közös komponens és a hibatag alapján

$$r_{ij} = \mathbf{z_i z_j} = \mathbf{c_i'c_j} + \mathbf{e_i'e_j} = \mathbf{c}_i'\mathbf{c}_j,$$

mivel a hibatagok függetlenek.

$\mathbf{c}_i'\mathbf{c}_j$ a két közös komponens közötti kovarianciával egyenlő. A közös komponensek varianciáját a megfelelő megfigyelt változók kommunalitása adja ($\mathbf{c}_i'\mathbf{c}_j = h_i^2$). Ha a kovarianciákat osztjuk a megfelelő varianciákkal, akkor a korrelációs együtthatóhoz jutunk. Mátrixjelölésekkel:

$$\mathbf{H}^{-1}(\mathbf{R}-\mathbf{U}^2)\mathbf{H}^{-1},$$

ami a megfigyelt változók közös komponenseinek a korrelációmátrixa.

A faktorelelemzés iteratív eljárását a kommunalitások becslésével kezdjük. A kommunalitások kezdeti becslései a változók többszörös korrelációs együtthatói, amelyeket a korrelációmátrix megfelelő indexű diagonális elemeivel kicserélünk, és az így módosított mátrixból kiindulva keressük azt a közös faktort, amelynek megbízhatósági együtthatója a legnagyobb. Ezután a korrelációs mátrixot adjusztáljuk, és az eljárást addig folytatjuk, amíg a kommunalitások nem konvergálnak. A faktorok egyre csökkenő megbízhatóságúak, az első faktornak lesz a legnagyobb a megbízhatósági együtthatója, a második faktornak a második legnagyobb, és így tovább.

A faktoroknak ez a tulajdonsága a rotálásnál megváltozik. Az elemzéshez annyi faktort tartunk meg, amennyinek a sajátértéke nagyobb 1-nél.

12.6. Maximum likelihood faktorelemzés

Négy különböző faktorelemző eljárás is a faktormátrix (faktorsúlyok) elemeinek lényegében azonos becsléséhez vezet: Lawley (1940, 1941) maximum likelihood módszere, Rao (1955) kanonikus korreláció elméletén alapuló módszere, Bargmann (1957) és Howe (1955) maximális determináns módszere, amely a faktorsúlyok becsléséhez maximalizálja a megfigyelt változók feltételes korrelációmátrixának determinánsát (a megfigyelt változóknak a faktorok kiszűrésével számított parciális korrelációmátrixának determinánsát), továbbá Jöreskog (1966) lineáris strukturális relációkra kidolgozott modellje (LISREL), amely maximum likelihood becslést ad a faktorsúly mátrix-elemeire.

Lawley likelihood egyenletei és Rao kanonikus faktoregyenletei azonos identifikációs feltételeket tartalmaznak: hogy $\mathbf{\Lambda}'\mathbf{U}^{-2}\mathbf{\Lambda}$ mátrix diagonális legyen, és hogy azonos egyértelmű megoldást adjanak a faktormátrixra.

A faktorelemzésnek az az általános feltétele, hogy az első faktor járuljon maximális mértékben hozzá a megfigyelt változók varianciáihoz, a második a maradék lehetőségek közül járuljon hozzá az összvarianciához maximális mértékben azzal a feltételezéssel, hogy független legyen az előző faktortól, és így tovább, azzal az identifikációs feltétellel függ össze, hogy $\mathbf{\Lambda}'\mathbf{U}^{-2}\mathbf{\Lambda}$ diagonális legyen (Lawley és Maxwell [1971], 4. fejezet).

A maximális determináns egyenleteknek nincs egyértelmű megoldása, de ortogonális transzformációval megkaphatjuk azt a megoldást, amelyik kielégíti a fenti identifikációs feltételeket.

Lawley eljárásáról Howe (1955) kimutatta, hogy sokkal lassabban konvergál, mint a maximális determináns módszer. Lord (1956) viszont azt találta, hogy bizonyos esetekben, ha a kezdeti becslés nem elég jó, Lawley eljárása megáll, és a faktorsúlyokra rossz becslést ad. Ezért Lawley eljárását nem használják.

Rao kanonikus faktorelemző eljárását Browne (1968) vizsgálta, és tapasztalata szerint az eljárás lassabban konvergál, mint a maximális determináns módszer, és előfordulhat olyan eset, hogy az egymást követő iterációs lépésekben az u_i^2 értékek között csak nagyon kicsi az eltérés akkor, amikor még ezek az értékek a valódi értékektől nagyon messze vannak.

Bargmann és Howe maximális determináns módszere a megfigyelt változók közötti korrelációs együtthatók mátrixának determinánsát maximálja, ahol a kontrollváltozók a faktorok.

A parciális korrelációmátrix:

$$\mathbf{R}_{(x|f)} = \mathbf{U}^{-1}(\mathbf{R} - \mathbf{\Lambda}\,\mathbf{\Lambda}')\mathbf{U}^{-1}. \tag{12.69}$$

A determináns:

$$|\mathbf{R}_{(x|f)}| = |\mathbf{R}||\mathbf{I} - \mathbf{\Lambda}'\mathbf{R}^{-1}\mathbf{\Lambda}||\mathbf{U}|^2. \tag{12.70}$$

A determináns akkor lesz maximális, ha $\mathbf{R}_{(x|f)}$ az egységmátrix.

Az eljárásban a kommunalitások kezdeti becslésének a többszörös korrelációs együtthatók becslésének a négyzetét (R_j^2) tekintik. Maximum likelihood becslés esetén:

$$R_j^2 = 1 - 1/r^{jj},$$

ahol r^{jj} az $\mathbf{R}^{-1}$ inverz mátrix j-edik diagonális elme.

Ennek alapján az egyedi varianciák kezdeti becslései:

$$u_j^2 = 1/r^{jj},$$

általában

$$\mathbf{U}^2 = (\text{diag } (\mathbf{R}^{-1}))^{-1}.$$

Bargman eljárásában az iterációt akkor állítják le, ha az egymás utáni lépésekben az u_i^2-ek eltérésének abszolút értéke nem nagyobb, mint 0,0001.

Annak a nullhipotézisnek a tesztelésére, hogy r számú faktor elégséges a korrelációmátrix reprodukálásához, a likelihood hányados statisztikát alkalmazzuk (lásd Howe, 1957):

$$\alpha_r = -k \log_e |\mathbf{R}_{(x|f)}|. \tag{12.71}$$

A k multiplikátort Bartlett (1950) adta meg:

$$k = n - (2m + 11)/6 - (2r)/3.$$

Az α_r statisztika aszimptotikusan $1/2((m - r)^2 - m - r)$ szabadságfokú khi-négyzet eloszlást követ.

Lawley (1940) a likelihood hányados statisztikának könnyen számolható közelítő formáját határozta meg:

$$\alpha_r = k \sum_{j=1}^{m} \sum_{i=1}^{j} r_{ij}^{*2},$$

ahol r_{ij}^{*2} a parciális korrelációmátrix i-edik sorának j-edik eleme.

A likelihood hányados statisztikát 0, 1, 2.... számú faktorra egymás után alkalmazzuk egészen addig, amíg a nullhipotézist már nem kell elvetnünk (az adott valószínűségi szint mellett). A faktorok számának becslése az első olyan lépésben levő faktorszám lesz, amelyben a nullhipotézist nem utasítjuk el.

Ha feltételezzük, hogy a megfigyelt változók többváltozós normális eloszlású valószínűségi változók, akkor a minta variancia-kovarianciamátrixa Wishart eloszlást követ $(n-1)$ szabadságfokkal (lásd például Mardia, Kent és Bibby [1979]). A faktormodellben a populáció variancia-kovarianciamátrixát ($\boldsymbol{\Sigma}$) a következő egyenlettel fejezzük ki:

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\,\boldsymbol{\Lambda}' + \mathbf{U}^2. \tag{12.72}$$

A likelihood függvény logaritmusa (Anderson [1958]) a megfigyelésektől függő konstans elhagyása után

$$\log_e L = 1/2(n-1)[\log_e |\boldsymbol{\Sigma}| + \text{trace}(\mathbf{S}\,\boldsymbol{\Sigma}^{-1})], \tag{12.73}$$

ahol $\mathbf{S}$ a megfigyelt változók mintából becsült variancia-kovarianciamátrixa.

A (12.73) és a (12.72) egyenletek alapján látható, hogy a likelihood-függvény logaritmusa a faktorsúlyok és a specifikus variancia függvénye. Ezen paramétereket becsülhetjük a likelihood-függvény logaritmusának maximalizálásával úgy, hogy $\boldsymbol{\Lambda}'\mathbf{U}^{-2}\boldsymbol{\Lambda}$ mátrix diagonális legyen. A gyakorlatban célszerűbb a likelihood-függvény logaritmusa helyett a következő függvényt minimalizálni:

$$F = \text{trace}(\mathbf{S}\,\boldsymbol{\Sigma}^{-1}) + [\log_e |\boldsymbol{\Sigma}| - \log_e |\mathbf{S}|] - m. \tag{12.74}$$

Az F függvény minimalizálása ekvivalens $\log_e L$ maximalizálásával, mivel $F = -(c_1 \log_e L + c_2)$, ahol c_1 és c_2 konstansok.

Ha $\mathbf{S}$ és $\boldsymbol{\Sigma}$ elemei hasonlóak, akkor az inverzeik is hasonlóak. Ezért $\mathbf{S}\,\boldsymbol{\Sigma}^{-1}$ közelít az egységmátrixhoz, ha $\mathbf{S}$ és $\boldsymbol{\Sigma}$ közelít egymáshoz. Mivel egy mátrix nyoma a diagonális elemeinek összege, az első tag m-hez közelít, ha $\mathbf{S}$ és $\boldsymbol{\Sigma}$ közelít egymáshoz. A második tag $\boldsymbol{\Sigma}$ és $\mathbf{S}$ determinánsa logaritmusainak különbsége, ezért a második tag 0-hoz közelít, ha $\boldsymbol{\Sigma}$ közelít $\mathbf{S}$-hez. A harmadik tag konstans, a megfigyelt változók számával egyenlő. Eszerint ha $\mathbf{S}$ és $\boldsymbol{\Sigma}$ egyenlő, az illeszkedést mérő függvény egyenlő 0-val.

Az F függvény minimalizálására Jöreskog (1967) javasolt egy sikeres eljárást, amely először a faktorsúlyok szerint minimalizálja az F függvényt a specifikus varianciák rögzített értékei mellett, majd a faktorsúlyok rögzítése mellett minimalizálja F értékét az $\mathbf{U}^2$ értékei szerint.

12.7. Legkisebb négyzetek módszere

A klasszikus legkisebb négyzetek módszere (unweighted least squares) minimalizálja a megfigyelt és a modell által reprodukált kovarianciák különbségeit:

$$F_{ULS} = \text{trace}[(\mathbf{S} - \boldsymbol{\Sigma})^2]. \tag{12.75}$$

A legkisebb négyzetek módszere $\boldsymbol{\Lambda}$ és $\mathbf{U}^2$ becsléséhez a reziduális variancia-kovarianciákat minimalizálja.

A legkisebb négyzetek módszerének legnagyobb előnye, hogy nem tételez fel a változók eloszlásáról semmit. Hátránya viszont, hogy egyrészt nem ad statisztikai próbát a modelllilleszkedésére, másrészt, hogy skálafüggő. Ha egy módszer skálafüggő, akkor megváltoztatva a megfigyelt változók mértékegységét (skáláját), az illeszkedést mérő függvény minimumhelye is megváltozik, és a becslésekben történő változás csupán a skála változását tükrözi. A legkisebb négyzetek módszere különböző eredményt adhat, ha a mértékegységeket változtatjuk. Ilyen esetben célszerű a változókat standardizálni, így a korrelációmátrixot elemezni (a variancia-kovarianciamátrix helyett).

12.8. Általánosított legkisebb négyzetek módszere

Az általánosított legkisebb négyzetek módszere abban különbözik a klasszikus legkisebb négyzetek módszerétől, hogy az $\mathbf{S}$ és $\boldsymbol{\Sigma}$ közötti különbséget súlyozza az $\mathbf{S}^{-1}$ elemeivel, így ez a módszer már invariáns a skála (mértékegység) megválasztására.

A modell illeszkedését mérő függvény:

$$F_{GLS} = \text{trace}[(\mathbf{S} - \boldsymbol{\Sigma})\mathbf{S}^{-1}]^2. \tag{12.76}$$

Amennyiben a megfigyelt változók együttes valószínűségeloszlása normális, az általánosított legkisebb négyzetek módszere aszimptotikusan ekvivalens a maximum likelihood módszerrel (lásd Lee, 1977, Bowne, 1974).

12.9. Faktorelemző eljárások összehasonlítása

A faktorelemzés különböző módszereit alapvetően aszerint hasonlíthatjuk össze, hogy az eljárás mit is optimalizál, hogyan definiálja az egyediségeket, hogyan becsüli a kommunalitásokat és a faktorértékeket. Az összehasonlításra Gorsuch (1983) táblázatát mutatjuk be.

Az eljárás neve	Az eljárás lényege	Az egyediségek definiálása	A kommunalitás becslése	Faktorértékek
Főkomponens-elemzés	Maximalizálja a magyarázott varianciát	Nincs	Nem szükséges	Számított
Főfaktorelemzés	Maximalizálja a magyarázott varianciát	Specifikus faktorok, véletlen hiba	Számos becslő eljárás	Becsült
Image-elemzés	Minimalizálja a reziduális image-eket	Minden változó azon része, amely nem korrelál a többi változóval	Többszörös korreláció négyzete	Számított
Alfa-elemzés	Maximalizálja a megbízhatósági együtthatót	Pszichometrikus hiba	Iteratív	Számított
Maximum-Likelihood (Lawley, Jöreskog)	A reprodukált korrelációs mátrix legjobb becslése	Specifikus faktorok, véletlen hiba	Iteratív	Becsült
Maximum-Likelihood (Rao-féle kanonikus faktorelemzés)	A megfigyelt változókkal maximálisan korreláló faktorok	Specifikus faktorok, véletlen hiba	Iteratív	Becsült

12.1. táblázat. Faktorelemző eljárások összehasonlítása

A faktorelemző eljárások összehasonlításakor azt kell elsősorban megállapítani, hogy amennyiben a kommunalitások értéke 1,0, a főkomponens-elemzés, a főfaktorok módszere, az alfa-elemzés és a maximum likelihood eljárás azonos eredményre vezet.

Amennyiben a változók száma növekszik, a korrelációmátrix elemei között a diagonális elemek aránya, így a diagonális elemek fontossága is csökken, ezért a kommunalitás becslése helyett a diagonális elemeken kívüli korrelációs együtthatók becslése válik fontosabbá, vagyis a maximum likelihood, image- és alfa-faktorelemző eljárás megoldásai pontosabb eredményre vezetnek.

Az empirikus összehasonlításokban Tucker, Koopman és Linn (1969) azt tapasztalta, hogy a főkomponens- és a főfaktor-eljárás azonos faktorokat eredményezett, amikor minden változónál (20 változó) magas volt a kommunalitás. Kallina és Hartman (1976) nem talált interpretálható különbséget a főkomponens és a főfaktor között. Dziuban és Harris (1973) más vizsgálatokkal ellentétben nagy különbségeket talált több módszer faktormátrixában, sőt azt a következtetést vonták le, hogy a főkomponens-elemzés bizonyos feltételek esetén nem elfogadható eljárás. Azonban, ahogyan ezt Velicer (1977) bizonyította, ez a következtetés nem volt helyes, mivel abból adódott, hogy a faktormátrixban túlságosan sok faktor szerepelt (az 1-nél nagyobb sajátértékű faktorok maradtak meg), és ha más teszteket felhasználva, kevesebb faktorral végezték volna el az össze-

hasonlítást, akkor nagyon hasonló eredményekre vezettek volna a különböző módszerek. Velicer (1977) saját vizsgálatában, amelyben a főkomponens, image- és a maximum likelihood módszert hasonlította össze, azt a következtetést vonta le, hogy a három eljárás lényegében azonos eredményt ad, ezen belül a maximum likelihood és az image-eljárás eredménye hasonlított legjobban egymáshoz, és a főkomponens és a maximum likelihood között tapasztalt nagyobb különbséget. A főkomponens-eljárásnak nagy előnye az egyszerűség, mind számítási, mind elméleti értelemben. Velicer szerint az image-eljárás adhatja a legelfogadhatóbb alternatívát. Browne (1968) a különböző eljárások pontosságát vizsgálta ismert faktorsúly-mátrix esetén, és eredményül azt kapta, hogy a maximum likelihood becslés pontosabb megoldást adott, de ez a pontosság csak nagy mintánál ($n = 1500$) volt igazán érzékelhető.

Az irodalmi tapasztalatok tanúsága szerint (lásd Gorsuch, 1983, 123) általánosságban az állítható, hogy viszonylag nagy (nagyobb mint 30) változószám esetén és viszonylag elfogadható kommunalitások mellett (nem kisebbek, mint 0,4), bármelyik exploratív faktorelemző eljárás praktikusan hasonlóan értelmezhető eredményekre vezet.

De mit lehet mondani általánosságban, ha a változók száma nem elég nagy? Sajnos azt, hogy nincs valódi alternatíva. Vagy alkalmas hipotézisünk van a közös faktorokról, és alkalmazhatjuk valamelyik közös faktoreljárást, vagy a kommunalitások nagyságában bízva a főfaktor-eljárást.

A főfaktor-eljárás a legkisebb négyzetek módszere elve szerinti becslést ad, míg a kanonikus faktorelemző eljárás maximum likelihood eljárás, ha a változók többdimenziós normális eloszlásúak. A maximum likelihood becslés általában jobb eredményre vezet, ha a változók eloszlása meghatározott. A legkisebb négyzetes becslésnél nincs statisztikai próbafüggvény az illeszkedés vizsgálatára, míg a maximum likelihood módszernél könnyen számolható tesztünk van. A főfaktor-eljárás a legjobb r faktort határozza meg, amelyek a legnagyobb mértékben járulnak hozzá a megfigyelt változók varianciájához, míg a kanonikus faktorelemzés olyan faktorokat keres, amelyek maximálisan korrelálnak a megfigyelt változókkal, így a varianciák helyett a megfigyelt változók közötti korrelációk reprodukciójára koncentrál.

12.10. Faktorstruktúrák összehasonlítása azonos minták esetén

A faktorelemzések eredményét nemcsak akkor hasonlíthatjuk össze, ha ugyanazon mintán azonos változók struktúráját különböző faktorelemző eljárással vizsgáljuk. Érdekes lehet összehasonlítani két különböző változóhalmazt egy adott minta esetén, ha az a feltételezésünk, hogy a két változóhalmazra azonos faktorok hatnak. Két nem azonos változóhalmaz faktorainak összehasonlítását Harman (1960) nyomán a faktorok invarianciájának mérésével végezhetjük el.

Eszerint ha a két vizsgálat faktorértékeinek mátrixai ($\mathbf{S}_1$ és $\mathbf{S}_2$) között korrelációt számítunk:

$$\begin{aligned}\mathbf{R} &= \frac{1}{n}\begin{pmatrix}\mathbf{S}_1' \\ \mathbf{S}_2'\end{pmatrix}[\mathbf{S}_1, \mathbf{S}_2] \\ &= \frac{1}{n}\begin{pmatrix}\mathbf{S}_1'\mathbf{S}_1 & \mathbf{S}_1'\mathbf{S}_2 \\ \mathbf{S}_2'\mathbf{S}_1 & \mathbf{S}_2'\mathbf{S}_2\end{pmatrix} \\ &= \begin{pmatrix}\mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22}\end{pmatrix},\end{aligned} \tag{12.77}$$

akkor $\mathbf{R}_{12}(= \mathbf{R}_{21}')$ a két faktormegoldás kongruencia együtthatóit tartalmazza, ezek közvetlenül mérik a faktorok hasonlóságát. Könnyen belátható, hogy ha a két megoldás faktorait transzformáljuk úgy, hogy azok maximálisan egybeessenek (kongruáljanak), és a faktorok megoldásonként ortogonálisak, akkor a kongruencia együttható egyenlő a kanonikus korrelációval.

12.11. Faktorstruktúrák összehasonlítása különböző minták esetén

Két különböző vizsgálat (minta) egy-egy faktora közötti hasonlóságot mérhetjük a faktorértékek (factor scores) között számított korrelációs együtthatóval, amit ebben az esetben invariancia együtthatónak nevezünk. Henrysson (1957) nyomán nevezhetjük ezt konfigurációs invarianciának is, tekintettel arra, hogy a faktorértékek között számított korrelációs együttható nem érzékeny a faktorsúlyok nagyságrendbeli különbségeire.

A faktorstruktúra hasonlóságát mérhetjük a faktorsúlyok között számított korrelációs együtthatóval is. Problematikus azonban ennél a mutatónál, hogy ha pl. az egyik faktor súlyai 0,0 és 0,85 között mozognak, a másiké pedig $-0{,}85$ és 0,85 között, akkor a korreláció számításánál a két standardizált intervallumban a 0 faktorsúly azonos értéket kap az erős negatív $-0{,}85$ faktorsúllyal. Ezért javasolta Pinneau és Newhause (1964), hogy a korrelációt a faktorsúlyok négyzetei között számoljuk. Ez a javaslat azonban ugyanazt a problémát veti fel, ti. két ellentétes értelmű súly ($-0{,}85$ és 0,85) azonos értéket vesz fel a számítás során.

Egy másik kézenfekvő mutató a faktorsúlyok eltéréseiből számított négyzetes középérték. Jelöljük az első megoldás faktorsúlyait ${}_1\lambda_{jk}$-val, a második megoldás faktorsúlyait pedig ${}_2\lambda_{jt}$-vel. Ekkor az eltérés négyzetes középértéke:

$$h_{kt} = \sqrt{\frac{1}{m}\sum_{j}^{m}({}_1\lambda_{jk} - {}_2\lambda_{jt})^2}. \tag{12.78}$$

Ez a mutató közelít 0-hoz, ha a faktorsúlyok hasonlók, de a mutató egyszerűségéből adódóan nehéz megmondani, hogy milyen értéke fejezi ki a „jó megegyezést".

Tucker (1951) javasolta a faktorok hasonlóságának mérésére a kongruencia-együtthatót, amelynek értelmezési tartománya –1 és +1 közé esik.

A kongruencia-együttható

$$\psi_{kt} = \frac{\sum_j {}_1\lambda_{jk}\,{}_2\lambda_{jt}}{\sqrt{(\sum_j {}_1\lambda_{jk}^2)(\sum_j {}_1\lambda_{jt}^2)}}. \tag{12.79}$$

Látható, hogy a fenti együttható a szorzat momentum-együtthatóhoz hasonló, de nem egyenlő a korrelációs együtthatóval.

Általánosságban, ha az ${}_1\mathbf{\Lambda}$ az első megoldás $(m \times p_1)$ típusú, ${}_2\mathbf{\Lambda}$ a második megoldás $(m \times p_2)$ típusú faktorsúly-mátrixa, és a faktorsúly-mátrixok oszlopvektorait normáljuk a következő módon:

$${}_1\mathbf{\Lambda}^* = {}_1\mathbf{\Lambda} \begin{pmatrix} \frac{1}{\sqrt{\sum_i^m {}_1\lambda_{i1}^2}} & & \\ & \vdots & \\ & & \frac{1}{\sqrt{\sum_i^m {}_1\lambda_{ip_1}^2}} \end{pmatrix}, \tag{12.80}$$

és hasonlóan számítjuk ${}_2\mathbf{\Lambda}^*$-t, akkor a kongruencia-együttható $(p_1 \times p_2)$ típusú mátrixa:

$$\mathbf{\Psi} = {}_1\mathbf{\Lambda}^{*'}\,{}_2\mathbf{\Lambda}^*. \tag{12.81}$$

A kongruencia-együtthatók megbízhatóságának vizsgálatára szignifikanciapróba nem áll rendelkezésre, és az értelmezhetősége is bizonytalan lehet. Nézzük például négy változónak két különböző faktorsúlyait: 0,01, 0,03, 0,02, 0,04, és 0,91, 0,93, 0,92, 0,94. A két faktorstruktúra azonos, bár közöttük 0,90 konstans különbség van, a kongruencia-együttható pedig csak 0,92.

Nézzünk egy másik példát: a két faktor súlyai, 0,2, 0,3, 0,1, 0,3 és 0,3, 0,2, 0,4, 0,1. Ebben az esetben a kongruencia-együttható 0,70, és ez mutatja a gyengébb egyezést (strukturálisan).

Általában a kongruencia-együttható képletéből adódóan magasabb kongruencia-együtthatót kapunk azoknál a faktoroknál, amelyeknél a megfelelő súlyok azonos előjelűek.

Ha két vizsgálat (vagy a két különböző faktorelemző eljárás) minden faktorpárjára kiszámítjuk a kongruencia-együtthatókat, akkor a kongruencia-együtthatók mátrixához jutunk (lásd a (12.81) egyenletet). Tekintsük a továbbiakban azt az esetet, amikor a két vizsgálat során azonos számú faktort határozunk meg, így a kongruenciamátrix kvadratikus lesz – $(p \times p)$ típusú – lesz.

Természetesen nem várhatjuk feltétlenül el, hogy az azonos indexű faktorok hasonlítsanak legjobban egymáshoz. Ezért ha a két faktormegoldás hasonlóságát egy összegző mutatóval akarjuk kifejezni, figyelembe kell venni, hogy az első megoldás egyes faktorai a második megoldás mely faktoraihoz hasonlítanak legjobban. Célszerűnek látszik a második megoldás faktorsúly-mátrixát permutálni úgy, hogy az az első faktormegoldás faktoraihoz igazodjon a maximális hasonlóság értelmében. Jelöljük a permutációs mátrixot **P**-vel. Ha a két faktormátrix azonos indexű faktorai hasonlítanak legjobban egymáshoz, a **P** mátrix diagonális lesz (**P** = **I**). Amennyiben az első faktormátrix i-edik faktora a második faktormátrix j-edik faktorához hasonlít legjobban, akkor a permutáló mátrix i-edik és j-edik diagonális eleme 0 lesz, és az (i, j), (j, i) cellákba kerül 1. Akkor is permutálni kell a második faktormátrix oszlopvektorait, ha az összehasonlítandó faktorsúlyok előjelben különböznek egymástól. Így, ha a kongruenciamátrix i-edik sorában a j-edik elem lesz maximális abszolút értékű, de az együttható negatív előjelű, akkor a

második faktormátrix j-edik faktorát –1-gyel kell szorozni és az i-edik faktorral kicserélni. Ekkor a permutációs mátrix (i, j)-edik eleme 1, a (j, i)-edik eleme pedig –1 lesz. Ha az így összeállított permutációs-mátrixszal jobbról megszorozzuk a második faktormátrixot, akkor a két faktormátrix közvetlenül összehasonlíthatóvá válik.

Jelölje a permutált faktormátrixot ${}_2\mathbf{\Lambda}^* = {}_2\mathbf{\Lambda}\,\mathbf{P}$.

Az összehasonlítandó két faktormátrix különbségmátrixa:

$$\mathbf{E} = {}_1\mathbf{\Lambda} - {}_1\mathbf{\Lambda}^*, \tag{12.82}$$

és a különbözőség összegző mutatója:

$$\gamma = \text{trace}(\mathbf{E}'\mathbf{E}). \tag{12.83}$$

A γ mutató tehát a két faktormegoldás különbözőségének mértékét fejezi ki. Tökéletes egyezés esetén értéke 0 lesz, hasonló megoldásoknál közel 0.

A γ mutatót függetlenné tehetjük a változók és a faktorok számától, ha elosztjuk a változók és a faktorok számának szorzatával, így különböző vizsgálatokban eltérő változó- és faktorszámra kapott különbözőségeket közvetlenül össze tudunk hasonlítani:

$$\overline{\gamma} = \gamma/\text{mp} = \frac{1}{\text{mp}}\text{trace}(\mathbf{E}'\mathbf{E}). \tag{12.84}$$

Mérhetjük a különböző faktormegoldások hasonlóságát a kongruenciamátrix elemeitől függő mutatóval is. A kongruenciamátrix diagonális elemeinek átlaga lehet egy ilyen mérőszám. Természetesen a kongruenciamátrix sor, illetve oszlopmaximumai nem feltétlenül a diagonális elemei, így szükséges lehet a fentiekben már definiált permutációs mátrix felhasználásával átrendezni a kongruenciamátrix oszlopvektorait.

Jelölje $\mathbf{\Psi}^* = \mathbf{\Psi}\,\mathbf{P}$ a permutált kongruenciamátrixot.

Két faktormátrix hasonlóságát méri a $\mathbf{\Psi}^*$ kongruenciamátrix diagonális elemeinek átlaga:

$$\mathbf{\Psi} = \frac{1}{p}\text{trace}(\mathbf{\Psi}^*). \tag{12.85}$$

A mutató értékkészlete 0 és 1 között van, maximális egyezés esetén az 1 értéket veszi fel.

Egy másik mutató lehet a kongruenciamátrix determinánsa:

$$\delta = |\mathbf{\Psi}^*|. \tag{12.86}$$

Amennyiben a két faktormátrix faktorai tökéletesen illeszkednek egymáshoz, akkor a kongruenciamátrix determinánsa egyenlő lesz 1-gyel.

A faktorstruktúrák hasonlóságának mérésére javasolt mutatók alkalmazására a következőkben mutatunk példát.

12.12. Különböző faktorelemző eljárások empirikus összehasonlítása

A következő faktorelemzési példa az MTA Szociológiai Kutatóintézetének Értékszociológiai Műhelye által 1982-ben végzett életút-értékrendszer vizsgálat anyagából készült.

A minta országos, a felnőtt lakosságot reprezentáló minta, a megfigyelési elemszám: $n = 1464$.

A kérdés, amire a választ kértük, a következő volt:

„Néhány tulajdonságot sorolunk fel, amire nevelni lehet a gyerekeket. Melyeket tartja Ön különösen fontosnak? Kérem, válassza ki az öt legfontosabbat!"

A következő tulajdonságok szerepeltek:

(01) – jó magaviselet
(02) – udvariasság
(03) – önállóság
(04) – a kemény munka szeretete
(05) – őszinteség
(06) – felelősségérzet
(07) – türelem
(08) – képzelőerő, fantázia
(09) – mások tisztelete, tolerancia
(10) – vezetőkészség
(11) – önfegyelem
(12) – takarékosság
(13) – határozottság, állhatatosság
(14) – vallásos hit
(15) – önzetlenség
(16) – engedelmesség
(17) – hűség, lojalitás

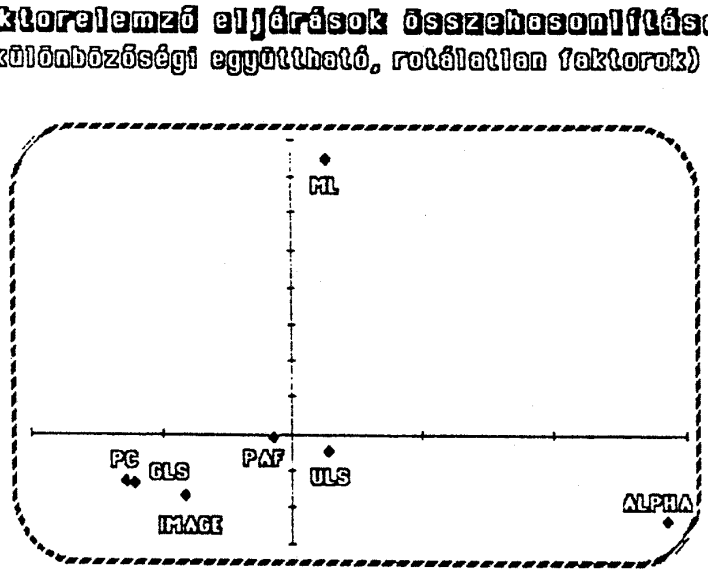
A válaszokat dichotom változókká alakítottuk, amelyek két kategóriája: választotta és nem választotta (kódjai 5 és 0). A válaszok mögött meghúzódó főbb latens értékválasztási dimenziók felderítésére a faktorelemzés eljárását alkalmaztuk. A számításokat az SPSS PC$^+$ programrendszerrel végeztük. A következőkben bemutatjuk a hét különböző faktorelemző eljárás eredményeit, majd a hét eljárás különbözőségét, ill. hasonlóságát a γ, ψ és δ mutatók alapján a MINISSA-módszer segítségével elemezzük. (A MINISSA-módszer és az ezen a néven ismert program az MDS(X) programcsomag egyik eljárása. Ezeket a számításokat az IBM 3031 számítógépen végeztük.)

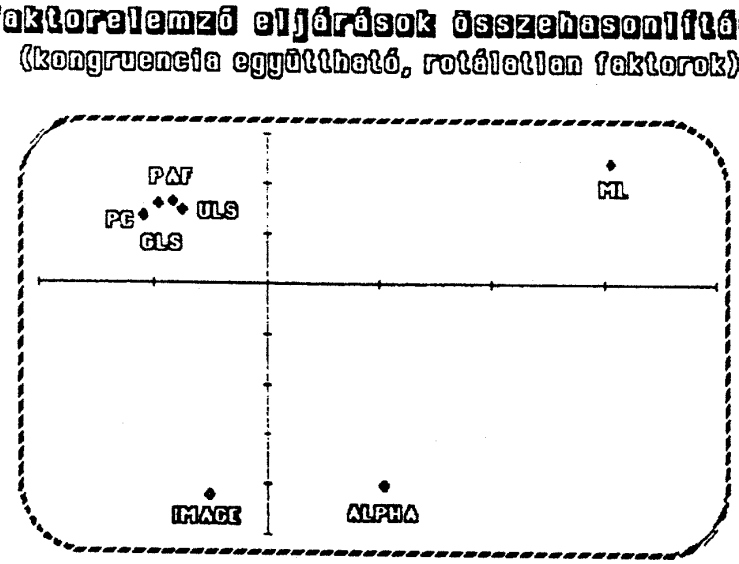
A hét faktorelemző eljárás:

1. PC Principal component analysis (főkomponens-elemzés)
2. PAF Principal axis factoring (Főfaktorelemzés)
3. ALPHA Alpha factoring (Alfa-faktorelemzés)
4. IMAGE Image factoring (Image-elemzés)
5. ULS Unweighted least squares (Legkisebb négyzetek módszere)
6. GLS Generalized least squares (Általánosított legkisebb négyzetek módszere)
7. ML Maximum likelihood (Maximum likelihood eljárás)

A hét faktoreljárás összehasonlításának eredményeit az 1–6 ábrák szemléletesen mutatják. Ezek a MINISSA-eljárás háromdimenziós megoldásainak az első két dimenzióját tartalmazzák. Tulajdonképpen a kétdimenziós megoldás illeszkedése is kitűnő volt, de néha túlságosan is közel, fedésbe kerültek az eljárásokat reprezentáló pontok azért, mert a harmadik dimenziónak a szórása is beleolvadt a kétdimenziós megoldásba. (Később még utalni fogunk arra, hogy mely módszert különített el leginkább a harmadik dimenzió.)

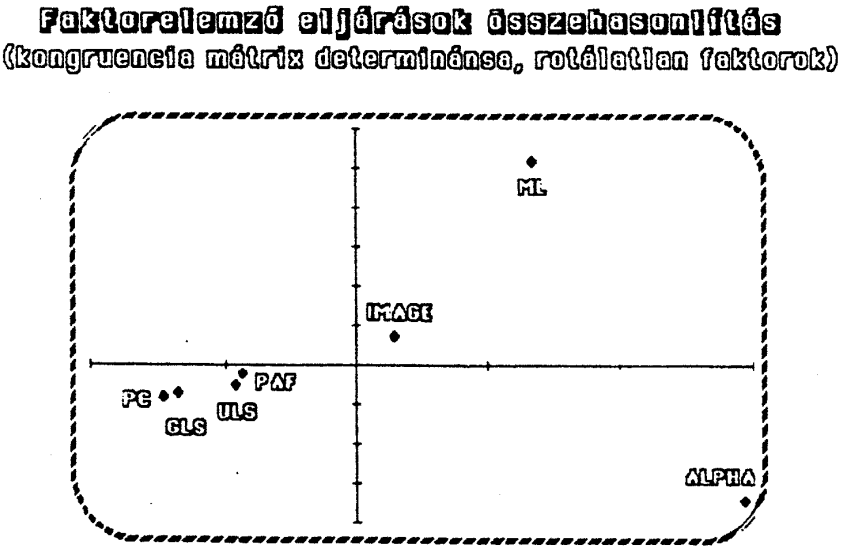
A háromdimenziós megoldás illeszkedését mérő mutató, a nyomaték (stress) értéke a különbözőségi, kongruenciaátlag és a kongruenciamátrix determinánsa együtthatóknál rendre a következő volt: a rotálatlan faktoroknál 0,0000021518, 0,0, 0,0000019681,

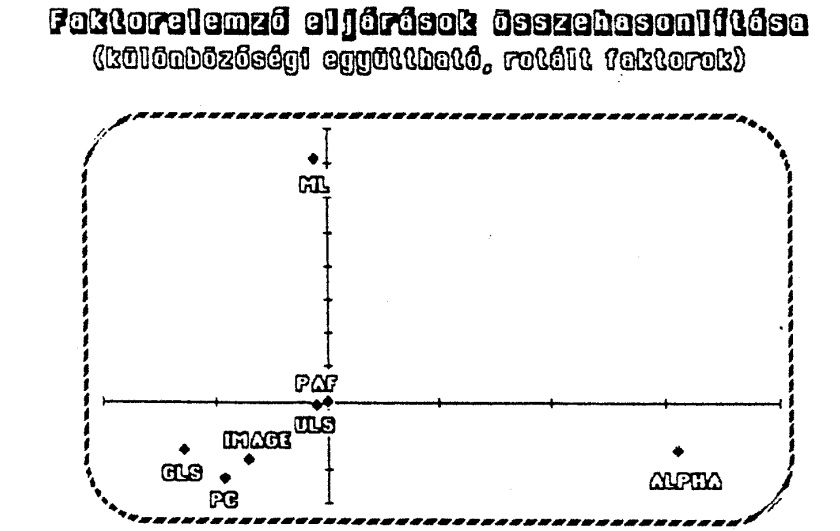




rotált faktoroknál 0,00053942, 0,0000018542, 0,0. Megállapíthatjuk, hogy a megoldások szinte tökéletes illeszkedést adtak.

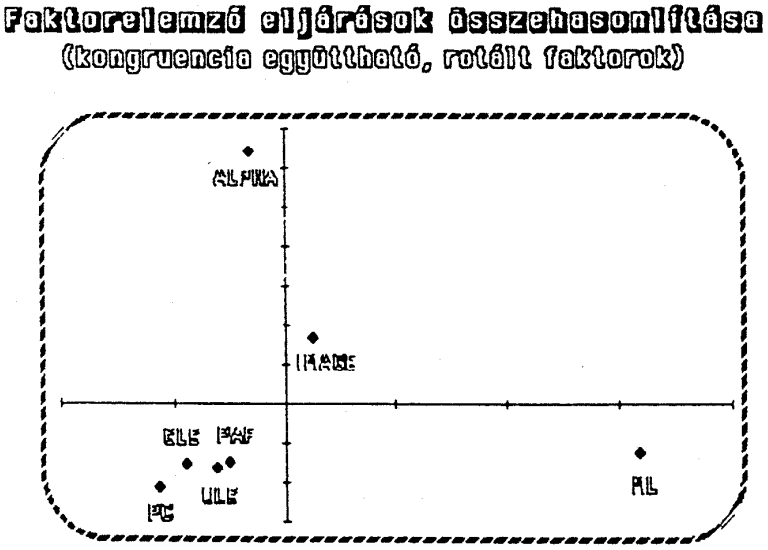
Összességében a leghasonlóbb, lényegében azonos megoldást adta a főkomponens (PC) és az általánosított legkisebb négyzetek (GLS) módszere (a negyedik tizedes jegyben voltak eltérések, néha a harmadikban), majdnem ugyanilyen mértékű egyezést találtunk a főfaktor (PAF) és a klasszikus legkisebb négyzetek módszere (ULS) között,

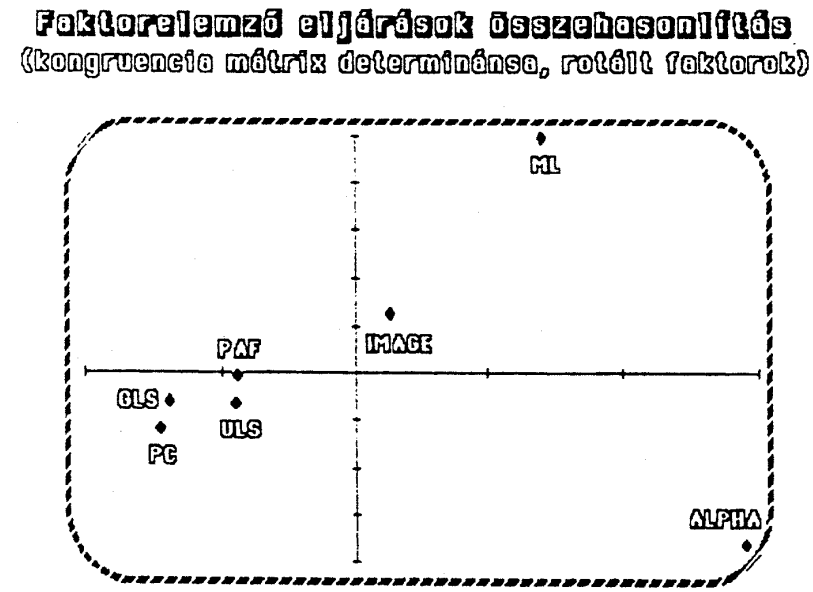




de ez a két páros egymáshoz is közel állt. A legnagyobb különbözőséget tapasztaltuk az alfa (ALPHA) és a maximum likelihood eljárás (ML) között, és ez a két módszer a többitől is hasonló mértékben tért el.

A fenti eredmény mindhárom vizsgált mutató esetében megegyező volt. Az image-eljárás (IMAGE) a PC, GLS, PAF, ULS és az ML, ALPHA módszerek között helyez-





kedett el, bár az átlagos kongruencia-együttható ítélete szerint az első csoporttól eltávolodott, és az ALPHA-eljáráshoz került közelebb (ez az eltérés azonban a rotációval gyengébbé vált).

Érdemes megemlíteni, hogy az IMAGE-eljárás a harmadik dimenzióban lényegesen elkülönült a többi eljárástól, és ez magyarázza azt, hogy a kétdimenziós vetületben a sík közepe tájékán mozgott – és ezért került relatíve közelebb a PC, GLS, PAF és ULS eljárásokhoz, vagyis az ML, az ALPHA és az IMAGE egymástól és a többi eljárástól is a legnagyobb különbözőséget adta. Meg kell említenünk, hogy a GLS-módszernek a várttól eltérő eredménye azzal magyarázható, hogy a vizsgált változók együttes eloszlása nem volt normális, még ha feltételezzük a választás folytonosságát, a durva mérés mögött akkor is csak egy erősen ferde, aszimmetrikus eloszlású változóhalmazt vizsgáltunk. A faktorok tartalmát a főkomponens-elemzés (PC) megoldásával mutatjuk be.

Először összefoglalóan megadjuk a legfőbb értékdichotómiákat, majd táblázatokba rendezve a faktorok súlyait.

Az első rotálatlan faktor a legfőbb értékpolaritást fejezi ki:

Modern személyiségértékek ⟷ Hagyományos közösségi értékek.

Ezt az alapvető értékdimenziót figyeltük meg a Rokeach-értékteszt elemzése során 1980-ban és 1982-ben is (intellektuális és autonómia-értékek álltak szemben a hagyományos közösségi és öröm-értékekkel; lásd Hankiss, et. al., 1983). Ebben a dimenzióban a „belülről irányított" ember típusa áll szemben a „tradícióktól irányított" ember tulajdonságaival.

A második rotálatlan faktor:

Munkaþ iztonságh it ⟷ Emberi kapcsolatok, autonómia.

A pozitív oldalon kétfajta érték keveredik: a munkával kapcsolatos és a vallásos értékek, ezen belül az elsőknek van nagyobb, meghatározóbb súlya. Ebben az összefüggésben a vallásos hit a kötelesség, felelősség, becsület fogalmaihoz kapcsolódik. A munka értéke bizonyos fajta biztonságra való törekvéssel fonódik össze: anyagi szempontból a takarékossággal, lelki szempontból a vallásos hittel és a hűséggel. A másik, negatív oldalon az emberi kapcsolatokat szabályozó tulajdonságok szerepelnek. Olyan értékek, amelyek a személyes autonómia fenntartása mellett a másik ember számára is könnyebbé, zökkenőmentesebbé teszik az együttélést.

A harmadik rotálatlan faktor dichotómiája:

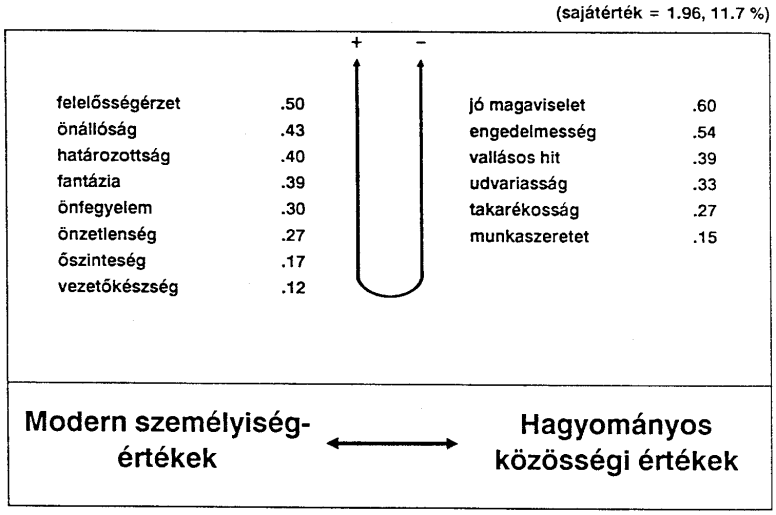
Hagyományos keresztény értékek ⟷ Protestáns értékek.

A pozitív póluson a hagyományos keresztény értékek szerepelnek, középpontban mások tisztelete, az alkalmazkodás értékei, mindössze a határozottság keveredik közéjük. Ennek a dimenziónak az ellentétes oldala az evilági élet, ahol az anyagi, protestáns értékek a fontosak.

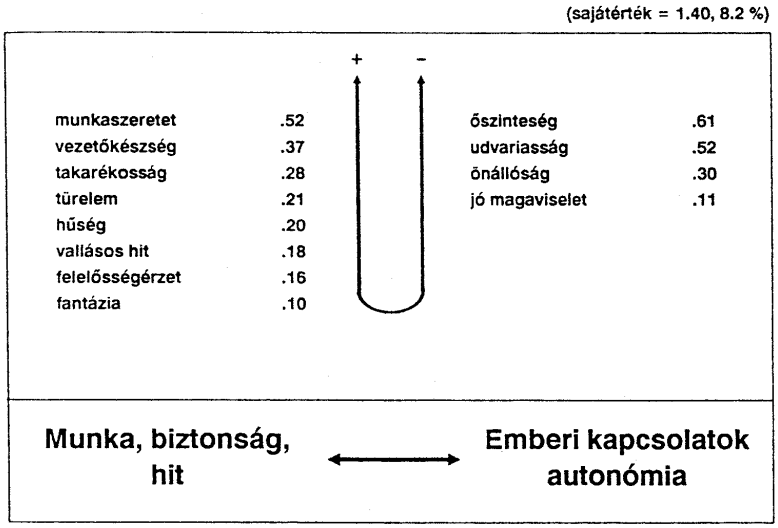
A harmadik rotált faktorban tisztábban jelenik meg az emberekhez, a társadalomhoz való igazodás kétféle változata, a hagyományos keresztény és a pragmatikusabb, materiálisabb, de az elfogadás fontosságát is valló szemlélet.

A továbbiakban a részletes faktormátrixokat közöljük.

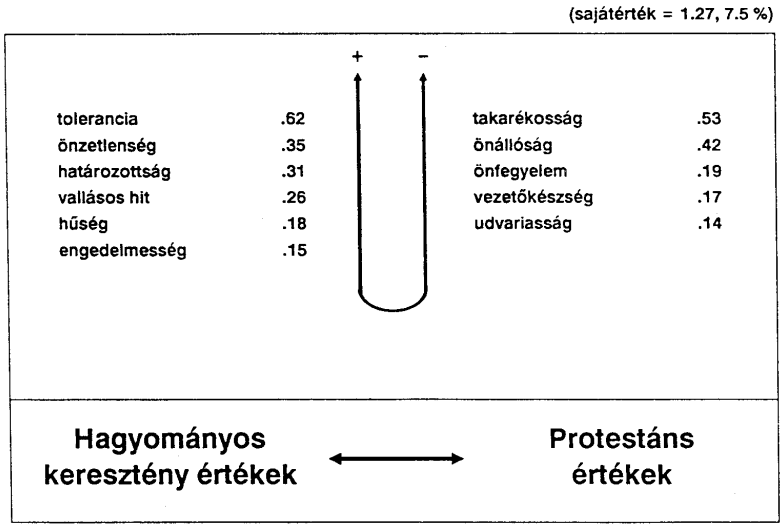
Gyermeknevelési elvek első, rotálatlan faktora (PC elemzés)



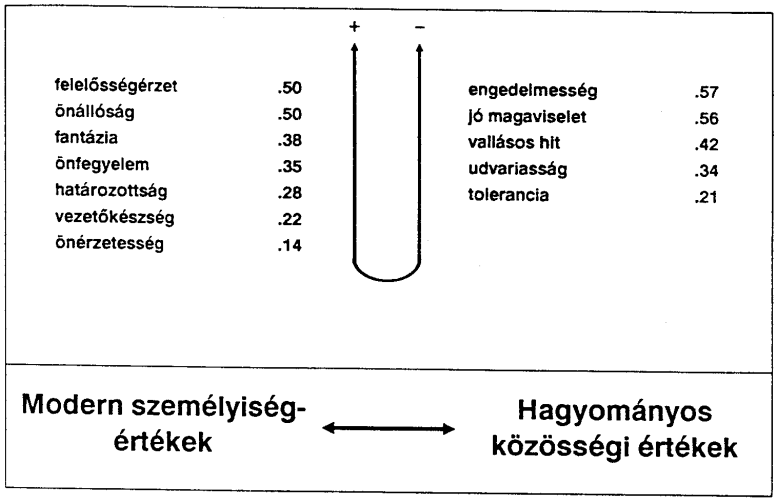
Gyermeknevelési elvek második, rotálatlan faktora (PC elemzés)



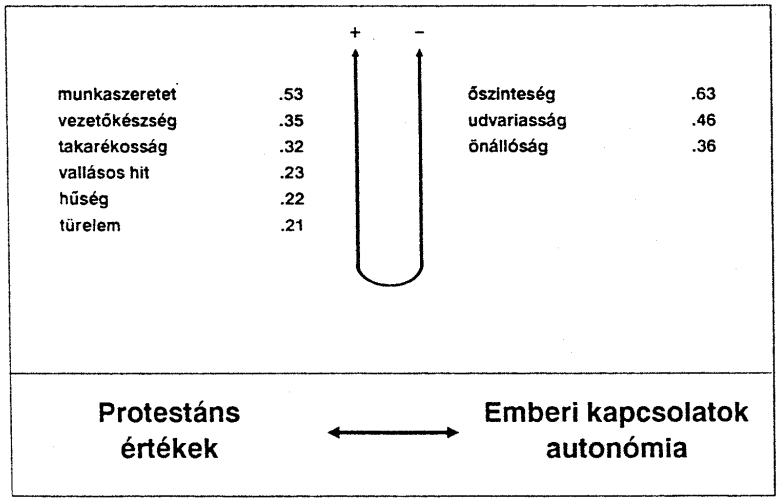
Gyermeknevelési elvek harmadik, rotálatlan faktora (PC elemzés)



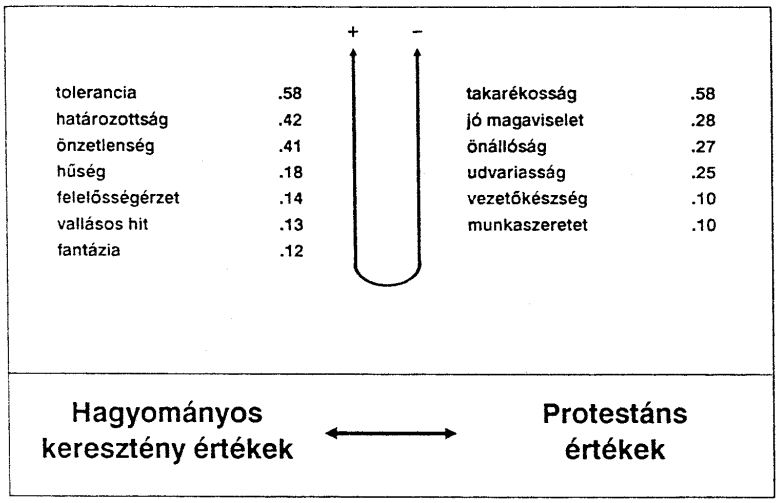
Gyermeknevelési elvek első, rotált faktora (PC elemzés)



Gyermeknevelési elvek második, rotált faktora (PC elemzés)



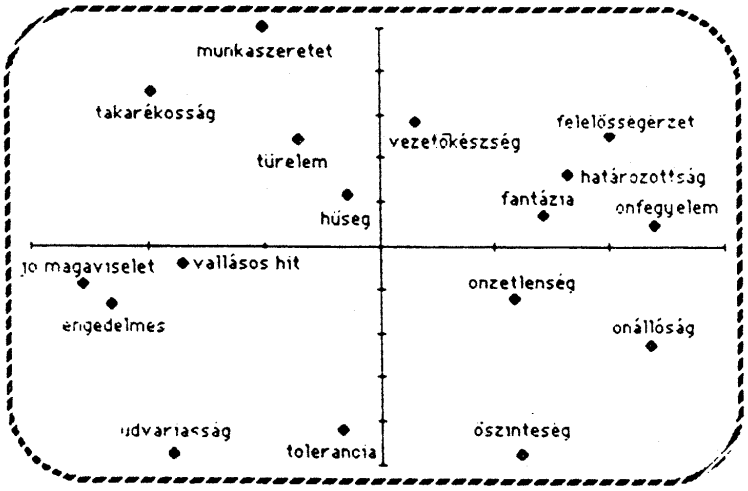
Gyermeknevelési elvek harmadik, rotált faktora (PC elemzés)



Gyermeknevelési elvek

(MINISSA eljárás, 2 dimenziós megoldás)

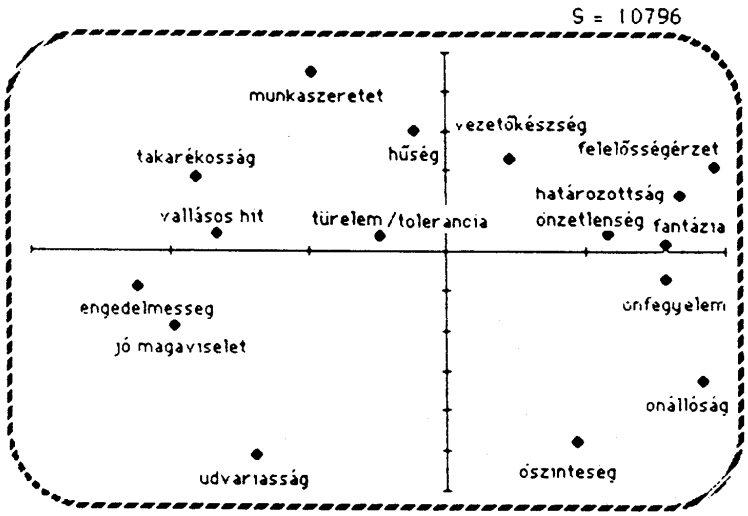
S = 1941

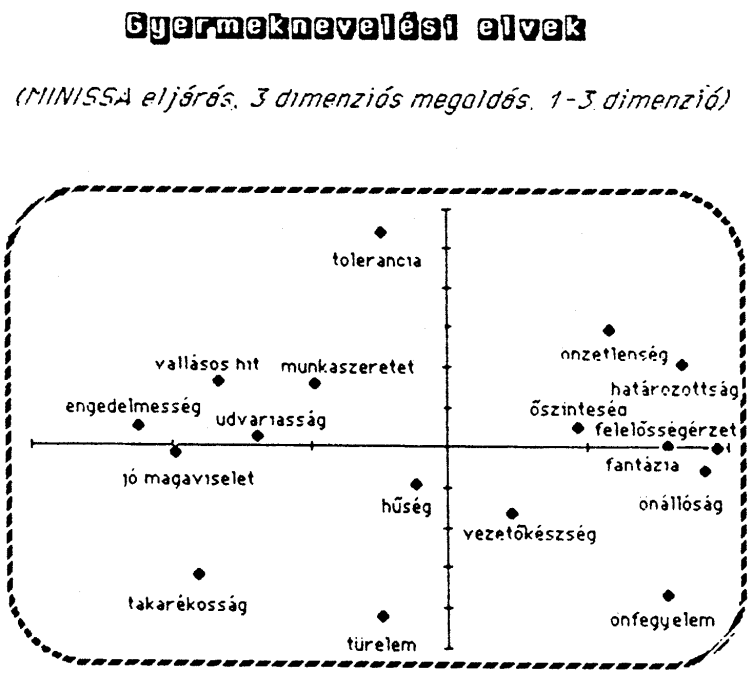


Gyermeknevelési elvek

(MINISSA eljárás, 3 dimenziós megoldás, 1-2 dimenzió)

S = 10796





Monte Carlo-vizsgálat

Az előző vizsgálatban egy empirikus elemzés mintabeli korrelációmátrixa alapján hasonlítottuk össze a kérdéses hét faktorelemző eljárást. A következőkben Monte Carlo-elemzéssel vizsgáljuk a különböző faktorelemző eljárásokat.

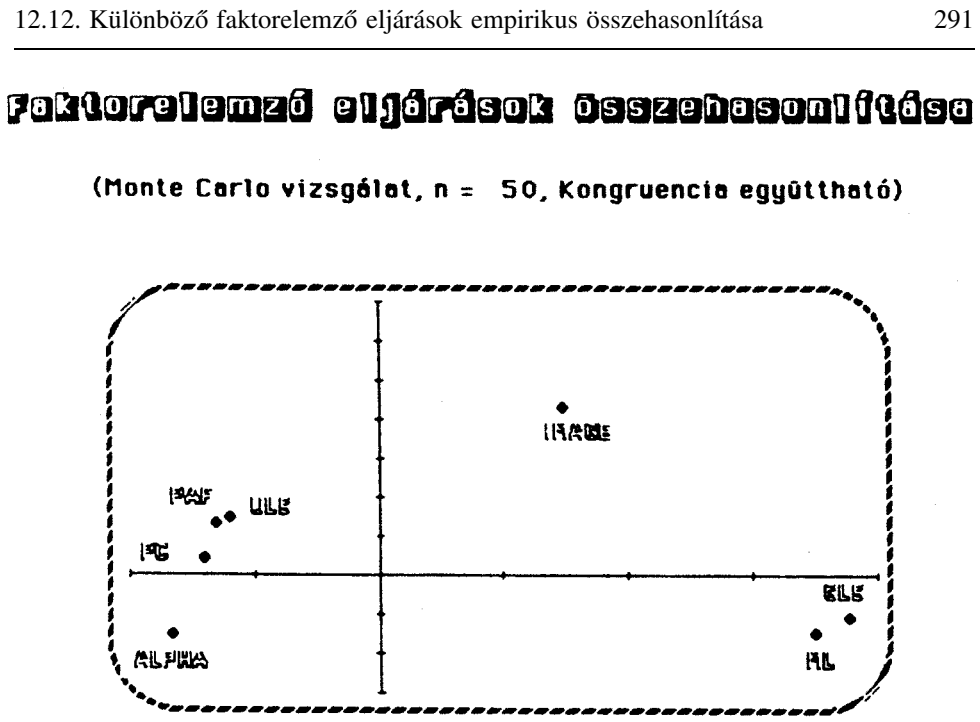
Kiindultunk két különböző faktorstruktúrából. A faktormátrixok alapján generáltunk az SPSS PC$^+$ programrendszer segítségével 50, 100, 200 és 400 elemű véletlen mintákat, amelyekre teljesült a változók normalitása. A négy különböző mintanagyságra azután elvégeztük a hét különböző faktorelemző eljárást, majd a kapott faktormátrixok hasonlóságát a γ (különbözőségi együttható), ψ (kongruencia-együttható) és a δ (a kongruenciamátrix determinánsa) mutatók alapján hasonlítottuk össze. A végeredményt a következő táblázatban közöljük.

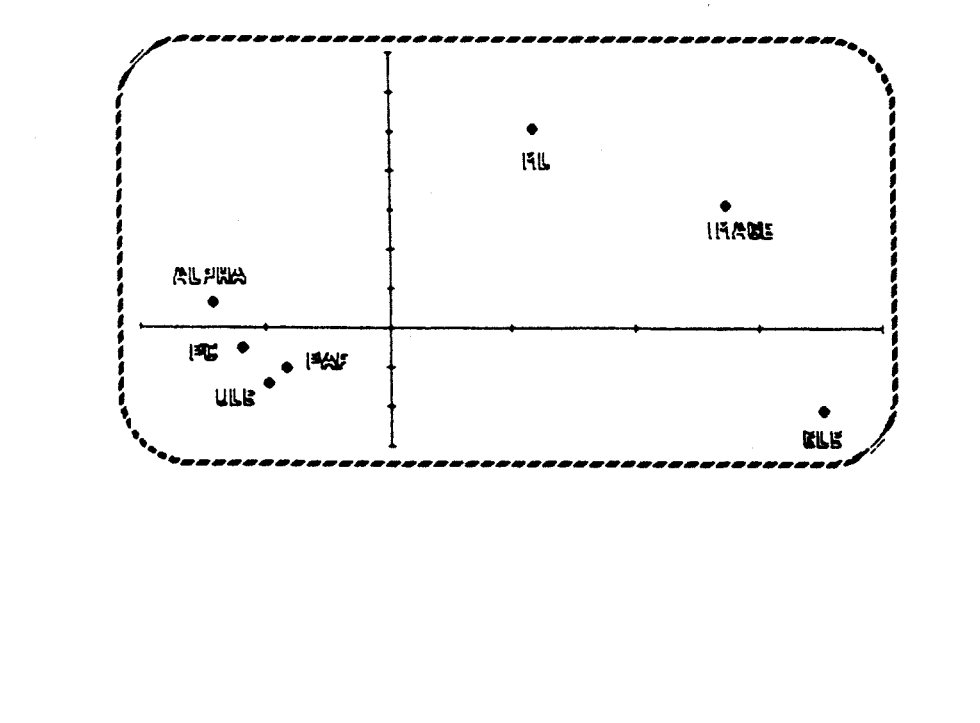
Faktorelemző eljárások összehasonlítása
két különböző faktorstruktúra és négy véletlen minta alapján
(a táblázatban azt jelöljük, hogy melyik faktorelemző eljárás adott a „várttól" eltérő eredményt)

Különböző faktor-struktúrák	Különbözőségi mutató	Kongruenciamutató	Kongruenciamátrix determinánsa
	50-es minta		
	1	2	3
1	OK	OK	ML, IMAGE
2	ML	ML	ML
	100-as minta		
	1	2	3
1	ML	ML	ML
2	ML	OK	ML
	200-as minta		
	1	2	3
1	OK	OK	ML, IMAGE
2	OK	OK	ML
	400-as minta		
	1	2	3
1	OK	OK	ML, IMAGE
2	OK	OK	ML

A hét faktorelemző eljárás hasonlóságait a 2-dimenziós MINISSA-térben ábrázolva jól láthatók a következő összefüggések (lásd a következő négy ábrát, amelyeket a kongruencia-együttható alapján készítettünk):

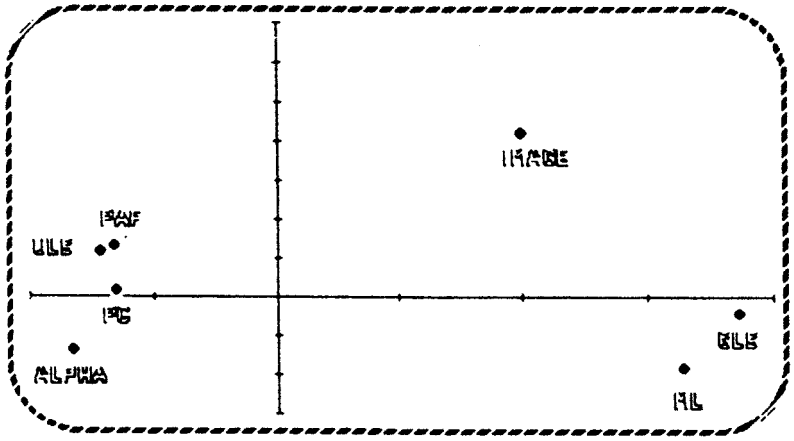
– lényegében azonos eredményeket adott a PC, ULS és PAF,
– az ALPHA is lényegében közeli struktúrát eredményezett, bár kissé elmozdult az ML irányába a második dimenzióban,
– a legnagyobb különbség a PC, ULS, PAF, ALPHA és az IMAGE, ML, GLS, módszerek között figyelhető meg, vagyis alapvető eltérést eszerinti csoportosításban találtunk,
– az IMAGE, ML, GLS csoporton belül az IMAGE-elemzés tért el legjobban a másik kettőtől is.


Faktorelemző eljárások összehasonlítása
(Monte Carlo vizsgálat, n = 50, Kongruencia együttható)
IMAGE
PAF
ULS
PC
ELS
ALPHA
ML


Faktorelemző eljárások összehasonlítása
(Monte Carlo vizsgálat, n = 100, Kongruencia együttható)
ML
IMAGE
ALPHA
PC
PAF
ULS
ELS

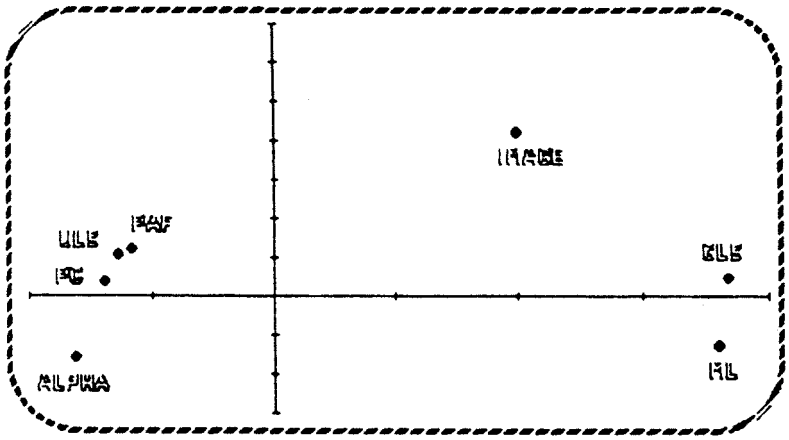
Faktorelemző eljárások összehasonlítása

(Monte Carlo vizsgálat, n = 200, Kongruencia együttható)



Faktorelemző eljárások összehasonlítása

(Monte Carlo vizsgálat, n = 400, Kongruencia együttható)

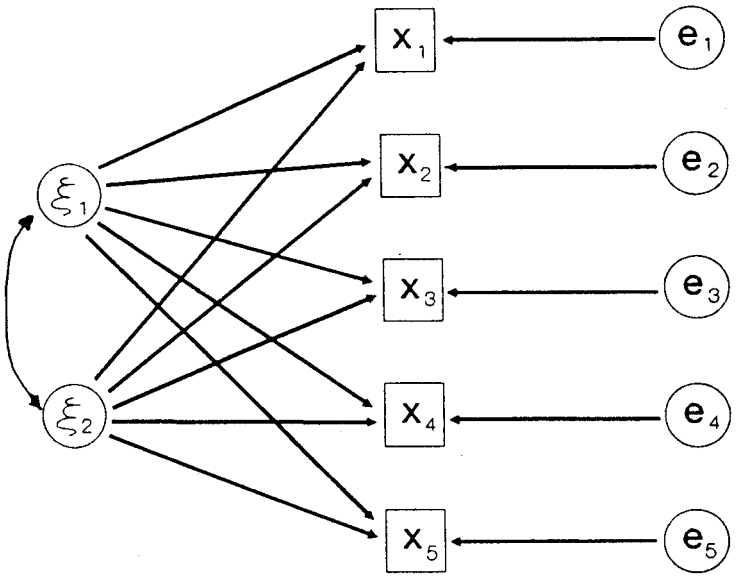


13. fejezet

Konfirmatív faktorelemzés

A faktorelemzést az előző fejezetben mint exploratív elemző eljárást tárgyaltuk. Abból az alapfeltételezésből indultunk ki, hogy a megfigyelt változók kifejezhetők latens változók, faktorok lineáris függvényeként. A manifeszt változók megfigyelt összefüggéseit a faktorok magyarázzák, így ha a faktorok hatásait a változókból kiszűrjük, akkor a változók korrelálatlanok lesznek.

Az exploratív faktorelemzés modelljét mutatja a következő ábra:



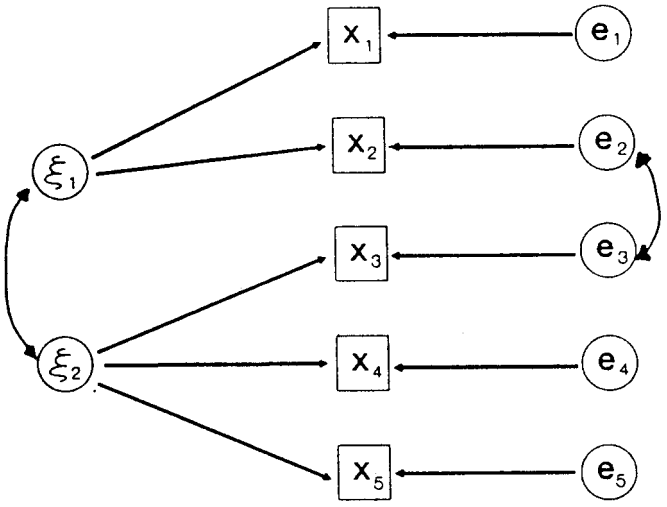
13.1. ábra. Az exploratív faktorelemzés modellje

Ahogyan a fenti ábrából is láthatjuk, az exploratív faktorelemzésnél feltételezzük, hogy

– a közös faktorok (ξ) korrelálnak egymással (vagy korrelálatlanok, más esetben),
– mindegyik közös faktor mindegyik változóra hatással van,
– az egyedi faktorok vagy hibakomponensek (e_i) korrelálatlanok egymással,
– mindegyik megfigyelt változóhoz tartozik egy egyedi faktor (hibakomponens),
– közös faktorok korrelálatlanok az egyedi faktorokkal.

Láthatjuk, hogy az exploratív faktorelemzés modellje nem tartalmaz feltételezéseket a modell struktúrájára, nincs *a priori* hipotézisünk, hogy a közös faktorok mely manifeszt változókra hatnak és melyekre nem, a közös faktorok között melyek vannak kapcsolatban egymással és melyek nem. Ha van *a priori* ismeretünk a manifeszt és a latens változók

közötti összefüggésekről, akkor ezeket beépíthetjük a modellbe, és az így felállított konfirmatív faktorelemzés modelljét illesztjük az adatokhoz. A következő ábra egy ilyen modellt mutat:



13.2. ábra. A konfirmatív faktorelemzés modellje

A konfirmatív faktorelemzés modelljében feltételezhetjük, hogy
– mely közös faktorpárok korrelálnak egymással,
– melyik közös faktor melyik megfigyelt változóra hat,
– melyik megfigyelt változóra hat egyedi faktor,
– az egyedi faktorok közül melyek korrelálnak egymással.

A fenti típusú feltételekkel felállított konfirmatív modell illeszkedését a megfigyelt adatokhoz statisztikai próba segítségével vizsgálhatjuk.

A konfirmatív faktorelemzést Jöreskog (1967, 1969), Jöreskog és Lawley (1968) fejlesztette ki, és Jöreskog nevéhez fűződik a LISREL (Linear Structural Relationships by the Method of Maximum Likelihood) számítógépes program is. A későbbiekben látni fogjuk, hogy a LISREL alkalmas a közös faktorok közötti strukturális kapcsolatok elemzésére is.

Matematikai jelölésekkel a modell a következőképpen írható fel:

$$\mathbf{x} = \mathbf{\Lambda}\,\mathbf{f} + \mathbf{e}, \tag{13.1}$$

ahol $\mathbf{x}$ a megfigyelt változók vektora ($m \times 1$ típusú), $\mathbf{f}$ a közös faktorok vektora ($r \times 1$ típusú), $\mathbf{e}$ az egyedi faktorokat vagy hibakomponenseket tartalmazza ($m \times 1$ típusú),

$\mathbf{\Lambda}$ a faktorsúlyok mátrixa ($m \times r$ típusú).

Feltételezzük, hogy $E(\mathbf{x}) = E(\mathbf{f}) = E(\mathbf{e}) = \mathbf{0}$, vagyis mind a megfigyelt, mind a latens változót a várható értéküktől való eltérésükkel jellemezzük. Feltételezzük továbbá, hogy $E(\mathbf{f}\,\mathbf{e}') = \mathbf{0}$, vagyis a közös faktorok és az egyedi faktorok (hibatagok) korrelálatlanok.

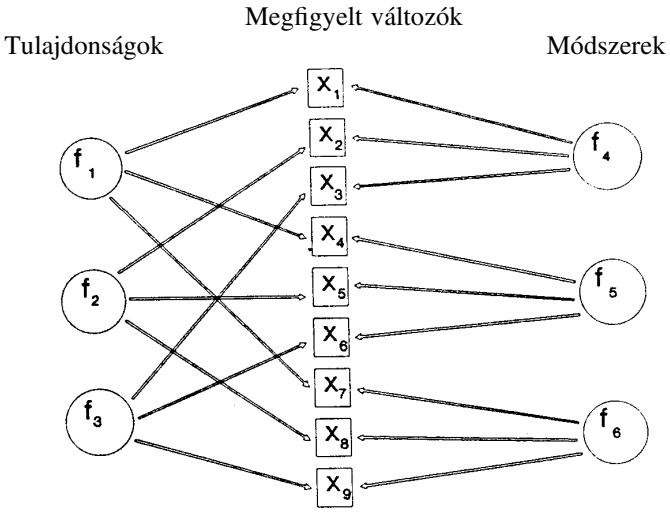
A modell feltétele alapján a megfigyelt változók variancia-kovarianciamátrixa ($\mathbf{\Sigma}$) a következőképpen fejezhető ki:

$$\mathbf{\Sigma} = \mathbf{\Lambda}\,\mathbf{\Phi}\,\mathbf{\Lambda}' + \mathbf{\Theta}, \tag{13.2}$$

ahol Φ a közös faktorok variancia-kovarianciamátrixa ($r \times r$ típusú), melyre $\Phi = E(\mathbf{f}\,\mathbf{f}')$, Θ az egyedi faktorok vagy hibatagok variancia-kovarianciamátrixa, amelyre $\Theta = E(\mathbf{e}\,\mathbf{e}')$.

A konfirmatív faktorelemző modellben a Λ, Φ és Θ paramétermátrixok elemeire teszünk feltételezéseket.

Például tekintsük a több módszer – több tulajdonság (Multi-Method Multi-Trait, MMMT-modellt), amelyben feltételezzük, hogy a latens jellemzőket, tulajdonságokat (trait) több módszerrel mérjük. Ha feltételezünk három latens jellemzőt (ezeknek megfelel az f_1, f_2, f_3 faktor) és mindegyiket három módszerrel mérjük (a módszereknek megfelelő faktorok f_4, f_5, f_6), valamint kilenc megfigyelt változónk közül x_1, x_2 és x_3 a megfelelő latens tulajdonság (trait) manifeszt indikátora az első módszerrel mérve, és az x_4, x_5, x_6 és x_7, x_8, x_9 változók a három latens trait indikátorai a másik két módszerrel mérve, akkor az MMMT-modell ábrája és együtthatómátrixai a következőképpen írhatók fel:



13.3. ábra. Az MMMT- (Multi-Method Multi-Trait) modell

A faktorsúlyok mátrixa:

$$\Lambda = \begin{array}{c} \begin{array}{cccccc} \multicolumn{3}{c}{\text{trait súlyok}} & \multicolumn{3}{c}{\text{módszersúlyok}} & \\ f_1 & f_2 & f_3 & f_4 & f_5 & f_6 & \end{array} \\ \left(\begin{array}{cccccc} \lambda_{11} & 0 & 0 & \lambda_{14} & 0 & 0 \\ 0 & \lambda_{22} & 0 & \lambda_{24} & 0 & 0 \\ 0 & 0 & \lambda_{33} & \lambda_{34} & 0 & 0 \\ \lambda_{41} & 0 & 0 & 0 & \lambda_{45} & 0 \\ 0 & \lambda_{52} & 0 & 0 & \lambda_{55} & 0 \\ 0 & 0 & \lambda_{63} & 0 & \lambda_{65} & 0 \\ \lambda_{71} & 0 & 0 & 0 & 0 & \lambda_{76} \\ 0 & \lambda_{82} & 0 & 0 & 0 & \lambda_{86} \\ 0 & 0 & \lambda_{93} & 0 & 0 & \lambda_{96} \end{array}\right) \begin{array}{c} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \\ x_9 \end{array} \end{array}$$

A faktorok variancia-kovarianciamátrixa

$$\underset{(6\times 6)}{\mathbf{\Phi}} = \begin{pmatrix} \text{trait/trait} \\ (3 \times 3) & \text{trait/módszer} \\ & (3 \times 3) \\ \text{módszer/trait} & \text{módszer/módszer} \\ (3 \times 3) & (3 \times 3) \end{pmatrix}.$$

A faktorok közötti kovarianciák közül az első részmátrixban találhatók a trait-trait-kovarianciák, amelyek mentesek a módszerhatástól. A változók közötti kovarianciákat könnyen felbonthatjuk a trait és a módszer hatásaival összefüggő tagokra.

Pl.

$$E(x_1, x_2) = \sigma_{12} = \lambda_{11}\lambda_{22}\phi_{12} + \lambda_{14}\lambda_{22}\phi_{24} + \\ + \lambda_{11}\lambda_{24}\phi_{14} + \lambda_{14}\lambda_{24}\phi_{44}.$$

Az MMMT-modellben a mérési módszert elkülönítjük a latens jellemző szubjektív vonásától (trait-től), és mindkét szisztematikus latens komponens hatását (faktorsúlyait) a konfirmatív faktorelemzéssel vizsgálhatjuk.

13.1. A faktormodell identifikálhatósága

A modell paramétereinek a becslését mindig meg kell hogy előzze a modell identifikálhatóságának vizsgálata.

A modell identifikálhatósága a modell paramétereinek egyértelmű meghatározhatóságát jelenti. Miután a $\mathbf{\Sigma}$ variancia-kovarianciamátrix a $\mathbf{\Lambda}$, $\mathbf{\Phi}$, $\mathbf{\Theta}$ paraméterek függvénye, a paraméterek akkor határozhatók meg egyértelműen, ha ez a függvény egyértékű függvény, vagyis $\mathbf{\Sigma}$-t egy és csakis egy paraméterhalmaz (paraméterstruktúra) generálja. Ha egy modell nem identifikálható, akkor a modell $\mathbf{\Lambda}$, $\mathbf{\Phi}$, és $\mathbf{\Theta}$ paraméterei a $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}' + \mathbf{\Theta}$ kovariancia-egyenletből ismert $\mathbf{\Sigma}$ esetén sem határozhatók meg egyértelműen, vagyis több paraméterstruktúrához is tartozhat ugyanaz a $\mathbf{\Sigma}$ mátrix.

Az identifikálhatóság alapvetően a megfigyelt adatokból származó információ (a faktormodellben a variancia-kovarianciamátrix független elemei) és a modell független paraméterei számának különbségével függ össze. A modell alulidentifikált (nem identifikálható), ha a paraméterei nem határozhatók meg egyértelműen. Az éppen identifikálható modellben a paraméterek száma pontosan megegyezik a megfigyelt varianciák és kovarianciák számával. A modell paraméterei ilyenkor becsülhetők (a paraméterek egyértelműen meghatározhatók), azonban egy ilyen modellnek nincs diszkriminatív ereje, tudományosan nem érdekes, mivel soha nem lehet elvetni, minden mintához illeszthető. Az a modell érdekes, amelyik túlidentifikált, paramétereinek (a független paramétereknek) száma kevesebb, mint a megfigyelt varianciák és kovarianciák száma, amelynek paraméterei becsülhetők, és amely statisztikai próba segítségével elvethető vagy elfogadható.

Könnyen belátható, hogy az exploratív faktorelemzés modellje nem identifikálható. Kiindulva az $\mathbf{x} = \mathbf{\Lambda f} + \mathbf{e}$ és $\mathbf{\Lambda}\,\mathbf{\Phi}\,\mathbf{\Lambda}' + \mathbf{\Theta}$ egyenletekből és a $\mathbf{\Lambda}$, $\mathbf{\Phi}$ és $\mathbf{\Theta}$ paramétermátrixokból, bármely ortogonális mátrixszal ($\mathbf{M}$, $(r \times r)$ típusú) a faktorokat és faktorsúlymátrixot transzformálhatjuk anélkül, hogy a modell két egyenletét megsértenénk.

Jelölje a transzformált faktorokat $\mathbf{f}^* = \mathbf{M}\,\mathbf{f}$, a transzformált faktorsúlyokat $\mathbf{\Lambda}^* = \mathbf{\Lambda}\,\mathbf{M}'$, a transzformált faktorok variancia-kovarianciamátrixát $\mathbf{\Phi}^* = \mathbf{M}\,\mathbf{\Phi}\,\mathbf{M}'$, ekkor

könnyen belátható, hogy mind a Λ, Φ, Θ, mind a Λ^*, Φ^* Θ mátrixok kielégítik a faktormodell alapegyenleteit:

$$\begin{aligned}\mathbf{x} = \mathbf{\Lambda}^*\mathbf{f}^* + \mathbf{e} &= (\mathbf{\Lambda}\,\mathbf{M}')\mathbf{M}\,\mathbf{f} + \mathbf{e} \\ &= \mathbf{\Lambda}(\mathbf{M}'\,\mathbf{M})\mathbf{f} + \mathbf{e} \\ &= \mathbf{\Lambda}\,\mathbf{f} + \mathbf{e},\end{aligned} \tag{13.3}$$

és

$$\begin{aligned}\mathbf{\Sigma} &= \mathbf{\Lambda}^*\,\mathbf{\Phi}^*\,\mathbf{\Lambda}^{*'} + \mathbf{\Theta} \\ &= (\mathbf{\Lambda}\,\mathbf{M}')(\mathbf{M}\,\mathbf{\Phi}\,\mathbf{M}')(\mathbf{M}\,\mathbf{\Lambda}') + \mathbf{\Theta} \\ &= \mathbf{\Lambda}(\mathbf{M}'\,\mathbf{M})\mathbf{\Phi}(\mathbf{M}'\,\mathbf{M})\mathbf{\Lambda}' + \mathbf{\Theta} \\ &= \mathbf{\Lambda}\,\mathbf{\Phi}\,\mathbf{\Lambda}' + \mathbf{\Theta}.\end{aligned} \tag{13.4}$$

A faktorok ilyen transzformációs lehetőségét a faktorok rotálásánál tárgyaljuk még. A faktorok rotálása az exploratív faktorelemzésnél a faktorok egyértelműbb értelmezését segíti. A fentiekben láthattuk, hogy pótlólagos feltételezések nélkül a faktorelemzés modellje nem oldható meg egyértelműen. Feltételezve a faktorok korrelálatlanságát és a véletlen komponensek korrelálatlanságát, megmutatható (lásd Lawley és Maxwell, 1971, 4. fejezet), hogy az exploratív faktorelemzésben az egyértelmű megoldás feltétele, hogy $\mathbf{\Lambda}'\,\mathbf{\Theta}^{-1}\,\mathbf{\Lambda}$ diagonális legyen. $\mathbf{\Lambda}'\,\mathbf{\Theta}^{-1}\,\mathbf{\Lambda}$ diagonális elemét a következőképpen értelmezhetjük. Ha a hibakomponensek varianciája egységnyi, akkor λ_{ik}^2/Θ_i az i-edik változó varianciájának a k-adik faktorral összefüggő részét fejezi ki. Ha figyelembe vesszük az összes változót, akkor a megfigyelt változók összes varianciájához a k-adik faktor $\Sigma_i(\lambda_{ij}^2/\Theta_i)$ mértékben járul hozzá. Ez a $\mathbf{\Lambda}'\,\mathbf{\Theta}^{-1}\,\mathbf{\Lambda}$ mátrix k-adik diagonális eleme. A faktorokat úgy választjuk, hogy az első faktor maximális mértékben járuljon hozzá a megfigyelt változók együttes varianciájához, a második faktor a maradék lehetőség közül magyarázza maximálisan a megfigyelt változók varianciáit, miközben korrelálatlan az első faktorral (ha Φ egységmátrix), és így tovább.

Az identifikálhatóság szükséges feltételét a független kovariancia-egyenletek és a független paraméterek számának az összehasonlításával határozhatjuk meg.

A kovariancia-egyenletek száma ((13.4) egyenlet) $\frac{1}{2}m(m+1)$.

A független paraméterek száma az r-faktoros modellben $mr + m - \frac{1}{2}r(r-1)$, ha Φ egységmátrix.

A különbség a kovarianciamátrix elemei és a független paraméterek száma között:

$$\begin{aligned}s &= \frac{1}{2}m(m+1) - \left\{mr + m - \frac{1}{2}r(r-1)\right\} \\ &= \frac{1}{2}[(m-r)^2 - (m+r)].\end{aligned} \tag{13.5}$$

Az exploratív faktorelemzés identifikálhatóságának szükséges feltétele, hogy $s \geq 0$.

A (13.5) egyenlet az identifikálható modellben a faktorok számára ad felső korlátot. Eszerint a faktorok száma nem lehet nagyobb, mint

$$r \leq \{2m + 1 - (8m+1)^{\frac{1}{2}}\} \tag{13.6}$$

A (13.5) feltétel az identifikálhatóságnak csak szükséges feltétele, és nem biztosítja, hogy a véletlen komponensek varianciái (Θ elemei) becslései ne legyenek negatívak. Kano (1986) közöl szükséges és elégséges feltételt a modell identifikálhatóságára a maximum likelihood és az általánosított legkisebb négyzetek becslési módszerek esetén.

A konfirmatív faktorelemzésnél az identifikálhatóság szükséges feltételénél nem tételezhetjük fel általánosan, hogy a Φ mátrix egységmátrix, így a Λ, Φ és Θ paramétermátrixok összesen $mr + \frac{1}{2}r(r+1) + \frac{1}{2}m(m+1)$ elemet tartalmaznak.

A konfirmatív faktorelemzés identifikálhatóságának szükséges feltétele:

$$\frac{1}{2}m(m+1) - \left\{mr + \frac{1}{2}r(r+1) + \frac{1}{2}m(m+1)\right\} \geq 0.$$

A paraméterek lehetséges számából természetesen le kell vonni az adott modell kötött paramétereinek számát, és a szabad független paraméterek számát kell kivonni a variancia-kovarianciamátrix független elemeinek számából.

Jöreskog és Sörbom (1981) az identifikáció vizsgálatát beépítették számítógépes programjukba (LISREL). Az ún. információs mátrix alapján – amely a maximum likelihood becslő függvény paraméterek szerinti második parciális deriváltjait tartalmazza, és függ a paraméterek becsléseinek variancia-kovarianciáitól – azt lehet mondani, hogy ha az információs mátrix pozitív definit, akkor a modell (majdnem biztosan) identifikálható, és ha az információs mátrix szinguláris, akkor a modell nem identifikálható (Jöreskog és Sörbom, 1981, 11. old.).

13.2. Skála-invariancia

A gyakorlatban a faktormodellt általában a megfigyelt változók mintabeli korrelációmátrixához illesztjük a variancia-kovariancia helyett. Ennek az oka elsősorban az, hogy a manifeszt változók mértékegysége gyakran önkényes a társadalomtudományokban, így indokoltabb a változókat standardizálni a mintabeli szórásokkal.

Ha a változók mérésénél a mértékegységeket változtatjuk, akkor ez hatással van a változók szórásaira. Ha a szórásokkal standardizált változók között számítjuk a kovarianciákat, akkor azok a korrelációs együtthatókkal lesznek egyenlők. Ez biztosítja, hogy az x változók mértékegységeinek (skálájának) változtatása nincs hatással az eredményekre. Azokban a korrelációs együtthatók együttes eloszlása nem azonos a kovarianciák együttes eloszlásával, így nem nyilvánvaló, hogy a becslések a valódi maximum likelihood becsléseket adják-e, és hogy az illeszkedés jóságát mérő statisztika érvényes-e.

Induljunk ki a faktorelemzés alapegyenleteiből, és nézzük meg, hogy a mérési skála mértékegységének változtatása hogyan változtatja meg a modellt. Legyen $\mathbf{C}$ diagonális mátrix, diagonális elemei legyenek pozitívok, és jelentsék az egyes változók skálatranszformációit, vagyis

$$\mathbf{x}^* = \mathbf{C}\,\mathbf{x}.$$

Az új, transzformált modell:

$$\mathbf{x}^* = \mathbf{C}\,\mathbf{\Lambda}\,\mathbf{f} + \mathbf{C}\,\mathbf{e} \tag{13.7}$$

$$\operatorname{cov}(\mathbf{x}^*) = \mathbf{\Sigma}^* = \mathbf{C}\,\mathbf{\Lambda}\,\mathbf{\Lambda}'\,\mathbf{C} + \mathbf{C}\,\mathbf{\Theta}\,\mathbf{C}.$$

Ha $\mathbf{C} = (\operatorname{diag}\mathbf{\Sigma})^{-\frac{1}{2}}$, vagyis az x változókat a szórásukkal osztjuk, akkor a (13.7) egyenletet a következőképpen írhatjuk:

$$\mathbf{x}^* = \mathbf{\Lambda}^*\mathbf{f} + \mathbf{e}^*, \tag{13.8}$$

és

$$\mathbf{\Sigma}^* = \mathbf{\Lambda}^*\mathbf{\Lambda}^{*'} + \mathbf{\Theta}^*,$$

ahol $\mathbf{\Lambda}^* = \mathbf{C}\,\mathbf{\Lambda}$ és $\mathbf{\Theta}^* = \mathbf{C}\,\mathbf{\Theta}\,\mathbf{C}$.

A fentiekből látható, hogy a $\mathbf{x}$ mérési skálaegységeinek változtatása nincs hatással $\mathbf{x}^*$ értékeire, és így a Λ^* és Θ^* paraméterek becslése is független $\mathbf{x}$ mérési skálájától, így ezek a paraméterek skála-invariánsak.

A Λ^* és Θ^* paramétereket becsülhetjük úgy, hogy először becsüljük a Λ és Θ paramétereket, és ezek alapján határozzuk meg Λ^* és Θ^* becsléseit:

$$\widehat{\Lambda}^* = (\operatorname{diag}\widehat{\Sigma})^{-\frac{1}{2}}\widehat{\Lambda} \quad \text{és} \quad \widehat{\Theta}^* = (\operatorname{diag}\widehat{\Sigma})^{-\frac{1}{2}}\widehat{\Theta}(\operatorname{diag}\widehat{\Sigma})^{-\frac{1}{2}}.$$

Egy becslési eljárás skálafüggetlen, ha a modell paramétereitől függő Σ mátrix és a mintabeli $\mathbf{S}$ mátrix illeszkedését mérő függvény minimuma nem függ a változók skálájától. Eszerint skálafüggetlen becslési eljárásnál az illesztett függvény minimuma megegyezik akár a kovarianciamátrixot, akár a korrelációmátrixot elemezzük.

A klasszikus legkisebb négyzetek módszere függ a manifeszt változó skálájától, így ez a becslési eljárás skálafüggő.

Könnyen belátható (lásd Krane és McDonald, 1978), hogy a maximum likelihood becslési eljárás invariáns a skála-transzformációval szemben.

13.3. A konfirmatív faktormodell paramétereinek becslése

Az identifikáció vizsgálata után – ha a modell identifikálhatónak bizonyult – kezdhetünk a modell paramétereinek becsléséhez. Keressük a paramétereknek azt a becslését, amelyekkel a lehető legjobban reprodukálni tudjuk a manifeszt változók mintabeli variancia-kovarianciamátrixát. A populáció variancia-kovarianciamátrixa (Σ) a faktormodell szerint a Λ, Φ és Θ paramétermátrixok függvénye: $\Sigma = \Lambda\,\Phi\,\Lambda' + \Theta$.

A populáció variancia-kovarianciamátrixának becsléséhez a populáció paramétereinek becslésével juthatunk: $\widehat{\Sigma} = \widehat{\Lambda}\,\widehat{\Phi}\,\widehat{\Lambda}' + \widehat{\Theta}$. A becsült paramétermátrixoknak ki kell elégíteniük az *a priori* feltételeket, és a becslési eljárásnak olyan Σ becslést kell generálnia, amelynek elemei a lehető legjobban illeszkednek a mintabeli variancia-kovarianciamátrix megfelelő elemeihez. Az illeszkedés jóságát különböző függvényekkel mérhetjük. Általánosan az illeszkedést mérő függvényt jelölje $F(\mathbf{S}, \Sigma)$ vagy $F(\mathbf{S}, \Sigma(\Lambda, \Phi, \Theta))$.

Azokat a Λ, Φ, Θ paramétermátrixokat tekintjük a sokaság paraméterei becslésének (ezeket $\widehat{\Lambda}$, $\widehat{\Phi}$, $\widehat{\Theta}$ jelöli), amelyek esetén az F függvény felveszi minimumát.

A konfirmatív faktorelemzésnél az illesztés során háromféle függvényt alkalmaznak. Ezek a függvények háromféle statisztikai becslési eljáráshoz kapcsolódnak:

1) a súlyozatlan vagy klasszikus legkisebb négyzetek módszere (ULS),

2) az általánosított legkisebb négyzetek módszere (GLS) és

3) a maximum likelihood (ML) módszere.

13.3.1A legkisebb négyzetek módszere

A legkisebb négyzetek módszere (ULS) a mintabeli variancia-kovarianciamátrixnak ($\mathbf{S}$) és a konfirmatív faktormodell variancia-kovarianciamátrixának (Σ) megfelelő elemei közötti eltérés négyzetösszegét minimalizálja:

$$F_{ULS}(\mathbf{S}, \Sigma) = tr[(\mathbf{S} - \Sigma)^2]. \tag{13.9}$$

Az ULS-becslés nem tételez fel a megfigyelt változók eloszlásáról semmit. Ez a legnagyobb előnye. Hátránya viszont, hogy így az illeszkedés jóságát nem tudjuk statisztikai próbával vizsgálni. Másik hátránya, hogy a becslési eljárás skálafüggő, vagyis függ a megfigyelt változók mértékegységétől.

13.3.2A z általánosított legkisebb négyzetek módszere

Az általánosított legkisebb négyzetek módszere (GLS) a mintabeli variancia-kovarianciamátrixnak ($\mathbf{S}$) és a konfirmatív faktormodell $\boldsymbol{\Sigma}$ mátrixának megfelelő elemei közötti különbséget súlyozza az $\mathbf{S}^{-1}$ mátrix elemeivel, és a súlyozott eltérés-négyzetösszeget számítja:

$$F_{GLS}(\mathbf{S}, \boldsymbol{\Sigma}) = tr[(\mathbf{S} - \boldsymbol{\Sigma})\mathbf{S}^{-1}]^2. \tag{13.10}$$

Bár az általánosított legkisebb négyzetek módszere nem tételezi fel, hogy a megfigyelt változók valamilyen meghatározott valószínűségeloszlást követnek, viszont nagy előnye, hogy ha $\mathbf{x}$ normális eloszlású, a GLS becslés aszimptotikus tulajdonságú (a mintabeli becslés várható értéke közelíti a populáció paraméterének értékét, valamint a standard hiba a lehető legkisebb), és aszimptotikusan közelít a maximum likelihood becsléshez.

13.3.3. A maximum likelihood becslés

A maximum likelihood (ML) becslésnél feltételezzük, hogy a megfigyelt változók együttes eloszlása normális:

$$f(\mathbf{x}) = (2\pi)^{-m/2}|\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\} = \tag{13.11}$$

$$= (2\pi)^{-m/2}|\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}tr[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}]\right\},$$

ahol $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\,\boldsymbol{\Phi}\,\boldsymbol{\Lambda}' + \boldsymbol{\Theta}$.

Általában feltételezzük, hogy a megfigyelt változókat várható értéküktől való eltérésükkel mérjük, így $\mu = 0$, és $f(\mathbf{x})$ egyszerűbb alakra hozható:

$$f(\mathbf{x}) = (2\pi)^{-m/2}|\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}tr(\mathbf{x}\,\mathbf{x}'\boldsymbol{\Sigma}^{-1})\right]. \tag{13.12}$$

Ha az x változók együttes eloszlása normális, a minta

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i',$$

variancia-kovarianciamátrixa $(n - 1)$ szabadságfokú Wishart-eloszlást követ, ahol n a minta elemszáma, $\mathbf{x}_i$ az i-edik megfigyelés m változóra vonatkozó mérési értékeit tartalmazza (az x változókról feltételezzük, hogy standardizáltak).

A likelihood-függvény logaritmusa:

$$\log_e L = -\frac{1}{2}(n-1)\{\log_e 2\pi + \log_e|\boldsymbol{\Sigma}| + tr(\mathbf{S}\boldsymbol{\Sigma}^{-1})\}. \tag{13.13}$$

A log-likelihood-függvény maximalizálása helyett praktikus megfontolások miatt a gyakorlatban a következő F_{ML} függvényt minimalizáljuk:

$$F_{ML} = tr(\mathbf{S}\boldsymbol{\Sigma}^{-1}) + [\log_e |\boldsymbol{\Sigma}| - \log_e |\mathbf{S}|] - m, \tag{13.14}$$

($F_{ML} = -(c_1 \log_e L + c_2)$, ahol c_1 és c_2 konstansok). F_{ML} minimalizálására Jöreskog (1967) javasolt jól működő eljárást. Az egyszerűség kedvéért tételezzük fel, hogy a közös faktorok korrelálatlanok ($\Phi = \mathbf{I}$), és így a $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \Theta$ modell paramétereit becsüljük. Jöreskog eljárása két lépcsős, először az F_{ML} függvényt $\boldsymbol{\Lambda}$ szerint minimalizáljuk rögzített Θ érték mellett, második lépésként a minimalizálást Θ szerint végezzük rögzített $\boldsymbol{\Lambda}$ mellett.

Az első fázisban feltételezzük, hogy létezik F_{ML} minimuma rögzített Θ mellett, ezt $f(\Theta)$-val jelöljük, és Λ_Θ-val azt a Λ-t, amelynél a minimum található.

$$f(\boldsymbol{\Theta}) = \min_{\boldsymbol{\Lambda}} F_{ML}(\mathbf{S}, \boldsymbol{\Sigma}(\boldsymbol{\Lambda}, \boldsymbol{\Theta})). \tag{13.15}$$

Második fázisban a Θ szerint minimalizálunk:

$$\min_{\Theta} f(\boldsymbol{\Theta}) = \min_{\boldsymbol{\Lambda}, \boldsymbol{\Psi}} F_{ML}. \tag{13.16}$$

Az adott Θ mellett F_{ML} minimalizálása visszavezethető (a parciális deriváltakat egyenlővé téve $\mathbf{0}$-val) egy sajátérték-sajátvektor feladatra. Eszerint $\boldsymbol{\Lambda}_\Theta$ kifejezhető $\Theta^{-\frac{1}{2}}\mathbf{S}\,\Theta^{-\frac{1}{2}}$ sajátértékeivel és sajátvektoraival:

$$\boldsymbol{\Lambda}_\Theta = \boldsymbol{\Theta}^{\frac{1}{2}}\boldsymbol{\Gamma}(\boldsymbol{\Delta} - \mathbf{I})^{\frac{1}{2}},$$

ahol Δ diagonális mátrix, amelynek diagonálisa a $\Theta^{-\frac{1}{2}}\mathbf{S}\,\Theta^{-\frac{1}{2}}$ mátrix első r sajátértékét tartalmazza csökkenő sorrendben,

A $\boldsymbol{\Gamma}\Theta^{-\frac{1}{2}}\mathbf{S}\Theta^{-\frac{1}{2}}$ sajátvektorait tartalmazza a sajátértékeknek megfelelő sorrendben.

Könnyen belátható, hogy $\boldsymbol{\Lambda}_\Theta$ kielégíti a következő két feltételt:

$$\begin{aligned}\boldsymbol{\Lambda}_\Theta\boldsymbol{\Lambda}_\Theta' &= \boldsymbol{\Theta}^{\frac{1}{2}}\boldsymbol{\Gamma}(\boldsymbol{\Delta} - \mathbf{I})^{\frac{1}{2}}(\boldsymbol{\Delta} - \mathbf{I})^{\frac{1}{2}}\boldsymbol{\Gamma}'\,\boldsymbol{\Theta}^{\frac{1}{2}} \\ &= \boldsymbol{\Theta}^{\frac{1}{2}}[\boldsymbol{\Theta}^{-\frac{1}{2}}\mathbf{S}\,\boldsymbol{\Theta}^{-\frac{1}{2}} - \mathbf{I}]\boldsymbol{\Theta}^{\frac{1}{2}} \\ &= \mathbf{S} - \boldsymbol{\Theta},\end{aligned}$$

ami a modell alapegyenletének egy becslése, és

$$\boldsymbol{\Lambda}_\Theta'\boldsymbol{\Theta}^{-1}\boldsymbol{\Lambda}_\Theta = (\boldsymbol{\Delta} - \mathbf{I})^{\frac{1}{2}}\boldsymbol{\Gamma}'\boldsymbol{\Gamma}(\boldsymbol{\Delta} - \mathbf{I})^{\frac{1}{2}} = (\boldsymbol{\Delta} - \mathbf{I}),$$

ami diagonális mátrix, és ezzel $\boldsymbol{\Lambda}_\Theta$ a modell identifikálhatóságára bevezetett pótlólagos feltételt is kielégíti.

$\boldsymbol{\Lambda}_\Theta$ meghatározása után az $f(\Theta)$ függvényt minimalizáljuk a Fletcher és Powell (1963) eljárás szerint. Az iteratív eljárás során az egymást követő $\Theta^{(1)}, \Theta^{(2)}, \ldots$ mátrixokra teljesül, hogy

$$f(\boldsymbol{\Theta}^{(h+1)}) < f(\boldsymbol{\Theta}^{(h)}),$$

és minden lépésben újraszámítjuk a $\Theta^{(h)}$-hoz tartozó $\boldsymbol{\Lambda}_\Theta^{(h)}$ mátrixot. Az eljárás gyorsan konvergál a végső becslés $\widehat{\boldsymbol{\Lambda}}$ és $\widehat{\boldsymbol{\Theta}}$ mátrixaihoz. Az eljárás első lépésénél szükségünk van a Θ mátrix kezdeti becslésére, amit Jöreskog a következőképpen határozott meg:

$$\psi_{ii}^{(0)} = \left(1 - \frac{1}{2}r/m\right)(1/s^{ii}), \tag{13.17}$$

ahol s^{ii} az $\mathbf{S}^{-1}$ mátrix i-edik diagonális eleme.

Az eljárásnak két lehetséges problémája van. Az egyik, hogy a lokális minimumot találjuk meg a globális minimum helyett. Ez a gyakorlatban tapasztalatunk szerint ritkán fordul elő (lásd pl. Jöreskog és Sörbom, 1981). A másik lehetséges probléma, hogy az f függvény minimumhelyén a reziduális varianciák között előfordul negatív érték. Ezt az esetet az irodalomban *Heywood-esetnek* nevezik (Heywood, 1931).

A Heywood-eset elsősorban a minta hibája, ezért elsősorban a minta elemszámával függ össze. Kis minta esetén ($n \leq 100$) nagy, nagy minta esetén ($n \geq 500$) kicsi az előfordulásának esélye.

Ha a modell egyébként helyesen definiált, a Heywood-eset előfordulásának valószínűsége közelít 0-hoz $n \longrightarrow \infty$ esetén.

Adott mintanagyság esetén a negatív variancia előfordulásának esélye csökken, ha a változók száma nő. Nő a Heywood-eset bekövetkezésének valószínűsége, ha a változópárok korrelációi között előfordulnak nagyon magas értékek. Az ilyen változók közül az egyiket minden különösebb veszteség nélkül elhagyhatjuk. Ez azonban nem mindig szünteti meg a problémát.

A Heywood-eset előfordulása összefügghet természetesen a modellel magával is. Legkézenfekvőbb, amikor több faktort szerepeltetünk a modellben, mint amennyi elméletileg lehetséges.

Összefügghet a negatív variancia előfordulása a változók normális eloszlására vonatkozó feltétel megsértésével is.

A Heywood-eset előfordulása lehet a helytelen hiányzó adat-kezelés következménye is. A gyakorlati alkalmazásokban a kovariancia- vagy korrelációmátrixot legtöbbször a páronkénti hiányzó adat-kezeléssel számítjuk. Ez azt jelenti, hogy minden megfigyelést figyelembe veszünk, ha az éppen vizsgált két változóra érvényes mérési eredménnyel rendelkeznek. Ebből következően minden kovarianciát vagy korrelációt más-más mintából becsülünk. Különösen akkor alkalmazzuk a páronkénti hiányzó adat-kezelést, ha a minta nagysága kicsi, ami még jobban növeli a Heywood-eset előfordulásának valószínűségét.

A páronkénti hiányzó adat-kezelést csak akkor alkalmazhatjuk, ha a hiányzó adatok száma mind a változóknál, mind a megfigyeléseknél kicsi.

Természetesen, ha pontosan tudjuk a Heywood-eset előfordulásának okát, azt könnyen megszüntethetjük. Egyébként a problémát megkerülve vagy azt tehetjük, hogy az iterációs eljárást leállítjuk még azelőtt, hogy a hibavariancia negatívvá válna (valamilyen rögzített kicsi Θ_i érték pl. 0,05 vagy 0,01 mellett), vagy ahhoz a változóhoz, amelynél a hibakomponens negatív varianciája előfordult, hozzáadunk egy ismert (σ^2) varianciájú véletlen változót, és az új változót alkalmazzuk a modellben:

$$\mathrm{var}(x_i^*) = \sum_j^r \lambda_{ij} + \Theta_i + \sigma^2,$$

ahol σ^2-et úgy kell meghatározni, hogy az x_i^* változók mellett a Heywood-eset ne forduljon elő.

Az x változó helyett az x^* változót alkalmazva az új korrelációmátrixot úgy kaphatjuk meg, hogy a diagonálison kívüli elemeket megszorozzuk $(1+\sigma^2)^{-\frac{1}{2}}$-nel. Ebben az esetben az eredeti és a módosított modell paraméterei között a kapcsolat:

$$\lambda_{ij} = \lambda_{ij}^*(1+\sigma^2)^{\frac{1}{2}} \quad \text{és} \quad \Theta_i = 1 - (1+\sigma^2)\sum_j^r \lambda_{ij}^{*2},$$

ahol λ_{ij}^*-ok jelölik a módosított modell faktorsúlyait.

13.4. A modell illeszkedésének vizsgálata

Az általánosított legkisebb négyzetek és a maximum likelihood módszer alkalmazása esetén – ha a becslési módszerek feltételei teljesülnek – khi-négyzet próbával tesztelhetjük a modell illeszkedését a mintabeli variancia-kovarianciamátrixhoz.

Legyen H_0 az adott modell nullhipotézise (a modell tökéletesen illeszkedik a populáció kovarianciamátrixához). H_0 alternatív hipotézise (H_1) azt fejezi ki, hogy $\boldsymbol{\Sigma}$ bármely pozitív definit, szimmetrikus mátrix lehet, amelynek a rangja m.

A khi-négyzet próba méri a modell becslésével kapott $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Lambda}}\,\widehat{\boldsymbol{\Phi}}\,\widehat{\boldsymbol{\Lambda}} + \widehat{\boldsymbol{\Theta}}$ és a mintabeli variancia-kovarianciamátrix ($\mathbf{S}$) közötti eltérést. A khi-négyzet statisztikához tartozó szabadságfok a konfirmatív faktormodellben: szfok=(a független paraméterek száma a H_1 hipotézis mellett)

– (a független paraméterek száma a H_0 hipotézis mellett)

$$\text{szfok} = \frac{1}{2}m(m+1) - t, \tag{13.18}$$

ahol m a megfigyelt változók száma, t a modell ismeretlen paramétereinek száma.

A modell becslésénél láttuk, hogy a paraméterek identifikálhatósága miatt fel kell tételeznünk, hogy a $\boldsymbol{\Lambda}'\boldsymbol{\Theta}^{-1}\boldsymbol{\Lambda}$ mátrix diagonális mátrix, így a modell $\frac{1}{2}r(r-1)$ pótlólagos feltételt tartalmaz a paraméterekre. Ha feltételezzük, hogy a közös faktorok korrelálatlanok ($\boldsymbol{\Phi} = \mathbf{I}$), akkor a $\boldsymbol{\Lambda}$, $\boldsymbol{\Theta}$ paramétermátrixok

$$m + mr - \frac{1}{2}r(r-1)$$

szabad, ismeretlen paramétert tartalmaznak, így a szabadságfok:

$$\begin{aligned}\text{szfok} &= \frac{1}{2}m(m+1) - [m + mr - \frac{1}{2}r(r-1)] \\ &= \frac{1}{2}[(m-r)^2 - (m+r)].\end{aligned} \tag{13.19}$$

Természetesen a konfirmatív faktormodellben általában nem tételezzük fel a közös faktorok függetlenségét, és a $\boldsymbol{\Lambda}$, $\boldsymbol{\Phi}$ és $\boldsymbol{\Theta}$ paramétermátrixok elemeire lehetnek pótlólagos feltételezéseink, pl. ha $\boldsymbol{\Phi}$ szimmetrikus mátrix, akkor a ϕ_{21} paramétert nem becsülhetjük, mivel $\phi_{12} = \phi_{21}$ a szimmetria miatt. Vagy feltételezhetjük, hogy bizonyos faktorsúlyok egyenlők egymással, pl. $\lambda_{11} = \lambda_{12} = \lambda_{13}$. Így egy adott modell független paramétereinek számát a konkrét feltételezések figyelembevételével határozhatjuk meg.

A becsült ($\widehat{\boldsymbol{\Sigma}}$) és a mintabeli variancia-kovarianciamátrix ($\mathbf{S}$) eltérésének statisztikai tesztelésére a hagyományos khi-négyzet statisztika (χ^2) mellett a modell becslésénél definiált likelihood függvény is alkalmas.

Jelölje Ω az összes lehetséges $\boldsymbol{\Sigma}$ mátrixok halmazát (a pozitív definit, szimmetrikus, m rangú mátrixok halmazát). Legyen ω ennek azon részhalmaza, amelyhez tartozó kovarianciamátrixok megfelelnek a H_0 hipotézisnek (a (13.2) egyenlettel definiált modellnek és a modell feltételeinek). Jelölje L_Ω és L_ω a megfelelő maximum likelihood függvény maximumát. Mivel az L függvény maximumát az Ω halmazon akkor veszi fel, ha $\boldsymbol{\Sigma} = \mathbf{S}$, így a (13.13) egyenlet a konstanstól eltekintve a következő lesz:

$$\log_e L_\Omega = -\frac{1}{2}(n-1)(\log_e |\mathbf{S}| + m). \tag{13.20}$$

L_ω-t megkaphatjuk, ha Σ-t helyettesítjük $\widehat{\Sigma} = \widehat{\Lambda}\widehat{\Lambda}' + \Theta$ becslésével (feltételezve az egyszerűség kedvéért megint, hogy $\Phi = \mathbf{I}$):

$$\log_e L_\omega = -\frac{1}{2}(n-1)[\log_e |\widehat{\Sigma}| + tr(\mathbf{S}\,\widehat{\Sigma}^{-1})]. \tag{13.21}$$

A likelihood-hányados:

$$\lambda = L_\omega / L_\Omega. \tag{13.22}$$

Bár λ pontos valószínűségeloszlását nem ismerjük, de nagy minták esetén ismert, hogy $-2\log_e \lambda$ eloszlása közelítőleg khi-négyzet eloszlású, ha H_0 igaz.

A (13.20) és (13.21) egyenletek alapján

$$-2\log_e \lambda = -2\log_e L_\omega + 2\log_e L_\Omega \tag{13.23}$$

$$= (n-1)[\log_e |\widehat{\Sigma}| + tr(\mathbf{S}\,\widehat{\Sigma}^{-1}) - \log_e |\mathbf{S}| - m].$$

A (13.23) egyenletet összehasonlítva a (13.14) egyenlettel világosan látszik, hogy a likelihood-hányados logaritmusának mínusz kétszerese $(-2\log\lambda)$ egyenlő F_{ML} minimumának $(n-1)$-szeresével. Vagyis a maximum likelihood becslés F_{ML} függvénye felhasználható a modell illeszkedésének statisztikai tesztelésére:

$$v = (n-1)\min F_{ML} \tag{13.24}$$

aszimptotikusan khi-négyzet eloszlású, a szabadságfok egyenlő a mintabeli variancia-kovarianciamátrix független elemei számának és a becsült független paraméterek számának különbségével, általánosan

$$\text{szfok} = m(m+1)/2 - t,$$

ahol t a modell becsült független paramétereinek száma.

A v statisztika a minta elemszámának, valamint $\widehat{\Sigma}$ és $\mathbf{S}$ illeszkedésének a függvénye.

Nagyon sok esetben fontos tudni, hogy a modell változtatása javított-e a modell illeszkedésén. Ha a vizsgált modellek hierarchikusak, vagyis ha a H_m modellt a H_b modellből kapjuk meg úgy, hogy a H_b modell egy vagy több paraméterére feltételeket teszünk, tehát ha a H_m modell lehetséges Σ_m variancia-kovarianciamátrixai részhalmazát képezik a H_b modell lehetséges Σ_b mátrixai halmazának (lásd pl. a ω és a Ω halmazokat), akkor a modelleknek megfelelő likelihood-hányados statisztikák különbsége is khi-négyzet eloszlást követ:

$$(n-1)(F_m - F_b) \tag{13.25}$$

khi-négyzet eloszlású (szfok_m - szfok_b) szabadságfokkal.

Mivel a likelihood-hányados statisztika (és a khi-négyzet statisztika is) függ a minta nagyságától, kis minta esetén még a nyilvánvalóan nem megfelelő modell is elfogadható illeszkedést mutat. Nagy minta esetén figyelembe kell venni majdnem minden lehetséges közös faktort, hogy elfogadhatóan illeszkedő modellt kapjunk, ami nem egyezik az elméleti törekvéssel, a khi-négyzet-statisztikát próbálták transzformálni olymódon, hogy kevésbé legyen érzékeny a minta elemszámára.

Bentler és Bonett (1980) javasolta, hogy a khi-statisztikák felhasználásával 0 és 1 értékek közé eső mutatót számoljunk a modell kiértékelésére.

Az illeszkedés Bentler–Bonett-féle normált mutatója:

$$\triangle_1 = \frac{\chi_b^2 - \chi_m^2}{\chi_b^2}, \tag{13.26}$$

vagy

$$\triangle_1 = \frac{F_b - F_m}{F_b}, \tag{13.27}$$

ahol χ_b^2 egy alapmodell (nullhipotézis, amely szerint a változók között nincs összefüggés) illeszkedését mérő χ^2 statisztika, χ_m^2 a hipotetikus, javított modell χ^2 értéke, F_b és F_m az illesztett függvény értékei a két modellnél.

$\triangle_1$ méri a hipotetikus, vizsgált modelllilleszkedésének javulását az alapmodellhez viszonyítva.

$\triangle_1$ értéke 0 és 1 között mozog, közeledve 1-hez jobb modelllilleszkedést mutat. Maximális értékét akkor veszi fel, ha χ_m^2 (vagy F_m) egyenlő nullával. χ_m^2 nagy minta esetén aszimptotikusan khi-négyzet eloszlást követ, és így χ_m^2 várható értéke közelítőleg a szabadságfokával (szf_m) lesz egyenlő.

A χ_b^2 becslő függvény csak bizonyos feltételek mellett (lásd Steiger, 1985) követ nemcentrális khi-négyzet eloszlást.

A ($\overline{\chi_b^2}$- szf_m) kifejezés (ahol $\overline{\chi}_b^2$ az alapmodell khi-négyzet értékének az átlaga) átlagértéket ad a vizsgált modell esetére, így a ($\chi_b^2 - \chi_m^2$) különbségét viszonyíthatjuk ehhez a várható különbséghez. A $\triangle_2$ mutató (a $\overline{\chi}_b^2$ átlag helyett annak becslését, χ_b^2-t helyettesítettük):

$$\triangle_2 = \frac{\chi_b^2 - \chi_m^2}{\chi_b^2 - \text{szf}_m}. \tag{13.28}$$

A (13.28) mutató magasabb értéket ad, ha kevesebb paraméterrel kapjuk ugyanazt a becsült χ_m^2 értéket.

Az illesztett függvényértékekre:

$$\triangle_2 = \frac{F_b - F_m}{F_b - (\text{szf}_m/(n-1))}, \tag{13.29}$$

ahol n a minta elemszáma.

Ha $\triangle_2$ számlálója és nevezője is pozitív, akkor $\triangle_2$ nagyobb mint $\triangle_1$. $\triangle_2$ nem normalizált mutató, ezért lehetséges, hogy a 0 és 1 határokon kívüli értéket vesz fel. $\triangle_2$ maximuma $\chi_b^2/(\chi_b^2 - \text{szf}_b)$, feltéve hogy $\chi_b^2 > \text{szf}_b$. 1 $\triangle_2$ minimuma 0 hierarchikus modellek esetén.

A $\triangle_2$ mutató általánosított formáját kapjuk, mikor a számlálóba az alapmodell helyett egy kevésbé korlátozott modell (de nem annyira, mint a vizsgált) khi-négyzet értékét tesszük.

Ekkor

$$\triangle_2 = \frac{\chi_r^2 - \chi_m^2}{\chi_b^2 - \text{szf}_m}.$$

Boolen (1986) közölt a $\triangle_1$-hez hasonló, a szabadságfokokhoz viszonyított eltérést mérő mutatót:

$$\rho_1 = \frac{\left(\frac{\chi_b^2}{\text{szf}_b}\right) - \left(\frac{\chi_m^2}{\text{szf}_m}\right)}{\left(\frac{\chi_b^2}{\text{szf}_b}\right)}, \tag{13.30}$$

vagy

$$\rho_1 = \frac{\left(\frac{F_b}{\text{szf}_b}\right) - \left(\frac{F_m}{\text{szf}_m}\right)}{\left(\frac{F_b}{\text{szf}_b}\right)}. \tag{13.31}$$

ρ_1 normált mutató, értéke 0 és 1 között mozog. Bentler és Bonett (1980) javasolta ρ_1 nem normalizált változatát:

$$\rho_2 = \frac{\left(\frac{\chi_b^2}{\text{szf}_b}\right) - \left(\frac{\chi_m^2}{\text{szf}_m}\right)}{\left(\frac{\chi_b^2}{\text{szf}_b}\right) - 1}, \tag{13.32}$$

$$\rho_2 = \frac{\left(\frac{F_b}{\text{szf}_b}\right) - \left(\frac{F_m}{\text{szf}_m}\right)}{\left(\frac{F_b}{\text{szf}_b}\right) - (1/(n-1))}. \tag{13.33}$$

A minta nagysága kétféleképpen hat az illeszkedés jóságát mérő mutatókra.

Az egyik hatás, amikor a minta elemszáma (n) közvetlenül szerepel a számításnál.

Belátható (lásd pl. Boolen, 1986), hogy ρ_2 és $\triangle_2$ közvetlenül függ a mintanagyságtól, míg ρ_1 és $\triangle_1$ nem.

A másik hatás, amikor az illeszkedést mérő mutató mintabeli eloszlásának várható értéke függ a minta elemszámától. Ha egy illeszkedő modell mellett különböző elemszámú véletlen mintákat veszünk, és kiszámoljuk az illeszkedést mérő mutatókat és azok eloszlásait, akkor azt tapasztaljuk, hogy $\triangle_1$ és ρ_1 mintabeli eloszlásának várható értéke pozitív kapcsolatban van a minta elemszámával (n-nel), míg $\triangle_2$ és ρ_2 függősége közel nulla.

Boolen (1989) közölt Monte Carlo-kísérletet a $\triangle_1$, $\triangle_2$, ρ_1, és ρ_2 mutatók eloszlásának vizsgálatára.

Vizsgálta az

$$\mathbf{x} = \mathbf{\Lambda f} + \mathbf{e}$$

modellt, ahol a megfigyelt változók (x) száma $m = 5$, a faktorok (f) száma $r = 2$ volt. Boolen a konfirmatív faktormodellt alkalmazta, ahol a modell paramétereire a következő feltételezéseket tette:

$$\mathbf{\Lambda}_x = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{\Phi} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix},$$

$$\text{diag}\mathbf{\Theta} = [1, 1, 1, 1, 1, 1],$$

ahol Φ a közös faktorok variancia-kovarianciamátrixa, Θ a hibakomponensek variancia-kovarianciamátrixa.

A megfigyelt változók eloszlása rendre normális eloszlású volt. Boolen három mintanagyságot vizsgált (75, 150, 300), és minden mintanagyságra tizennégyszer végezte el a számításokat a maximum likelihood becslést alkalmazva (F_{ML}). A következő táblázat mutatja a modell illeszkedését mérő mutatók átlagait és szórásait:

Az illeszkedés jóságát mérő mutatók	Várható érték (szórás) Minta		
	75	150	300
$\triangle_2$	1,003	1,002	1,000
	(0,019)	(0,012)	(0,008)
$\triangle_1$	0,967	0,985	0,992
	(0,018)	(0,012)	(0,008)
ρ_2	1,008	1,005	1,001
	(0,052)	(0,030)	(0,021)
ρ_1	0,918	0,963	0,980
	(0,046)	(0,030)	(0,020)

A táblázatból látható, hogy $\triangle_1$ és ρ_1 várható értéke függ a minta elemszámától, míg $\triangle_2$ és ρ_2 várható értéke közel esik 1-hez mindegyik mintanagyság esetén.

A másik jellemző tanulság a fenti vizsgálatból, hogy a $\triangle$ mutatók szórása lényegesen kisebb (kevesebb mint fele), mint a ρ mutatóké.

Ezenkívül látható még, hogy a $(\triangle_2 - \triangle_1)$ és a $(\rho_2 - \rho_1)$ különbségek csökkennek a minta elemszámának növekedésével.

13.5. A faktorok értelmezése és transzformálása

Miután a faktormodellt illesztettük az adatokhoz, a paraméterek becslése ismeretében a paraméterek és a faktorok értelmezése következhet. A továbbiakban is feltételezzük, hogy a közös faktorok korrelálatlanok.

Az értelmezést a következő egyenlettel kezdjük:

$$var(x_i) = \lambda_{i1}^2 + \lambda_{i2}^2 + \ldots + \lambda_{ir}^2 + \Theta_i \tag{13.34}$$

$$var(x_i) = \sum_{k=1}^{r} \lambda_{ik}^2 + \Theta_i \quad i = 1 \ldots, m.$$

A (13.34) egyenlet az i-edik megfigyelt változó varianciáját két tagra bontja. Az első tag az i-edik változó varianciájának az r-számú közös faktorral magyarázható része, a második tag a hibakomponens (vagy egyedi faktor) által magyarázott rész. Az első tag az adott változó kommunalitása. Általában a változókat standardizáljuk, így

$$1 = var(x_i) = \sum_{k=1}^{r} \lambda_{ik}^2 + \Theta_i .$$

Ha az együttes varianciát vizsgáljuk, összegezve a (13.34) egyenleteket $i = 1, \ldots, m$-re, a következőket kapjuk:

$$\sum_{i}^{m} var(x_i) = \sum_{i}^{m} \lambda_{i1}^2 + \sum_{i}^{m} \lambda_{i2}^2 + \ldots + \sum_{i}^{m} \lambda_{ir}^2 + \sum_{i}^{m} \Theta_i , \tag{13.35}$$

ahol a $\sum_i^m \lambda_{ik}^2$ a k-adik faktor hozzájárulását fejezi ki a teljes varianciához. Ezt kifejezhetjük relatív formában is:

$$\frac{\sum_i^m \lambda_{ik}^2}{\sum_i^m var(x_i)}.$$

Szokás ezt a kifejezést százalékos formában is használni. Ilyenkor azt mondjuk, hogy az adott faktor az összvariancia

$$\frac{\sum_i^m \lambda_{ik}^2}{\sum_i^m var(x_i)} \cdot 100\%\text{-át magyarázza.}$$

A faktorok értelmezését a λ_{ik} faktorsúlyok segítségével adhatjuk meg. A λ_{ik} faktorsúlyt mint standardizált regressziós együtthatót értelmezhetjük. A λ_{ik} kifejezhető az i-edik változó és a k-adik faktor közötti korrelációs együtthatóval, ha a közös faktorok korrelálatlanok. A k-adik faktor értelmezéséhez a λ_{ik} faktorsúlyok relatív nagyságát és előjelét kell figyelembe vennünk. Minél nagyobb a λ_{ik} érték, annál szorosabban kapcsolódik az i-edik változó a k-adik faktorhoz. Az azonos előjelű, nagy relatív súlyú változókat összegyűjtve az ezen változókban lévő közös tulajdonság, közös jellemző fejeződik ki a faktorban. Ha minden változó azonos mértékben és előjellel kapcsolódik a faktorhoz, akkor az adott faktort általános faktornak nevezhetjük. Gyakran előfordul, hogy a változók egy része pozitívan, egy másik része negatívan kapcsolódik az adott faktorhoz. Ilyenkor az adott faktor egyik irányban azt fejezi ki, amiben az egyik változóhalmaz közös ugyanakkor, amikor a másik változóhalmaz közös jellemzője hiányzik, míg másik irányban fordítva mér. Ezt a faktort bipoláris faktornak nevezzük.

Azokat a faktorokat lehet könnyen értelmezni, amelyekben λ egy része közel esik 0-hoz, a másik része hasonló nagyságú és előjelű. Ha a faktorsúlyokat nem lehet eszerint szeparálni, akkor a faktorteret transzformálhatjuk úgy, hogy a faktorok értelmezése lehetséges (vagy könnyebb) legyen.

A paraméterek becslése – mint láttuk – kielégíti azt a feltételt, hogy a $\mathbf{\Lambda}'\mathbf{\Theta}^{-1}\mathbf{A}$ mátrix diagonális legyen. Ez azt jelenti, hogy a faktorokat a főtengelyek mentén határozzuk meg. Az első faktor a pontok legnagyobb szóródása irányába mutat, a második a lehetséges második legnagyobb szóródás irányába, de merőlegesen az elsőre, és így tovább. A gyakorlatban nem biztos, hogy ez a legjobb választás, és így a faktorok értelmezése meglehetősen nehéz lehet. A faktorteret transzformálhatjuk úgy, hogy a modell illeszkedése ne változzon meg, de a faktorsúlyok mátrixa lehetőleg egyszerű struktúrát mutasson. Egy faktormátrixot egyszerű struktúrának nevezünk, ha a változók a lehető legkevesebb faktorral kapcsolódnak a 0-tól különböző súllyal.

Tekintsük az irodalomban sokszor szereplő példát a faktortér transzformációjára (lásd pl. Everitt, 1984).

A következő táblázat a hat tantárgy iskolai eredményei közötti korrelációt mutatja egy 220 iskolás fiú adatait tartalmazó mintában.

	Iskolai tantárgyak					
	1	2	3	4	5	6
1 Francia nyelv	1,00					
2 Angol nyelv	0,44	1,00				
3 Történelem	0,41	0,35	1,00			
4 Számtan	0,29	0,35	0,16	1,00		
5 Algebra	0,33	0,32	0,19	0,59	1,00	
6 Geometria	0,25	0,33	0,18	0,47	0,46	1,00

13.1. táblázat. Iskolai tantárgyak korrelációmátrixa

A faktormodell illesztésekor két közös faktort találtak. Az eredményeket a következő táblázat tartalmazza.

Tantárgyak	Faktorsúlyok		Kommunalitások
1 Francia nyelv	0,55	0,43	0,49
2 Angol nyelv	0,57	0,29	0,41
3 Történelem	0,39	0,45	0,36
4 Számtan	0,74	−0,27	0,62
5 Algebra	0,72	−0,21	0,57
6 Geometria	0,59	−0,13	0,37

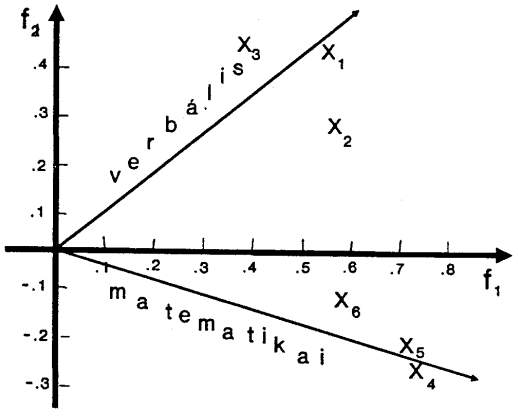
13.2. táblázat. Iskolai tantárgyak 2-faktoros modelljének eredményei

A fenti faktormátrixból látható, hogy az első faktor, amelyik általános faktornak tekinthető, mindegyik változóval pozitívan, viszonylag erősen korrelál. Ezt általános intelligencia-faktornak nevezhetjük el.

A második faktor a változókat két részre bontja, egyik részével pozitívan, másikkal negatívan korrelál, ezért ez a faktor bipoláris, és a verbális kontra matematikai képességet fejezi ki.

A második faktor azt mutatja, hogy egy adott általános intelligenciaszinten azok a tanulók, akik jó eredményeket érnek el a verbális tárgyakból, kevésbé jók a matematikai tárgyakból, és fordítva.

A faktorsúlyok mátrixát ábrázolhatjuk úgy, hogy a tengelyek a faktorokat reprezentálják, a faktorsúlyok a változóknak a faktorokra vonatkozó koordinátái:



A faktorok grafikus rotálásánál a tengelyeket úgy forgatjuk, hogy lehetőleg elkerüljük a negatív faktorsúlyokat, és a lehető legkevesebb nullától különböző súly maradjon.

A faktortengelyeket rotálva a két változócsoporthoz, eredményül két korrelált faktort kapunk, az egyiket a verbális képességek, a másodikat a matematikai képességek faktorának nevezhetjük.

Láttuk korábban, hogy a faktorok ortogonális transzformációja a modell illeszkedését nem változtatja meg. Általánosságban a faktorok grafikus rotálását az

$$\mathbf{M} = \begin{pmatrix} \cos\Theta & \sin\Theta \\ -\sin\Theta & \cos\Theta \end{pmatrix}$$

ortogonális mátrix segítségével végezhetjük Θ forgásszöggel az óra járásával azonos irányban. Az új faktorsúlyok mátrixa:

$$\mathbf{\Lambda}^* = \mathbf{\Lambda}\,\mathbf{M}.$$

A forgásszöget megkaphatjuk úgy, hogy az ábráról lemérjük. Az óra járásával ellentétes irányú forgatásmódot ad az alábbi ortogonális transzformációsmátrix:

$$\mathbf{M} = \begin{pmatrix} \cos\Theta & -\sin\Theta \\ \sin\Theta & \cos\Theta \end{pmatrix}.$$

A rotálás numerikus módszerei közül a leggyakrabban használt a *varimax*-módszer. A varimax-módszer célja olyan súlyokat találni, hogy abszolút értékben lehetőleg csak viszonylag nagy vagy viszonylag kicsi faktorsúlyok forduljanak elő. A varimax-eljárás minden változóra maximalizálja a faktorsúlyok négyzeteinek varianciáját:

$$S_v = \sum_k^r \left[\frac{m \sum_i^m (\lambda_{ik}^2)^2 - \left(\sum_i^m \lambda_{ik}^2 \right)^2}{m^2} \right] \tag{13.36}$$

A varimax-módszer részletes leírását lásd pl. Lawley és Maxwell (1971) 6. fejezetében. A *varimax* minimalizálja azon változók számát, amelyeknek nagy súlya van az adott faktorban.

A varimax-rotálás ortogonális faktorokat eredményez. Hasonlóan ortogonális transzformációs eljárás a *quartimax*. A quartimax-rotálásnál a faktorsúlyok negyedik hatványának

$$S_q = \sum_i^m \sum_k^r \lambda_{ik}^4 \tag{13.37}$$

összegét maximalizáljuk. A *Quartimax* minimalizálja a faktorok számát, amelyek szükségesek a változók magyarázatához.

A nem ortogonális transzformációk közül, amelyek korrelált, ferdeszögű faktorokat eredményeznek, a leggyakrabban az oblimin-rotálást alkalmazzuk. Az oblimin-rotálásnál a különböző faktorok faktorsúlyai négyzeteinek kovarianciáit minimalizáljuk:

$$S_{d0} = \sum_\ell^r \sum_{k \neq \ell}^r \left[\sum_i^m \lambda_{i\ell}^2 \lambda_{ik}^2 - \frac{1}{m} \sum_i^m \lambda_{i\ell}^2 \sum_i^2 \lambda_{ik}^2 \right] \tag{13.38}$$

Ha a faktorok korrelálnak egymással, akkor a faktorsúlyokat már nem értelmezhetjük úgy, mint a változók és a faktorok közötti korrelációs együtthatókat.

13.6. Ipszatív változók faktorelemzése

Az ipszatív transzformáció fogalmát Catell (1944) vezette be (a latin ipse= ő maga szóból) egy változóhalmaz értékeinek a megfigyelési egységek átlagaihoz történő centrírozására. Az ipszatív transzformációt Horst (1965) jobb oldali centrírozásnak nevezte, megkülönböztetve a változók normatív transzformálásától (amikor a változókat saját átlagukhoz centrírozzuk), amit bal oldali centrírozásnak nevezett. Amikor a változók értékeit mind a két módon transzformáljuk, akkor dupla centrírozásról beszélünk.

Az ipszatív transzformálás eredményeként a változók értékeinek összege minden megfigyelési egységnél egyenlő egy konstanssal. A konstans a gyakorlatban egyenlő 0-val, mivel a megfigyelési egységek ipszatív értékei a vizsgált változóhalmaz értékeinek eltérései az adott megfigyelési egység átlagától. Általában az ipszatív fogalmat használjuk tekintet nélkül a konstans értékére, ha a változók értékeinek az összege minden megfigyelésnél egyenlő a konstans értékkel.

A gyakorlatban egy változóhalmaznak lehet ipszatív tulajdonsága akkor is, ha az nem ipszatív transzformáció eredménye, például rangsorolt változók esetében.

Legyen $\mathbf{y}$ m megfigyelt valószínűségi változó vektora. Tételezzük fel, hogy mindegyik változót azonos mértékegységgel, azonos skálán mérjük, és hogy $\mathbf{y}$ variancia-kovarianciamátrixa pozitív definit.

Az $\mathbf{y}$ változó ipszatív transzformációját általánosan a következőképpen írhatjuk:

$$\mathbf{x} = \mathbf{y} - \mathbf{1}\, w, \tag{13.39}$$

ahol $\mathbf{1}$ összegzővektor, w skalár, $\mathbf{x}$ $(m \times 1)$-típusú vektor, ipszatív tulajdonsággal.

Általában $w = m^{-1}(\mathbf{1}'\mathbf{y} - c)$, ahol c konstans, a változók értékeinek összege minden megfigyelésnél. c értéke általában indifferens, így célszerű 0-nak választani. Ekkor $w = m^{-1}\mathbf{1}'\mathbf{y}$. A (13.39) transzformációban a változók átlagát választjuk skalárnak, és így

$$\begin{aligned} \mathbf{x} &= \mathbf{y} - \mathbf{1}(m^{-1}\mathbf{1}'\mathbf{y}) \\ &= \mathbf{y} - m^{-1}\mathbf{1}\mathbf{1}'\mathbf{y} \\ &= (\mathbf{I} - m^{-1}\mathbf{U})\mathbf{y}, \end{aligned} \tag{13.40}$$

ahol $\mathbf{U} = \mathbf{1}\mathbf{1}'$ $(m \times m)$ típusú összegzőmátrix.

Az ipszatív transzformáció $(\mathbf{I} - \frac{1}{m}\mathbf{U})$ mátrixa szimmetrikus, idempotens mátrix, a rangja $(m - 1)$. (Lásd pl. Horst, 1965, 288–289).

Jelöljük az ipszatív transzformációmátrixát $\mathbf{A}$-val:

$$\mathbf{A} = \left(\mathbf{I} - \frac{1}{m}\mathbf{U}\right), \qquad \mathbf{x} = \mathbf{A}\,\mathbf{y}.$$

Az $\mathbf{x}$ variancia-kovarianciamátrixa:

$$\mathbf{\Sigma}_x = \mathbf{A}\,\mathbf{\Sigma}_y\mathbf{A}, \tag{13.41}$$

ahol $\mathbf{\Sigma}_y$ a megfigyelt változók variancia-kovarianciamátrixa.

Mivel $\mathbf{A}$ rangja $(m - 1)$, (13.41) $\mathbf{\Sigma}_y$ szinguláris transzformációja, ezért $\mathbf{\Sigma}_x$ is szinguláris lesz, így nem elemezhető ugyanúgy, mint $\mathbf{\Sigma}_y$.

Az $\mathbf{\Sigma}_x$ ipszatív variancia-kovarianciamátrix tulajdonságai (lásd Cleanans, 1956):

– $\mathbf{\Sigma}_x$ sorainak (vagy oszlopainak) az összege nullával egyenlő,

– egy ipszatív változóhalmaz és egy kritérium, függő változó kovarianciáinak az összege egyenlő 0-val,

– ha m, a változók száma nő, $\mathbf{\Sigma}_x$ közelít $\mathbf{\Sigma}_y$-hoz.

Mivel $\mathbf{\Sigma}_x$ nem pozitív definit, nem alkalmazhatjuk a faktorelemzés modelljét közvetlenül a $\mathbf{\Sigma}_x$ mátrixra.

Tételezzük fel viszont, hogy a megfigyelt változókra a faktorelemzés modellje illeszthető:

$$\mathbf{y} = \mathbf{\Lambda}_y \mathbf{f} + \mathbf{e}. \tag{13.42}$$

A modell feltételei:

$$E(\mathbf{y}) = E(\mathbf{f}) = E(\mathbf{e}) = \mathbf{0}$$

$$E(\mathbf{f}\,\mathbf{e}') = \mathbf{0}$$

$$E(\mathbf{f}\,\mathbf{f})' = \mathbf{\Phi}_y$$

$$E(\mathbf{e}\,\mathbf{e}') = \mathbf{\Theta}_y.$$

A faktormodell kovariancia egyenlete:

$$E(\mathbf{y}\,\mathbf{y}') = \mathbf{\Sigma}_y = \mathbf{\Lambda}_y \mathbf{\Phi}_y \mathbf{\Lambda}_y + \mathbf{\Theta}_y. \tag{13.43}$$

A (13.40) egyenlet szerinti ipszatív transzformációt elvégezve, az ipszatív változók faktormodelljét kapjuk:

$$\begin{aligned} \mathbf{x} &= (\mathbf{\Lambda}_y \mathbf{f} + \mathbf{e}) - \mathbf{1} m^{-1} \mathbf{1}' (\mathbf{\Lambda}_y \mathbf{f} + \mathbf{e}) \\ &= \mathbf{\Lambda}_y \mathbf{f} + \mathbf{e} - \mathbf{1}(\overline{\lambda}' \mathbf{f} + \overline{\mathbf{e}}), \end{aligned} \tag{13.44}$$

ahol $\overline{\lambda}'$ az $\mathbf{\Lambda}_y$ faktorsúlymátrix oszlopainak átlagvektora, $\overline{\mathbf{e}}$ a hibakomponensek átlaga.

Tovább egyszerűsítve a (13.44) egyenletet, az ipszatív változók faktormodelljét a következőképpen írhatjuk:

$$\mathbf{x} = (\mathbf{\Lambda}_y - \mathbf{1}\,\overline{\lambda}')\mathbf{f} + \mathbf{A}\,\mathbf{e}, \tag{13.45}$$

ahol $\mathbf{A}\,\mathbf{e} = \mathbf{e} - \mathbf{1}\,\overline{\mathbf{e}}$, $\mathbf{A} = (\mathbf{I} - m^{-1}\mathbf{U})$, vagy

$$\mathbf{x} = \mathbf{\Lambda}_x \mathbf{f} + \mathbf{A}\,\mathbf{e}, \tag{13.46}$$

ahol $\mathbf{\Lambda}_x = (\mathbf{\Lambda}_y - \mathbf{1}\,\overline{\lambda}')$, $\mathbf{A} = (\mathbf{I} - m^{-1}\mathbf{U}) = (\mathbf{I} - m^{-1}\mathbf{1}\,\mathbf{1}')$.

Az ipszatív változók faktorsúlymátrixa az $\mathbf{\Lambda}_y$ faktormátrixnak és az átlagos faktorsúlyok mátrixának a különbsége. A hibakomponensek az ipszatív modellben korrelálnak egymással (mivel $\mathbf{A}$ nem diagonális mátrix).

Az ipszatív $\mathbf{x}$ változók kovarianciamátrixa:

$$E(\mathbf{x}\,\mathbf{x}') = \mathbf{\Sigma}_x = \mathbf{\Lambda}_x \mathbf{\Phi}_y \mathbf{\Lambda}_x + \mathbf{A}\,\mathbf{\Theta}_y \mathbf{A}. \tag{13.47}$$

Látható, hogy az ipszatív változók faktormodelljében a faktorok megegyeznek a preipszatív modell faktoraival, ugyanígy a faktorok kovarianciamátrixa is változatlan maradt. Az ipszatív faktorsúlyok mátrixa az eredeti faktormátrix elemeinek eltérését tartalmazza a faktorsúlymátrix megfelelő oszlopainak átlagától. A hibakomponensek az ipszatív transzformáció hatása miatt korreláltakká váltak.

A gyakorlatban ipszatív tulajdonságú változókhoz általában nem ipszatív transzformációval, hanem elsősorban rangsorolással jutunk akkor, amikor a vizsgált személyektől azt kérjük, hogy valamilyen szempont szerint rangsorolják a változóhalmaz elemeit. Így a mérés eredménye rangszám lesz, és a változóhalmaz ipszatív tulajdonságú.

Mivel $\mathbf{\Sigma}_x$ ebben az esetben szinguláris, a $\mathbf{\Sigma}_x$ mátrixból el kell hagynunk egy sort és egy oszlopot, önkényesen elhagyva az $\mathbf{x}$ változók közül egyet. Jelölje $\mathbf{x}^*$ az eggyel csökkentett számú ipszatív változókat ($\mathbf{x}^*$ $(m-1)$ változót tartalmaz). A $\mathbf{x}^*$ változók $\mathbf{\Sigma}_{x^*}$ kovarianciamátrixához illeszthetjük a faktormodellt. Az elhagyott változó faktorsúlyát a faktormátrix minden oszlopában megkaphatjuk úgy, hogy az $(m-1)$ változó faktorsúlya

összegének az ellentettjét vesszük (–1-gyel szorozzuk) minden oszlopban, és ezt feleltetjük meg a faktormátrix hiányzó változójának súlyával.

Alwin és Jackson (1981) három feltételt nevezett meg, amely teljesülése esetén az ipszatív tulajdonságú változókra alkalmazhatjuk a faktorelemzés modelljét: (1) az adott ipszatív tulajdonságú változóhalmaz mellett létezik a populációban egy ezzel összefüggő, hipotetikus nem ipszatív változóhalmaz, (2) létezik az ipszatív transzformáció a két változóhalmaz között, (3) teljesülnek a faktormodell feltételei a hipotetikus nem ipszatív változókra.

13.7. Dichotom változók faktorelemzése

Jelölje $\mathbf{y}^*$ a megfigyelt manifeszt változókat (számuk m), η pedig a faktorokat (számuk r).

A faktorelemzés modellje:

$$\mathbf{y}^* = \mathbf{\Lambda}\,\boldsymbol{\eta} + \boldsymbol{\epsilon},$$

ahol $\mathbf{\Lambda}$ a faktorsúlyok mátrixa ($m \times r$) típusú, ϵ az m számú reziduális jelöli.

A $\mathbf{y}^*$ megfigyelt változó variancia-kovarianciamátrixa a szokásos feltételezésekkel:

$$V(y^*) = \mathbf{\Sigma} = \mathbf{\Lambda}\,\mathbf{\Phi}\,\mathbf{\Lambda}' + \mathbf{\Theta},$$

ahol Φ a faktor kovarianciamátrix, $\Theta = \mathbf{I} - \mathrm{diag}(\mathbf{\Lambda}\,\mathbf{\Phi}\,\mathbf{\Lambda}')$ Θ a reziduális kovarianciamátrix (diagonális).

Mivel a faktorokat és a reziduálisokat folytonos változóknak tekintjük, így y_i^*-k is folytonosak, azonos skálájú változók. Tegyük fel, hogy minden y_i^* változóhoz tartozik egy dichotom ((0, 1) értékű) y változó. A klasszikus faktorelemzés alkalmazása esetén feltételezhető, hogy $y = y^*$ minden változóra. Ez azonban eléggé önkényes feltételezés, mivel a modell jobb oldalán folytonos, intervallummérési skálájú változók, a bal oldalon diszkrét, dichotom változók állnak. Ez a probléma jól ismert a regresszióelemzés alkalmazásánál, és mint az OLS-regresszió versus logisztikus regresszió polémia jelenik meg az irodalomban. A probléma lényege az, hogy $\mathbf{y}^*$ és η között a modell lineáris, míg $\mathbf{y}$ és η között nemlineáris a kapcsolat.

Tegyük fel, hogy y_i^* változó egy pozitív irányú latens hatást tükröz. Ha ez a hatás eléri az adott változóra jellemző küszöbértéket, akkor a változó kifejez egy szimptómát, egyébként nem. Ha r_i jelöli ezt a küszöbértéket, akkor ezt a következőképpen írhatjuk:

$$y_i = \begin{cases} 0, & \text{ha } y_i^* < r_i \\ 1, & \text{egyébként.} \end{cases}$$

Az y_i változó binomiális eloszlású, feltételezve, hogy a reziduálisok függetlenek η-tól és normális vagy logisztikus eloszlásúak. Az y_i és η közötti nemlineáris kapcsolat:

$$P(y_i = 0 \mid \boldsymbol{\eta}) = P(y_i^* < r_i) = F(r_i - \boldsymbol{\lambda}'\boldsymbol{\eta}),$$

ahol F standard normális, vagy logisztikus eloszlásfüggvény, és feltételezzük, hogy a reziduális variancia standardizált.

Ha az $y = y^*$ megfeleltetést vesszük, és a faktorelemzés modelljét alkalmazzuk, akkor a Pearson-féle korrelációt alkalmazzuk a dichotom változókra, ami azt jelenti, hogy a lineáris modell alapján becsüljük a paramétereket.

Tudjuk, hogy dichotom változók esetén a Pearson-féle ϕ együtthatók (product moment coefficients) nem mérik jól az asszociációt, a maximumuk függ a peremeloszlásoktól, (a ϕ együttható nemcsak a változók kapcsolatának erősségétől függ, hanem a változók várható értékétől is, az 1 értéket csak akkor veszi fel, ha a két változónak azonos a várható értéke). Azonban, ha a változók rendre pozitívan kapcsolódnak egymáshoz, és az eloszlásuk közel azonos irányba és mértékben ferdül, a faktorelemzés a ϕ együtthatók alkalmazása esetén lényegében nem ad torz eredményeket, a faktorstruktúra közel azonos lesz, bár a faktorsúlyok kisebbek lesznek, így a megbízhatóságról hamis képet kapunk. Ha a nemlineáris modellt alkalmazzuk, akkor a tetrakorikus korrelációt kell számítanunk a megfigyelt dichotom változók között, a faktormodellt ehhez kell illesztenünk.

A modell paramétereit ($\mathbf{r}$, λ, ϕ) becsülhetjük a következő függvény minimalizálásával:

$$F = (\mathbf{s} - \boldsymbol{\sigma})' \mathbf{W}^{-1} (\mathbf{s} - \boldsymbol{\sigma}),$$

ahol $\mathbf{s}$ első m eleme a küszöbértékeket, a további $m(m-1)/2$ elem a tetrakorikus korrelációkat tartalmazza, $\boldsymbol{\sigma}$ elemei az $\mathbf{s}$ megfelelői a teljes populációban, $\mathbf{W}$ az $\mathbf{s}$ mintabeli kovarianciamátrixának a becslése.

Ha $\mathbf{W} = \mathbf{I}$, akkor a klasszikus legkisebb négyzetek módszerével egyezik F minimalizálása, egyébként a reziduálisokat súlyozzuk a mintabeli ingadozással, így a paraméterek standard hibája kisebb lesz.

Az általánosított legkisebb négyzetek módszerével egyezik F minimalizálása, egyébként a reziduálisokat súlyozzuk a mintabeli ingadozással, így a paraméterek standard hibája kisebb lesz.

Az általánosított legkisebb négyzetek módszerének előnye továbbá, hogy az F minimuma χ^2 eloszlású (Mathén, 1978). F minimalizálásánál nem okoz ugyan gondot, ha a tetrakorikus korrelációmátrix nem pozitív definit, de ez jelzi, hogy a változók mögötti eloszlás nem normális eloszlású. Ilyenkor fel kell tételeznünk azt is, hogy a mögöttes latens faktorok eloszlása nem normális. Muthén (1989) javasolta a többváltozós probit regresszió alkalmazását erre az esetre, és példát is ad mindkét modell alkalmazására.

13.8. Szimultán faktorelemzés a faktoriális invariancia vizsgálatára

Ha azonos változóhalmaz faktormodelljét becsüljük több mintában, felmerül a faktorok összehasonlításának problémája, a faktorok hasonlóságának mérése.

Az exploratív faktorelemzésnél vizsgáltuk a faktorok hasonlóságának mérésére kidolgozott mutatókat. Ezeket elsősorban az exploratív elemzés eredményeinek összehasonlításánál alkalmazhatjuk. A faktoriális invariancia problémája a konfirmatív faktorelemzés témaköréhez kötődik. Jöreskog (1971), Lawley és Maxwell (1971), Sörbom és Jöreskog (1976) javasolta a szimultán faktorelemzés módszerét a faktoriális invariancia vizsgálatára.

Ha például két országban vizsgáljuk a mérési modellt, a következő hipotéziseket tesztelhetjük:

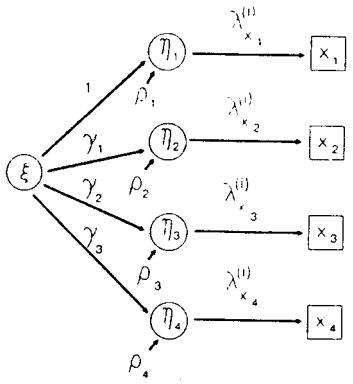
a) $\boldsymbol{\Lambda}^{(1)} = \boldsymbol{\Lambda}^{(2)}$ a két országban azonos a mérési modell,

b) $\boldsymbol{\Theta}^{(1)} = \boldsymbol{\Theta}^{(2)}$ a hibakomponensek varianciái egyenlők,

c) $\boldsymbol{\Phi}^{(1)} = \boldsymbol{\Phi}^{(2)}$ a faktorok variancia-kovarianciamátrixai egyenlők.

A két ország különbségét a congeneric mérési modellt feltételezve úgy írjuk le, hogy különbség van a két ország között variabilitásban, de egyébként a faktormodell invariáns a két országban. Ez azt jelenti, hogy a szórások különbözhetnek az országok között, de a korrelációmátrix megegyezik. Másképpen fogalmazva, a kovarianciamátrix csak a megfigyelt változók skálafaktoraiban különbözik.

Ezt mutatja a következő ábra:



ahol γ-k a faktorsúlyok, λ-k a skálafaktorok.

13.8.1. Szimultán faktorelemzés

Feltételezzük, hogy a különböző populációkban (vagy részpopulációkban) azonos m manifeszt változó faktormodelljét a következőképpen írhatjuk:

$$\mathbf{x}_g = \boldsymbol{\nu}_g + \boldsymbol{\Lambda}_g \boldsymbol{\xi}_g + \boldsymbol{\epsilon}_g, \tag{13.48}$$

ahol $g = 1, 2, \ldots, s$,

$\boldsymbol{\nu}_g$ az m manifeszt változó mérési skáláinak középpontjait, nullpontjait tartalmazó oszlopvektor,

$\boldsymbol{\xi}_g$ a latens (hipotetikus) közös faktorok ($r \times 1$) típusú vektora,

$\boldsymbol{\epsilon}_g$ az egyedi faktorok (hibakomponensek) ($m \times 1$) típusú vektora,

$\boldsymbol{\Lambda}_g$ a faktorsúlyok ($m \times r$) típusú mátrixa.

$\boldsymbol{\nu}_g$ általánosan a változók mérési skálájának középpontja, kiindulópontja, nullpontja, a gyakorlatban legtöbbször az m változó várható értékeinek vektora (Sörbom és Jöreskog [1976] lokális paraméternek nevezi).

Általában azt szokták feltételezni, hogy $E(\boldsymbol{\xi}_g) = \mathbf{0}$, a következőkben azonban a faktorokat értelmezhető skálán mérjük.

A faktormodell feltételei a g-edik populációban:

1. $E(\boldsymbol{\xi}_g) = \boldsymbol{\theta}_g$ a faktorok várható értékvektora
2. $E(\boldsymbol{\epsilon}_g) = \mathbf{0}$
3. $E(\boldsymbol{\epsilon}_g \boldsymbol{\epsilon}_g') = \Theta_g$ diagonális
4. $E(\boldsymbol{\xi}_g \boldsymbol{\epsilon}_g') = \mathbf{0}$.

A 2–4. feltételek figyelembevételével $\mathbf{x}_g$ variancia-kovarianciamátrixát következőképpen írhatjuk:

$$\mathbf{\Sigma}_g = \mathbf{\Lambda}_g \mathbf{\Phi}_g \mathbf{\Lambda}_g' + \mathbf{\Theta}_g, \tag{13.49}$$

ahol $E(\boldsymbol{\xi}_g \boldsymbol{\xi}_g') = \mathbf{\Phi}_g$ a latens közös faktorok variancia-kovarianciamátrixa, $E(\boldsymbol{\epsilon}_g \boldsymbol{\epsilon}_g') = \mathbf{\Theta}_g$ az egyedi faktorok (hibakomponensek) variancia-kovarianciamátrixa (feltételezzük, hogy diagonális).

A modell paramétereinek becslése előtt a modell identifikálhatóságát kell vizsgálni ahhoz, hogy a paraméterek egyértelmű becslését megkaphassuk. Az identifikálhatóságnál a $\mathbf{\Lambda}_g, \mathbf{\Phi}_g, \mathbf{\Theta}_g$ paramétermátrixok és a $\mathbf{\Sigma}_g$ variancia-kovarianciamátrix egyértelmű megfeleltetését kell biztosítani. Elégséges feltétele az identifikálhatóságnak, hogy a modell egyenleteinek száma nagyobb vagy egyenlő legyen független paramétereinek számánál.

A becsléshez feltételezzük, hogy $\mathbf{x}_g$ többdimenziós normális eloszlású. A g-edik mintában a likelihood-függvény logaritmusa

$$\log_e L_g = -\frac{1}{2}(n_g - 1)\left[\log_e |\mathbf{\Sigma}_g| + tr(\mathbf{S}_g \mathbf{\Sigma}_g^{-1})\right],$$

ahol n_g a g-edik minta elemszáma, $\mathbf{S}_g$ a mintabeli variancia-kovarianciamátrix.

Mivel a minták függetlenek, az összes mintára a likelihood-függvény logaritmusa:

$$\log_e L = \sum_{g=1}^{s} \log_e L_g.$$

A gyakorlatban $\log_e L$ maximalizálása helyett az

$$F = \sum_{g=1}^{s} (n_g - 1)\left[tr(\mathbf{S}\,\mathbf{\Sigma}_g^{-1}) + (\log |\mathbf{\Sigma}_g| - \log |\mathbf{S}_g|) - m_g\right]$$

függvényt minimalizáljuk. A Fletcher és Powell (1963) -féle eljárással minimalizáljuk az F függvényt $\mathbf{\Lambda}_g$, $\mathbf{\Phi}_g$ és $\mathbf{\Theta}_g$ ismeretlen paraméterei szerint.

A modell illeszkedését khi-négyzet statisztikával tesztelhetjük.

$$v = \min F(\mathbf{S}_g, \mathbf{\Sigma}_g(\mathbf{\Lambda}_g, \mathbf{\Phi}_g, \mathbf{\Theta}_g))$$

nagy minták esetén khi-négyzet eloszláshoz tart

$$d = \sum_{g=1}^{s} \frac{1}{2} m_g(m_g + 1) - t$$

szabadságfok mellett, ahol t a független paraméterek száma a modellben.

Az $\mathbf{x}_g$ megfigyelt változók kovarianciastruktúrája független a változók mérésétől, ν_g-től és a faktorok várható értékétől, θ_g-től.

A megfigyelt változók várható értéke a faktorok várható értékének lineáris függvénye:

$$E(\mathbf{x}_g) = \boldsymbol{\mu}_g = \boldsymbol{\nu}_g + \mathbf{\Lambda}_g \boldsymbol{\theta}_g. \tag{13.50}$$

Ha $\theta_g = \mathbf{0}$ (a faktorok várható értéke nulla), akkor $\mu_g = \nu_g$, (a megfigyelt változók várható értéke ν_g-vel egyenlő). Ebben a modellben a faktorok mérése érdektelen. Ha $\nu_g = 0$, akkor $\mu_g = \mathbf{\Lambda}_g \theta_g$, vagyis a megfigyelt változók várható értéke a faktorok várható értékének lineáris függvénye, de a lineáris függvényben nincs additív konstans. Ezt a modellt akkor használjuk, amikor a faktorok várható értékei fontosak, de a megfigyelt változók mérésének lokalizálása érdektelen. A továbbiakban feltételezzük, hogy $\nu_g = \mathbf{0}$.

A közös faktormodell paramétereinek becsléséhez feltételezzük, hogy a modell identifikálható. Az identifikálhatóság szükséges feltétele, hogy a Λ_g, Φ_g és Θ_g paraméterekre a modell legalább r^2 független feltételt tartalmazzon (lásd Jöreskog, 1971, Sörbom, 1974, Sörbom és Jöreskog, 1976).

13.8.2. Faktoriális invariancia

A faktoriális invariancia vizsgálata négy kérdést jelent. Az első kettő a mérési struktúra (Λ_g és Θ_g) invarianciáját vizsgálja, a másik kettő a faktortérben elemzi a modell (θ_g és Φ_g) invarianciáját.

A faktoriális invariancia Thurstone-féle vizsgálata (Thurstone, 1947) leszűkül a faktorsúlymátrix invarianciájának vizsgálatára (Harmat, 1967; Rummel, 1970), a Λ_g mátrix azonosságának vizsgálatára két vagy több populációban. Az exploratív faktorelemzésnél a faktorok hasonlóságának mérőszámait a standardizált faktorokra alkalmaztuk. Ha $\mathbf{S} = (\text{diag}\Phi)^{-\frac{1}{2}}$ és $\mathbf{D} = (\text{diag}\Sigma)^{-\frac{1}{2}}$, a Λ faktorsúlymátrixot transzformálhatjuk, hogy elmei $[-1;\ +1]$ között mozogjanak, ahogyan az exploratív faktorelemzésnél ezt feltételezzük:

$$\mathbf{\Lambda}^* = \mathbf{D}\,\mathbf{\Lambda}\,\mathbf{S}^{-1}. \tag{13.51}$$

A Λ^* „standardizált” faktorsúlymátrix függ mind a Λ, mind a Φ mátrixtól (és természetesen $\mathbf{D}$-től, a változók varianciáitól is), így a Λ^* mátrix hasonlóságának vizsgálata nem különíti el Λ és Φ hatásait, összemossa Λ és Φ invarianciájának vizsgálatát.

Először Jöreskog (1971) és Sörbom (1974) a szimultán faktorelemzés modelljében különítette el Λ_g, Φ_g és Θ_g invarianciájának vizsgálatát, ezt egészítette ki Alwin és Jackson (1981) θ_g invarianciájának vizsgálatával.

A következőkben a faktoriális invariancia hipotéziseit vizsgáljuk.

1.1. Hipotézis

Az első hipotézis, hogy a modell három paramétermátrixa, Λ_g, Φ_g és Θ_g invariáns a különböző populációkban, amiből következik az a hipotézis, hogy a populációk megfigyelt kovarianciastruktúrái egyenlőek:

1.1. $$H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \ldots = \mathbf{\Sigma}_s.$$

A hipotézis teszteléséhez a statisztikát Jöreskog (1971) dolgozta ki.

A statisztika közelítőleg χ^2 eloszlást követ $\frac{1}{2}(m-1)p(p+1)$ szabadságfok mellett (χ^2_Σ).

Ha a hipotézist nem vetjük el, akkor bizonyos hiba megengedésével azt tételezhetjük fel, hogy

$$\mathbf{\Lambda}_1 = \mathbf{\Lambda}_2 = \ldots = \mathbf{\Lambda}_s, \quad \mathbf{\Phi}_1 = \mathbf{\Phi}_2 = \ldots = \mathbf{\Phi}_s \quad \text{és}$$

$$\mathbf{\Theta}_1 = \mathbf{\Theta}_2 = \ldots = \mathbf{\Theta}_s.$$

A továbbiakban ezeknek a mátrixoknak az invarianciáját teszteljük.

1.2. Hipotézis

Ha elfogadjuk azt a hipotézist, hogy a különböző populációk kovarianciastruktúrája egyenlő, vagyis egyenlőek a faktorsúlymátrixok, megvizsgálhatjuk a faktorok várható értékeinek egyenlőségét:

1.2. $$H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = \ldots = \boldsymbol{\theta}_s.$$

A hipotézist úgy teszteljük, hogy megvizsgáljuk a

$$H_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \ldots = \boldsymbol{\Sigma}_s \quad \text{és} \quad \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_s$$

hipotézisnek megfelelő statisztikát, amely $\frac{1}{2}(s-1)m(m+1) + m(s-1)$ szabadságfokú χ^2 eloszlást követ. Jelöljük $\chi^2_{\boldsymbol{\Sigma}\,\boldsymbol{\theta}}$-vel, és értékét összehasonlítjuk a $\chi^2_{\boldsymbol{\Sigma}}$ értékkel (1.1. hipotézis), és így az 1.2. H_0 hipotézist a

$$\chi^2_{\boldsymbol{\theta}|\boldsymbol{\Sigma}} = \chi^2_{\boldsymbol{\Sigma}\,\boldsymbol{\theta}} - \chi^2_{\boldsymbol{\Sigma}} \quad m(s-1) \quad \text{szabadságfokú}$$

χ^2 statisztikával teszteljük.

Ha $\chi^2_{\boldsymbol{\theta}|\boldsymbol{\Sigma}}$ szignifikáns, akkor a faktorok várható értékei különböznek. Ha a statisztika nem szignifikáns, akkor mivel $\boldsymbol{\mu}_g = \boldsymbol{\Lambda}_g \boldsymbol{\theta}_g$, azt mondhatjuk, hogy $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = \ldots = \boldsymbol{\theta}_s$.

2.1. Hipotézis

Ha az 1.1 hipotézist elvetjük, akkor külön-külön kell megvizsgálni a paramétermátrixok invarianciáját. Először a faktorsúlymátrix invarianciáját vizsgáljuk. A nullhipotézis

2.1. $$H_0 : \boldsymbol{\Lambda}_1 = \boldsymbol{\Lambda}_2 = \ldots = \boldsymbol{\Lambda}_s.$$

Ez a hipotézis feltételezi, hogy a populációkban a közös faktorok száma megegyezik, és a faktorok struktúrája is azonos.

Ha χ^2_r jelöli azt a statisztikát, amely a faktorok azonos számát teszteli, $\chi^2_{\boldsymbol{\Lambda}}$ pedig annak a modellnek az illeszkedését méri, amelynek paramétermátrixai $\boldsymbol{\Lambda}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots \boldsymbol{\theta}_s$, $\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2, \ldots, \boldsymbol{\Phi}_s, \boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \ldots, \boldsymbol{\Theta}_s$, akkor a 2.1. H_0 hipotézist a

$$\chi^2_{\boldsymbol{\Lambda}|r} = \chi^2_{\boldsymbol{\Lambda}} - \chi^2_r$$

$\text{szf}_{\boldsymbol{\Lambda}|r} = \text{szf}_{\boldsymbol{\Lambda}} - \text{szf}_r$ szabadságfokú χ^2 eloszlású statisztika teszteli, ahol $\text{szf}_r = s\left[\frac{1}{2}m(m+1) + m - mr + q - \frac{1}{2}r(r-1) - m - r\right]$, q a $\boldsymbol{\Lambda}_g$-ben rögzített elemek száma, $\text{szf}_{\boldsymbol{\Lambda}} = \frac{1}{2}sm(m+1) - mr + q - \frac{1}{2}sr(r+1) - sm$.

Ha a χ^2 érték szignifikánsan nagyobb a feltételeket tartalmazó modellben, mint a feltételeket nem tartalmazó modellben ($\chi^2_{\boldsymbol{\Lambda}|r} > 0$), akkor a faktorsúlymátrixok invarianciájának hipotézisét elvetjük.

2.2. Hipotézis

Ha a 2.1 hipotézist nem vetettük el, akkor megvizsgáljuk a faktorok várható értékei egyenlőségét. A hipotézis:

2.2. $$H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = \ldots = \boldsymbol{\theta}_s.$$

Először illesztjük azt a modellt, amelynek paramétermátrixai $\boldsymbol{\Lambda}$, $\boldsymbol{\theta}$, $\boldsymbol{\Phi}_1$, $\boldsymbol{\Phi}_2$, …, $\boldsymbol{\Phi}_s$, $\boldsymbol{\Theta}_1$, $\boldsymbol{\Theta}_2$, …, $\boldsymbol{\Theta}_s$. A modell illeszkedését a $\chi^2_{\boldsymbol{\Lambda}\,\boldsymbol{\theta}}$ statisztika méri. A $\chi^2_{\boldsymbol{\Lambda}\,\boldsymbol{\theta}}$ statisztikának és a 2.1. hipotézist tesztelő $\chi^2_{\boldsymbol{\Lambda}}$ statisztikának a

$$\chi^2_{\boldsymbol{\theta}|\boldsymbol{\Lambda}} = \chi^2_{\boldsymbol{\Lambda}\boldsymbol{\theta}} - \chi^2_{\boldsymbol{\Lambda}}$$

$$\text{szf}_{\boldsymbol{\theta}|\boldsymbol{\Lambda}} = \text{szf}_{\boldsymbol{\Lambda}\boldsymbol{\theta}} - \text{szf}_{\boldsymbol{\Lambda}}$$

különbsége teszteli a 2.2. hipotézist, ahol $\text{szf}_{\boldsymbol{\theta}|\boldsymbol{\Lambda}} = \frac{1}{2}sm(m+1) - mr - r + q - \frac{1}{2}sr(r+1)$, q a $\boldsymbol{\Lambda}$ mátrix kötött elemeinek a száma.

2.3. Hipotézis

Ha a faktorstruktúra invariáns, akkor a faktorok várható értékei azonossága után (mellett) a hibakomponensek kovarianciastruktúrájának invarianciáját vizsgálhatjuk:

2.3. $$H_0: \boldsymbol{\Theta}_1 = \boldsymbol{\Theta}_2 = \ldots = \boldsymbol{\Theta}_s.$$

Ezt a hipotézist is feltételes hipotézisként értelmezzük. Illesztjük a $\boldsymbol{\Lambda}$, $\boldsymbol{\Theta}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_s$, $\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2, \ldots, \boldsymbol{\Phi}_s$ paraméterű modellt. Az illeszkedés jóságát a $\chi^2_{\boldsymbol{\Lambda}\,\boldsymbol{\Theta}}$ statisztika méri, $\mathrm{szf}_{\boldsymbol{\Lambda}\,\boldsymbol{\Theta}} = 1/2sm(m+1) + sm - mr + q - 1/2sr(r+1) - m - sr$.

A 2.3. H_0 hipotézist a

$$\chi^2_{\boldsymbol{\theta}|\boldsymbol{\Lambda}} = \chi^2_{\boldsymbol{\Lambda}\,\boldsymbol{\Theta}} - \chi^2_{\boldsymbol{\Lambda}}$$

$$\mathrm{szf}_{\boldsymbol{\Theta}|\boldsymbol{\Lambda}} = \mathrm{szf}_{\boldsymbol{\Lambda}\boldsymbol{\Theta}} - \mathrm{szf}_{\boldsymbol{\Lambda}} \quad \text{szabadságfokú}$$

χ^2 statisztikával teszteljük.

2.4. Hipotézis

Ha a faktorstruktúra invarianciájának hipotézisét (2.1) elfogadjuk, a harmadik paramétermátrix, amit tesztelnünk kell, a $\boldsymbol{\Phi}$. A nullhipotézis:

2.4. $$H_0: \boldsymbol{\Phi}_1 = \boldsymbol{\Phi}_2 = \ldots = \boldsymbol{\Phi}_s.$$

Az előzőekhez hasonlóan először a $\boldsymbol{\Lambda}, \boldsymbol{\Phi}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_s$, $\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \ldots, \boldsymbol{\Theta}_s$ modellt illesztjük, és a $\boldsymbol{\Phi}_s$-re vonatkozó invariancia hipotézist a következő statisztikával tesztelük:

$$\chi^2_{\boldsymbol{\Phi}|\boldsymbol{\Lambda}} = \chi^2_{\boldsymbol{\Lambda}\,\boldsymbol{\Phi}} - \chi^2_{\boldsymbol{\Lambda}}$$

$$\mathrm{szf}_{\boldsymbol{\Phi}|\boldsymbol{\Lambda}} = \mathrm{szf}_{\boldsymbol{\Lambda}\,\boldsymbol{\Phi}} - \mathrm{szf}_{\boldsymbol{\Lambda}},$$

ahol $\mathrm{szf}_{\boldsymbol{\Lambda}\,\boldsymbol{\Phi}} = \frac{1}{2}sm(m+1) - mr + m - \frac{1}{2}r(r+1) - sr$.

Ha $\chi^2_{\boldsymbol{\Phi}|\boldsymbol{\Lambda}}$ szignifikáns, a faktorok kovarianciastruktúrájának invarianciájára vonatkozó hipotézist el kell vetnünk.

3.1. Hipotézis

Jöreskog (1971) javasolta a fenti hipotézisek mellett $\boldsymbol{\Phi}_g$ invarianciájának vizsgálatát a $\boldsymbol{\Lambda}_g$ és $\boldsymbol{\Theta}_g$ invariáns paraméterek feltételezésével.

Először a $\boldsymbol{\Lambda}, \boldsymbol{\Theta}, \boldsymbol{\Phi}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_s$ modellt illesztjük, majd a $\boldsymbol{\Lambda}, \boldsymbol{\Theta}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_s$, $\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2, \ldots, \boldsymbol{\Phi}_s$ modellt, és a két modell illeszkedésének különbségével teszteljük $\boldsymbol{\Phi}_g$ invarianciáját. A statisztika:

$$\chi^2_{\boldsymbol{\Phi}|\boldsymbol{\Lambda}\,\boldsymbol{\Theta}} = \chi^2_{\boldsymbol{\Lambda}\,\boldsymbol{\Theta}\,\boldsymbol{\Phi}} - \chi^2_{\boldsymbol{\Lambda}\,\boldsymbol{\Theta}}$$

$$\mathrm{szf}_{\boldsymbol{\Phi}|\boldsymbol{\Lambda}\,\boldsymbol{\Theta}} = \mathrm{szf}_{\boldsymbol{\Lambda}\,\boldsymbol{\Theta}\,\boldsymbol{\Phi}} - \mathrm{szf}_{\boldsymbol{\Lambda}\,\boldsymbol{\Theta}},$$

ahol $\mathrm{szf}_{\boldsymbol{\Lambda}\,\boldsymbol{\Theta},\,\boldsymbol{\Phi}} = \frac{1}{2}sm(m+1) + sm - mr + q - \frac{1}{2}sr(r+1) - m - sr$.

A fenti hipotéziseket áttekintve azt láthatjuk, hogy az előzetesen említett faktoriális invariancia vizsgálat két csoportba sorolt négy kérdése előtt két előzetes hipotézist is megvizsgáltunk, nevezetesen $\boldsymbol{\Sigma}_g$ és μ_g invarianciájának hipotéziseit. A faktoriális invariancia tesztelése csak két előzetes hipotézis elfogadása után értelmes.

A mérési modell invarianciájának tesztelése a $\boldsymbol{\Lambda}_g$ és $\boldsymbol{\Theta}_g$ paramétermátrixok invarianciájának tesztelését jelenti. Először $\boldsymbol{\Lambda}_g$ invarianciáját teszteltük (2.1. hipotézis), majd $\boldsymbol{\Theta}_g$ invarianciáját azzal a feltételezéssel, hogy a 2.1. hipotézist nem vetettük el. A faktorok várható értékének ($\boldsymbol{\theta}_g$) invarianciáját fogalmazta meg a 2.2. hipotézis, míg a faktorkovariancia-struktúra invarianciáját ($\boldsymbol{\Phi}$) a 2.4. hipotézis. Ez utóbbi kettő már nem a mérési modellel, hanem a latens faktorok tulajdonságaival függ össze. Jöreskog (1971)

javasolta, hogy Φ_g invarianciáját mind Λ_g, mind Θ_g invarianciájának feltevésével teszteljük. Ezt fejezte ki a 3.1. hipotézis.

Összefoglalóan a faktoriális invariancia vizsgálatára a következő hipotéziseket fogalmaztuk meg:

Előzetes hipotézisek:

1.1. $H_0 : \Sigma_1 = \Sigma_2 = \ldots = \Sigma_s$

1.2. $H_0 : \Sigma_1 = \Sigma_2 = \ldots = \Sigma_s$ és $\mu_1 = \mu_2 = \ldots = \mu_s$.

Specifikus invariancia hipotézisek:

2.1. $H_0 : \Lambda_1 = \Lambda_2 = \ldots = \Lambda_s$

2.2. $H_0 : \Lambda_1 = \Lambda_2 = \ldots = \Lambda_s$ és $\theta_1 = \theta_2 = \ldots = \theta_s$

2.3. $H_0 : \Lambda_1 = \Lambda_2 = \ldots = \Lambda_s$ és $\Theta_1 = \Theta_2 = \ldots = \Theta_s$

2.4. $H_0 : \Lambda_1 = \Lambda_2 = \ldots = \Lambda_s$ és $\Phi_1 = \Phi_2 = \ldots = \Phi_s$.

3.1. $H_0 : \Lambda_1 = \Lambda_2 = \ldots = \Lambda_s$ és $\Theta_1 = \Theta_2 = \ldots = \Theta_s$ és $\Phi_1 = \Phi_2 = \ldots = \Phi_s$.

A 2.2., 2.3. és 2.4. hipotézisek feltételezik, hogy a 2.1. hipotézist nem utasítottuk el, a 3.1. hipotézisnél pedig azt tételeztük fel, hogy a 2.3. hipotézist nem utasítottuk el.

13.9. A faktorértékek becslése

A faktormodell Λ, Φ, Θ paramétereinek becslése és értelmezése után érdekelhet bennünket, hogy a közös faktorok latens értékei hogyan becsülhetők, a közös faktorokon az egyes megfigyelések hol helyezkednek el. Elméletileg ez a probléma a megfigyelési egységek $\mathbf{x}$ vektorainak leképezését jelenti a ξ faktortérbe. Ilyen értelemben nem is helyes a faktorértékek becsléséről beszélni, sokkal inkább a ξ valószínűségi változó $\mathbf{x}$ valószínűségi változóra vonatkozó feltételes *a posteriori* eloszlásán alapuló leképezésekről. Először természetesen fel kell tételeznünk a faktorok *a priori* eloszlását. A közös faktorok eloszlásáról feltételezzük, hogy közelítőleg normális eloszlású, $\mathbf{0}$ várható értékkel, $\mathbf{I}$ varianciakovarianciamátrixszal (vagyis a faktorok korrelálatlanok és egységnyi varianciájúak):

$$\boldsymbol{\xi} \sim N_r(\mathbf{0}, \mathbf{I}). \tag{13.52}$$

A faktormodell ($\mathbf{x} = \mu + \Lambda\,\xi + \epsilon$) feltételezésével az $\mathbf{x}$ valószínűségi változó feltételes eloszlása normális eloszlást követ:

$$\mathbf{x} \mid \boldsymbol{\xi} \sim N_m(\boldsymbol{\mu} + \boldsymbol{\Lambda}\,\boldsymbol{\xi}, \boldsymbol{\Theta}), \tag{13.53}$$

ahol Λ a faktorsúlyok ($m \times r$) típusú mátrixa,

Θ a hibakomponensek varianciáinak diagonális mátrixa,

μ a megfigyelt változók várható értékeinek vektora.

A megfigyelt változók valószínűségeloszlása több dimenziós normális eloszlású:

$$\mathbf{x} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Lambda}\,\boldsymbol{\Lambda}' + \boldsymbol{\Theta}), \tag{13.54}$$

feltételezve a közös faktorok korrelálatlanságát.

A fenti eloszlások alapján a ξ közös faktorok *a posteriori* valószínűségeloszlása szintén normális:

$$\boldsymbol{\xi} \mid \mathbf{x} \sim N_r\{\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}), (\boldsymbol{\Lambda}'\boldsymbol{\Theta}^{-1}\boldsymbol{\Lambda} + \mathbf{I})^{-1}\}. \tag{13.55}$$

A közös faktorok feltételes várható értéke:

$$E(\boldsymbol{\xi} \mid \mathbf{x}) = \boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = (\mathbf{I} + \boldsymbol{\Gamma})^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Theta}^{-1}(\mathbf{x} - \boldsymbol{\mu}), \tag{13.56}$$

ami a következő azonosság felhasználásával könnyen igazolható:

$$\boldsymbol{\Lambda}'\boldsymbol{\Theta}^{-1}\boldsymbol{\Sigma} = \boldsymbol{\Lambda}'\boldsymbol{\Theta}^{-1}(\boldsymbol{\Lambda}\,\boldsymbol{\Lambda}' + \boldsymbol{\Theta}) = (\mathbf{I} + \boldsymbol{\Gamma})\boldsymbol{\Lambda}'.$$

ahol $\boldsymbol{\Gamma} = \boldsymbol{\Lambda}'\boldsymbol{\Theta}^{-1}\boldsymbol{\Lambda}$.

A faktorértékeket tehát megkaphatjuk ξ *a posteriori* várható értéke alapján. Természetesen a paramétermátrixoknak csak a becsléseit ismerjük, ezért a faktorértékeknek is egy becslését kaphatjuk így meg.

A faktorértékek azonos becsléséhez jutunk a Thomson (1951) által javasolt módszer alapján, amelyik a faktorértékek regressziós becslését keresi. Induljunk ki az $\mathbf{x} = \boldsymbol{\Lambda}\,\boldsymbol{\xi} + \boldsymbol{\epsilon}$ faktormodellből. Most az egyszerűség kedvéért feltételezzük, hogy $E(\mathbf{x}) = \boldsymbol{\mu} = \mathbf{0}$, és hogy a közös faktorok korrelálatlanok. Ekkor

$$E(\mathbf{x}\,\boldsymbol{\xi}') = E((\boldsymbol{\Lambda}\,\boldsymbol{\xi} + \boldsymbol{\epsilon})\boldsymbol{\xi}') = \boldsymbol{\Lambda}\,E(\boldsymbol{\xi}\,\boldsymbol{\xi}') = \boldsymbol{\Lambda},$$

vagyis korrelálatlan faktorok esetén a faktorsúlyok mátrixa ($\boldsymbol{\Lambda}$) a változók és a faktorok közötti korrelációs együtthatókat tartalmazza.

Továbbá a megfigyelt vektorok variancia-kovarianciamátrixa:

$$E(\mathbf{x}\,\mathbf{x}') = \boldsymbol{\Sigma} = \boldsymbol{\Lambda}\,\boldsymbol{\Lambda}' + \boldsymbol{\Theta}.$$

A faktorelemzés $\mathbf{x} = \boldsymbol{\Lambda}\,\boldsymbol{\xi} + \boldsymbol{\epsilon}$ egyenletéből a faktorokat kifejezhetjük mint a megfigyelt változók lineáris függvényét:

$$(\mathbf{B}\,\boldsymbol{\Lambda})^{-1}\mathbf{B}\,\mathbf{x} = \boldsymbol{\xi} + (\mathbf{B}\,\boldsymbol{\Lambda})^{-1}\mathbf{B}\boldsymbol{\epsilon}, \tag{13.57}$$

vagy

$$\widehat{\boldsymbol{\xi}} = \mathbf{A}\,\mathbf{x} = \boldsymbol{\xi} + \mathbf{u},$$

ahol $\mathbf{A} = (\mathbf{B}\,\boldsymbol{\Lambda})^{-1}\mathbf{B}$ és $\mathbf{A}\,\boldsymbol{\varepsilon} = \mathbf{u}$

Keressük tehát azt a $\widehat{\boldsymbol{\xi}}$-t, amelyik jó becslését adja a $\boldsymbol{\xi}$-nek.

Tekintsük a k-adik faktor, ξ_k becslését:

$$\widehat{\xi}_k = \mathbf{a}_k'\mathbf{x} = \mathbf{x}'\mathbf{a}_k,$$

ahol $\mathbf{a}_k'$ az $\mathbf{A}$ együtthatómátrix k-adik sorvektora, m-elemű.

Keressük azt az $\mathbf{a}_k$ vektort, amely mellett $(\widehat{\xi}_k - \xi_k)$ varianciája minimális.

Ezt a varianciát a következőképpen írhatjuk:

$$F_k = E(\widehat{\xi}_k - \xi_k)^2 = E(\mathbf{x}'\mathbf{a}_k - \xi_k)^2.$$

Az F_k függvényt deriváljuk $\mathbf{a}_k$ szerint:

$$\frac{\partial F_k}{\partial \mathbf{a}_k} = E[2\mathbf{x}(\mathbf{x}'\mathbf{a}_k - \xi_k)] = 2(\boldsymbol{\Sigma}\,\mathbf{a}_k - \boldsymbol{\lambda}_k),$$

ahol $\boldsymbol{\lambda}_k$ a $\boldsymbol{\Lambda}$ mátrix k-adik oszlopvektora.

A parciális deriváltakat egyenlővé tesszük 0-val:

$$\boldsymbol{\Sigma}\,\mathbf{a}_k = \boldsymbol{\lambda}_k$$

$$\mathbf{a}_k = \boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda}_k,$$

és így a k-adik faktor becslése

$$\widehat{\xi}_k = \boldsymbol{\lambda}_k'\boldsymbol{\Sigma}^{-1}\mathbf{x}.$$

Ha $\widehat{\boldsymbol{\xi}}$ jelöli a $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_r$ faktorbecslések vektorát, akkor a faktorértékek regressziós becslése:

$$\widehat{\boldsymbol{\xi}} = \boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\mathbf{x}. \tag{13.58}$$

A $\boldsymbol{\Lambda}\,\boldsymbol{\Sigma}^{-1}$ együtthatómátrix – mint azt a (13.56) egyenletnél láttuk – egyenlő a $(\mathbf{I} + \boldsymbol{\Gamma})^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Theta}^{-1}$-vel, így a (13.58) egyenlet alternatív formája

$$\widehat{\boldsymbol{\xi}} = (\mathbf{I} + \boldsymbol{\Gamma})^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Theta}^{-1}\mathbf{x}, \tag{13.59}$$

ahol $\boldsymbol{\Gamma} = \boldsymbol{\Lambda}'\boldsymbol{\Theta}^{-1}\boldsymbol{\Lambda}$.

A becslések $E(\widehat{\boldsymbol{\xi}}\,\widehat{\boldsymbol{\xi}}')$ kovarianciamátrix és $E(\widehat{\boldsymbol{\xi}}\,\boldsymbol{\xi}')$ megegyezik:

$$E(\widehat{\boldsymbol{\xi}}\,\widehat{\boldsymbol{\xi}}') = E(\widehat{\boldsymbol{\xi}}\,\boldsymbol{\xi}') = \boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} = \mathbf{I} - (\mathbf{I} + \boldsymbol{\Gamma})^{-1}. \tag{13.60}$$

A becslés hibája $(\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi})$, és a hiba varianciamátrixa:

$$E[(\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi})(\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi})'] = \mathbf{I} - \boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} = (\mathbf{I} + \boldsymbol{\Gamma})^{-1}. \tag{13.61}$$

A fenti egyenletből látható, hogy a becslés akkor lesz jó, ha $(\mathbf{I} + \boldsymbol{\Gamma})^{-1}$ diagonális elemei kicsik. A $\boldsymbol{\Gamma}$ mátrix ($\boldsymbol{\Gamma} = \boldsymbol{\Lambda}'\boldsymbol{\Theta}^{-1}\boldsymbol{\Lambda}$) a faktormodell identifikációs feltételei szerint diagonális, és a diagonális elemei lehetőleg nagyok. Ez igaz az első elemekre, és így az első faktorok értékeinek becslései jók lesznek, de ahogyan csökken a faktorok varianciája, úgy romlik a faktorértékek becslésének jósága.

A $\widehat{\boldsymbol{\xi}}$ regressziós becslésnek az elméleti $\boldsymbol{\xi}$ értékekre vonatkozó feltételes várható értéke eltér az elméleti értékektől ($\boldsymbol{\xi}$-től), így ez a becslés torzított:

$$\begin{aligned} E(\widehat{\boldsymbol{\xi}} \mid \boldsymbol{\xi}) &= E(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\mathbf{x} \mid \boldsymbol{\xi}) \\ &= (\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})\xi \\ &= \boldsymbol{\xi} - (\mathbf{I} + \boldsymbol{\Gamma})^{-1}\boldsymbol{\xi}, \end{aligned} \tag{13.62}$$

ahol felhasználtuk, hogy

$$E(\mathbf{x} \mid \boldsymbol{\xi}) = \boldsymbol{\Lambda}\,\boldsymbol{\xi},$$

mivel a hibatagok korrelálatlanok a közös faktorokkal.

Ha feltételezzük, hogy a faktorok korrelálnak egymással és kovarianciamátrixuk $\boldsymbol{\Phi}$, akkor a fenti azonosságok a következőképpen változnak:

$$\begin{aligned} E(\boldsymbol{\xi}\,\boldsymbol{\xi}') &= \boldsymbol{\Phi} \\ E(\mathbf{x}\,\boldsymbol{\xi}') &= E[(\boldsymbol{\Lambda}\,\boldsymbol{\xi} + \mathbf{e})\boldsymbol{\xi}'] = \boldsymbol{\Lambda}\,\boldsymbol{\Phi} \\ E(\mathbf{x}\,\mathbf{x}') &= \boldsymbol{\Sigma} = \boldsymbol{\Lambda}\,\boldsymbol{\Phi}\,\boldsymbol{\Lambda}' + \boldsymbol{\Theta}. \end{aligned}$$

A faktorok becslései:

$$\widehat{\boldsymbol{\xi}} = \boldsymbol{\Phi}\,\boldsymbol{\Lambda}\,\boldsymbol{\Sigma}^{-1}\mathbf{x}, \tag{13.63}$$

vagy

$$\widehat{\boldsymbol{\xi}} = \boldsymbol{\Phi}(\mathbf{I} + \boldsymbol{\Gamma}\,\boldsymbol{\Phi})^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Theta}^{-1}\mathbf{x}.$$

Továbbá

$$\begin{aligned} E(\widehat{\boldsymbol{\xi}}\,\widehat{\boldsymbol{\xi}}') = E(\widehat{\boldsymbol{\xi}}\,\boldsymbol{\xi}') &= \boldsymbol{\Phi}\,\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\,\boldsymbol{\Phi} = \boldsymbol{\Phi} - \boldsymbol{\Phi}(\mathbf{I} + \boldsymbol{\Gamma}\,\boldsymbol{\Phi})^{-1} \\ E[(\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi})(\widehat{\boldsymbol{\xi}} - \widehat{\boldsymbol{\xi}})'] &= \boldsymbol{\Phi}(\mathbf{I} + \boldsymbol{\Gamma}\,\boldsymbol{\Phi})^{-1}, \end{aligned} \tag{13.64}$$

és

$$\begin{aligned} E(\widehat{\boldsymbol{\xi}} \mid \boldsymbol{\xi}) &= \boldsymbol{\Phi}(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})\boldsymbol{\xi} \\ &= \boldsymbol{\Phi}(\mathbf{I} + \boldsymbol{\Gamma}\,\boldsymbol{\Phi})^{-1}(\boldsymbol{\Lambda}'\boldsymbol{\Theta}^{-1}\boldsymbol{\Lambda})\boldsymbol{\xi} \\ &= (\mathbf{I} + \boldsymbol{\Phi}\,\boldsymbol{\Gamma}\,\boldsymbol{\xi} \\ &= \boldsymbol{\xi} - (\mathbf{I} + \boldsymbol{\Phi}\,\boldsymbol{\Gamma})^{-1}\boldsymbol{\xi}. \end{aligned} \tag{13.65}$$

Bartlett (1938) a faktorértékek becslésére egy másik eljárást javasolt, amely nem használja az eloszlásokra vonatkozó feltevést. Az eljárás a legkisebb négyzetek elvét alkalmazza, és a standardizált reziduálisok négyzetösszegét minimalizálja:

$$F = \sum_{i}^{m} (\epsilon_i^2/\theta_i) = \boldsymbol{\epsilon}'\boldsymbol{\Theta}^{-1}\boldsymbol{\epsilon} = (\mathbf{x} - \boldsymbol{\Lambda}\,\boldsymbol{\xi})'\boldsymbol{\Theta}^{-1}(\mathbf{x} - \boldsymbol{\Lambda}\,\boldsymbol{\xi}). \tag{13.66}$$

Az F függvényt minimalizáljuk $\boldsymbol{\xi}$ szerint, az első derivált:

$$-2\boldsymbol{\Lambda}'\boldsymbol{\Theta}^{-1}(\mathbf{x} - \boldsymbol{\Lambda}\,\boldsymbol{\xi}) = 2(\boldsymbol{\Gamma}\,\boldsymbol{\xi} - \boldsymbol{\Lambda}'\boldsymbol{\Theta}^{-1}\mathbf{x}),$$

amelyet egyenlővé teszünk $\mathbf{0}$-val, és $\boldsymbol{\xi}$ egyenletet kielégítő becslést $\widehat{\boldsymbol{\xi}}^*$-vel jelölve

$$\boldsymbol{\Gamma}\,\widehat{\boldsymbol{\xi}}^* = \boldsymbol{\Lambda}'\boldsymbol{\Theta}^{-1}\mathbf{x},$$

amiből

$$\widehat{\boldsymbol{\xi}}^* = \boldsymbol{\Gamma}^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Theta}^{-1}\mathbf{x}. \tag{13.67}$$

Ez a becslés nem függ attól, hogy a faktorok korrelálnak-e egymással vagy nem.

Ha a $\widehat{\boldsymbol{\xi}}^*$ becslést összehasonlítjuk a $\widehat{\boldsymbol{\xi}}$ becsléssel (először a faktorok korrelálatlanságát feltételezve, (13.67) és (13.59) egyenletek), akkor a különbség az, hogy $(\mathbf{I} + \boldsymbol{\Gamma})^{-1}$-t az (13.59) egyenletben kicseréltük $\boldsymbol{\Gamma}^{-1}$ -zel, vagyis a kétfajta becslés csak egy skálafaktorban különbözik.

A faktorok korreláltságát feltételezve az összefüggés a két becslés között:

$$\widehat{\boldsymbol{\xi}}^* = (\mathbf{I} + \boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}^{-1})\widehat{\boldsymbol{\xi}}, \tag{13.68}$$

vagy

$$\widehat{\boldsymbol{\xi}} = \boldsymbol{\Phi}\,\boldsymbol{\Gamma}(\mathbf{I} + \boldsymbol{\Phi}\,\boldsymbol{\Gamma})^{-1}\widehat{\boldsymbol{\xi}}^*.$$

A $\widehat{\boldsymbol{\xi}}^*$ kovarianciamátrixai

$$E(\widehat{\boldsymbol{\xi}}^*\,\widehat{\boldsymbol{\xi}}^{*'}) = \boldsymbol{\Phi} + \boldsymbol{\Gamma}^{-1}$$

$$E(\widehat{\boldsymbol{\xi}}^*\,\boldsymbol{\xi}') = \boldsymbol{\Phi}$$

$$E[(\widehat{\boldsymbol{\xi}}^* - \boldsymbol{\xi})(\widehat{\boldsymbol{\xi}}^* - \boldsymbol{\xi}')] = \boldsymbol{\Gamma}^{-1}.$$

A $\widehat{\boldsymbol{\xi}}^*$ becslés torzítatlan becslése a $\boldsymbol{\xi}$ elméleti értéknek:

$$\begin{aligned} E(\widehat{\boldsymbol{\xi}}^* \mid \boldsymbol{\xi}) &= E(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Theta}^{-1}\mathbf{x} \mid \boldsymbol{\xi}) \\ &= \boldsymbol{\Gamma}^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Theta}^{-1}\boldsymbol{\xi} = \boldsymbol{\xi}. \end{aligned} \tag{13.69}$$

Annak, hogy a $\widehat{\boldsymbol{\xi}} = \mathbf{A}'\mathbf{x}$ becslés torzítatlan legyen, szükséges feltétele, hogy

$$\mathbf{A}'\boldsymbol{\Lambda} = \boldsymbol{\Lambda}'\mathbf{A} = \mathbf{I}, \tag{13.70}$$

mivel

$$E(\widehat{\boldsymbol{\xi}} \mid \boldsymbol{\xi}) = \mathbf{A}'\boldsymbol{\Lambda}\,\boldsymbol{\xi}.$$

14. fejezet

MDS-modell

(MultiDimensionalScaling, Sokdimenziós skálázás)

A következőkben először a többdimenziós output gondolata köré csoportosuló skálázó eljárások néhány alapelvét és az egyes módszerek rövid összefoglaló leírását adjuk meg.

A sokdimenziós skálázás feltételezése, hogy létezik az objektumnak egy kvantitatív reprezentációja.

A skálázó modellekben az objektumok az állapottér pontjaiként jelennek meg olyan módon, hogy a hasonló objektumok kerülnek közel egymáshoz.

A nemmetrikus MDS-modell alapgondolata:

(i) Vizsgáljuk az objektumok halmazát, ennek elemeit a sokdimenziós tér egy-egy pontjával reprezentáljuk, s a köztük lévő kapcsolatokat a különbözőségeik empirikus mérésével fejezzük ki:

$$\delta(c_i, c_j) = \delta_{ij}.$$

(ii) a megoldás a pontok konfigurációja $[X]$ az r-dimenziós térben, ahol a pontok távolsága:

$$d(c_i, c_j) = d_{ij}.$$

(iii) a nemmetrikus MDS eljárásnak célja, hogy találjuk meg a pontoknak egy olyan halmazát (X) egy minimális dimenziószámú térben, hogy a különbözőségek monoton függvényei legyenek a távolságoknak:

$$\begin{aligned} &\text{ha} && \delta_{ij} < \delta_{k\ell}, \\ &\text{akkor} && d_{ij} \leq d_{k\ell}. \end{aligned}$$

Az iterativ eljárásban a normalizált reziduális négyzetösszeget (S) minimalizáljuk:

$$S = \left[\frac{\sum_{i<j} \left(d_{ij} - \widehat{d}_{ij}\right)^2}{\sum_{i<j} d_{ij}^2} \right]^{1/2},$$

ahol $\widehat{d}_{ij}$ a becsült távolságokat jelöli.

Az MDS-modelleket három kritérium szerint különböztethetjük meg:

– milyen adatmátrixot elemeznek,

– milyen a modell, amelyik meghatározza az adatok térbeli reprezentációjának pontos módját,

– milyen a függvény, amelyik az eredeti adatok és a kapott megoldás kapcsolatát méri.

MINISSA: (Michigan-Israel-Nijmegen-Integrated-Small-Space-Analysis)

Az induló adatokat egy $(n \times n)$ méretű input adatmátrix tartalmazza, melyben az n objektum δ_{ij} különbözőségei foglalnak helyet. Ha alapvető információink m változóval jellemzett n objektum alakjában állnak rendelkezésre, akkor a δ_{ij} különbözőségek lehetnek például az állapottér pontjainak euklideszi távolságai is.

Az eljárás célkitűzése: előállítani a n pont koordinátáit egy m-nél kisebb, r dimenziószámú térben úgy, hogy az output pontok közötti távolságok sorrendisége megegyezzen az input különbözőségek sorrendjével.

A vázolt gondolatmenet alapján lehetőség nyílik arra, hogy pl. egy 15 dimenziós térbeli pontfelhőt leképezzük a kétdimenziós síkba, az alapvető struktúra megőrzése mellett.

MINIRSA: (MINI Rectangular Smallest Space Analysis)

Legyen Q az objektumok, P pedig a személyek halmaza: A P halmazhoz tartozó n személy mindegyike a Q objektumhalmaz m elemét valamilyen közös tulajdonságuk alapján sorbarendezi. A modell lényege, hogy az objektumokat és a személyeket az r-dimenziós tér pontjaiként ábrázolja úgy, hogy a személyeknek az objektumoktól mért távolságai d_{ij} megfeleljenek a személyek preferencia-rendezéseinek.

A megfelelés jóságát egy hibafüggvénnyel mérjük, amit az eljárás során minimalizálunk. A MINIRSA eljárás során az adatmátrix soraihoz és oszlopaihoz (a személyekhez és objektumokhoz) szimultán jelölünk ki pontokat az r-dimenziós térben úgy, hogy a személyeknek az objektumoktól mért távolságai és a személyeknek az objektumokra vonatkozó rangsorai között monoton kapcsolat álljon fenn.

MRSCAL (MetRic SCALing)

Az eljárás abból a feltevésből indul ki, hogy a pontoknak létezik egy olyan minimális dimenziószámú tere, amelyben a pontok közötti távolságok (d_{ij}) az eredeti térben mért különbözőségeknek (δ_{ij}) lineáris vagy logaritmikus transzformációjával állíthatók elő. Az eljárás célja megtalálni az n pont (objektum) koordinátáinak (X) azt a becslését az adott dimenziószámú térben, amelyben a számított távolságok (d_{ij}) lineáris függvényei a mintatérben mért különbözőségeknek (δ_{ij}).

INDSCAL (INdividual Differences SCALing)

A modellben – más sokdimenziós eljárásokhoz hasonlóan – feltételezzük, hogy az adatok különbözősége a származtatott pontok közötti távolságok monoton függvénye. Az INDSCAL-módszere háromdimenziós adatmátrixból kiindulva két eredménymátrixot határoz meg: a csoport teret és az egyedi teret. Az INDSCAL-modell megadja a dimenziók relatív fontosságát (súlyát) minden egyed szemszögéből. Az egyedre vonatkozó dimenzió súlyok azt jelzik, hogy mennyire kell megnyújtani az egyes tengelyeket ahhoz, hogy az objektumok közötti távolságok maximálisan korreláljanak az objektumok közötti hasonlóságokkal.

PREFMAP (PREFerence MAPing)

Az eljárás egyedeknek vagy ezek kategóriákba sorolt megfigyelési egységeinek keresi az ideális pontjait egy *a priori* térben a tér pontjaira vonatkozó preferencia értékek alapján.

– Rendelkezünk az egyedeknek a jellemzőkre vonatkozó megfigyelési értékeivel, amelyek valamilyen preferencia skálán vannak értelmezve.

– Rendelkezünk az objektumoknak r-dimenziós konfigurációjával.

Az objektumok adott *a priori* terébe a PREFMAP-eljárás négy különböző modellel illeszti az egyedeket ill. a megfigyelési egységeket.

PARAMAP (PARAmetric MAPing)

Az eljárás az objektumok sokdimenziós skálázását végzi el, az objektumoknak m változóra vonatkozó megfigyelési értékei alapján. Az eljárás az objektumokat a megfigyelési térből átviszi az r-dimenziós származtatott térbe úgy, hogy a folytonosságot maximalizálja. A PARAMAP-eljárás feltételezi, hogy a mért változók kifejezhetők r számú latens változó nemlineáris függvényével. Így az objektumok az m-dimenziós térből átvihetők egy redukált r-dimenziós térbe. A PARAMAP-eljárás a nemlineáris faktorelemzés egy változatának tekinthető.

MDPREF (Multi Dimensional PREFerence Scaling)

Az eljárás preferencia rendezések elemzését végzi vagy páronkénti összehasonlítások adataiból, vagy a preferenciákat kifejező mérési eredményekből kiindulva. A Q halmazhoz tartozó objektumokra adott a P halmazhoz tartozó személyek preferencia rendezése. Az MDPREF-modell a személyeket és az objektumokat egy r-dimenziós közös térbe illeszti úgy, hogy az objektumok személyek vektoraira vetített értékeinek relatív nagyságai összhangban legyenek a személyek objektumokra vonatkozó preferenciáival.

Az MDPREF-eljárás preferencia illesztése hasonlít a PROFIT-eljárásra , azonban míg a PROFIT-módszernél az objektumok *a priori* terébe illesztjük a személyek preferenciáinak legjobban megfelelő vektort, az MDPREF-eljárásnál az objektumok terét szimultán, a személyek terével együtt határozzuk meg.

HICLUS (HIerarchial CLUStering)

Ez az eljárás az objektumok optimálisan homogén csoportjait kereső technika, amelyik a csoportok hierarchikus rendszerét állítja elő az adatok monoton transzformációjára invariáns algoritmussal. A HICLUS-eljárás az objektumok strukturáját az objektumok, illetve azok csoportjainak hierarchikus elrendezésével próbálja felismerni. A hierarchiában először minden egyes objektum különálló klaszterbe kerül. A hierarchiában a klaszterek optimálisan homogének, és a hierarchia független az adatok monoton transzformációjától.

UNICON (UNIdimensional CONjoint measurement)

Az eljárás maximum öt független változónak egy függő változóra vonatkozó összetett, közös hatását határozza meg additív, szubtraktív, vagy multiplikatív, illetve ezen műveletek kombinációjával felállított modell segítségével.

A jól ismert ANOVA-modellben feltételezzük, hogy a függő változó intervallummérési szintű, míg az UNICON ordinális mérési szintű változóra határozza meg a legjobb additív becslést.

PROFIT (PROperty FITting)

Ez az eljárás feltételezi, hogy rendelkezésünkre áll az objektumról az *a priori* konfiguráció, és a megfigyelési egységeknek az objektumokra vonatkozó mérési eredménye.

A PROFIT-eljárás feladata, hogy minden megfigyelési egységhez találjon egy vektort az r-dimenziós térben, amelyek maximálisan korrelálnak a megfigyelési egységek objektumokra vonatkozó mérési adataival. Másképpen keressük az objektumok r-dimenziós terében a megfigyelési egységeknek legjobban megfelelő vektorokat.

14.1. A sokdimenziós skálázás (MultiDimensional Scaling, MDS)

A sokdimenziós skálázás a matematikai statisztikai eljárások azon csoportjához tartozik, amelyek az adatok rejtett struktúráját vizsgálják. A sokdimenziós skálázás rövid múltra tekinthet vissza annak ellenére, hogy a klasszikus módszert Torgerson már 1958-ban kifejtette. Torgerson a színek érzékelését tanulmányozta ezzel a módszerrel. Személyeket kért meg, hogy válasszák ki melyik két szín a leghasonlóbb egy harmadikhoz. A sokdimenziós skálázást mégis egy másik példával szokták bevezetni. Vegyünk egy térképet, amelyik mondjuk valamelyik országot ábrázolja. Feladatunk, hogy a térkép alapján határozzuk meg a nagyobb városok közötti távolságokat. A feladat nagyon egyszerű, veszünk egy vonalzót, megmérjük a városok közötti távolságokat, az így kapott értékeket átszámítjuk tényleges távolságokká, és a feladatot ezzel megoldottuk.

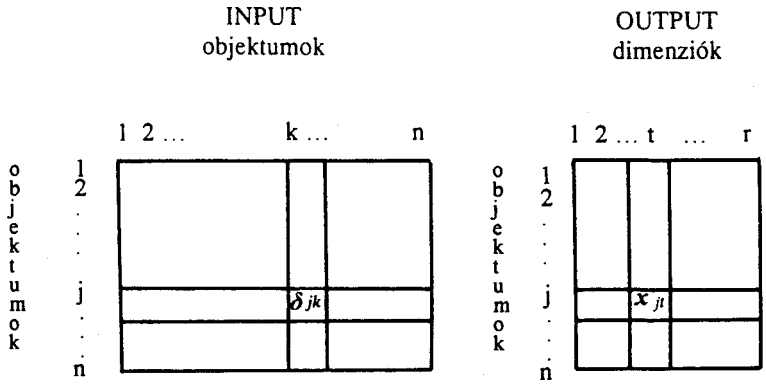
Képzeljük el, hogy a feladatunk éppen ennek a fordítottja. Megkapjuk a városok közötti távolságokat, és a térképet kell elkészítenünk. Ez sem megoldhatatlan feladat, de sokkal nehezebb mint az előző. A sokdimenziós skálázás tulajdonképpen ilyen fordított feladatok megoldására szolgáló módszer. Rendelkezésünkre áll bármilyen dologról, egyedekről, jellemzőkről, stimulusokról a köztük lévő hasonlóság vagy különbözőség. A sokdimenziós skálázás módszerének outputja egy olyan térbeli ábra, amely a térképhez hasonlóan tartalmazza a pontok geometriai alakzatát. Ez az alakzat az adatok belső struktúrájára vonatkozik, segíti az adatok kapcsolódási rendszerét feltárni. Másként fogalmazva, a sokdimenziós skálázás feladata, hogy a *minimális dimenziószámú térben* olyan ponthalmazt találjon, hogy a térbeli távolságok monoton függvényei legyenek az adatok közötti különbözőségeknek. Ha az i-edik és a j-edik egyed közötti különbözőség (δ_{ij}) kisebb, mint a k és l egyed közötti különbözőség, (δ_{kl}), akkor i és j közötti távolság legalább olyan kicsi legyen, mint k és ℓ közötti távolság ($d_{k\ell}$),

$$\text{vagyis ha } \delta_{ij} < \delta_{k\ell}, \quad \text{akkor } d_{ij} \leq d_{k\ell}.$$

Az ilyen ún. gyenge monotonitási kritériumot teljesítő módszereket nem metrikus módszernek nevezzük. Ezzel szemben, ha a különbözőségeket (δ) függvényszerűen feleltetjük meg a távolságokkal, akkor metrikus sokdimenziós skálázásról beszélünk, ekkor $d = f(\delta)$. Torgerson 1958-ban a $d = a + b\delta$ alakú függvénnyel számolt.

14.1.1. A sokdimenziós skálázás inputja és outputja

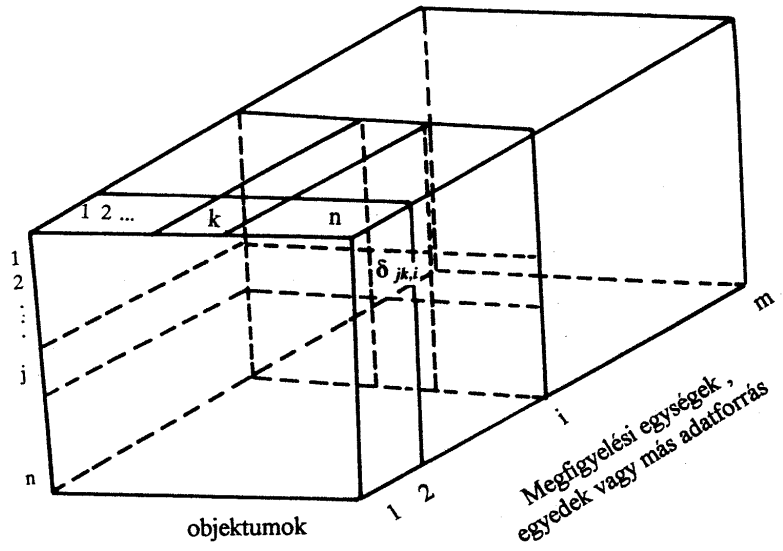
A sokdimenziós skálázás (MDS) a különbözőségeket kifejező adatok mátrixából indul ki. A különbözőségeket objektumok között mérjük, amelyek lehetnek stimulusok, jellemzők, változók. A különbözőségeket az objektumok megfigyelési egységekre (egyedekre) vonatkozó mérési értékei alapján számítjuk. A megfigyelési egység is jelenthet eltérő adatforrást, így lehetnek emberek (vizsgált személyek), különböző társadalmi csoportok, vagy szervezetek, időperiódusok stb. A kétutas (two-way) módszerek tipikus adatmátrixát (inputját) és eredménymátrixát (outputját) a következő ábra mutatja.



14.1. ábra. A kétutas MDS-módszerek inputja és outputja

A $\{\delta_{jk}\}$ az input különbözőségek $n \times n$-es, négyzetes, szimmetrikus mátrixa. A (δ_{jk}) a j-edik objektum vagy stimulus és a k-adik objektum vagy stimulus közötti különbözőséget méri. A szimmetrikusság miatt ($\delta_{jk} = \delta_{kj}$ minden j-re és k-ra) elég megadni a fődiagonális alatti vagy feletti elemeket. A kétutas MDS outputja egy $n \times r$-es mátrix, amely az n-objektum (stimulus) r-output dimenzióra vonatkozó koordinátáit tartalmazza. Az x_{jt} a j-edik stimulus t-edik dimenzióra (tengelyre) vonatkozó koordinátája.

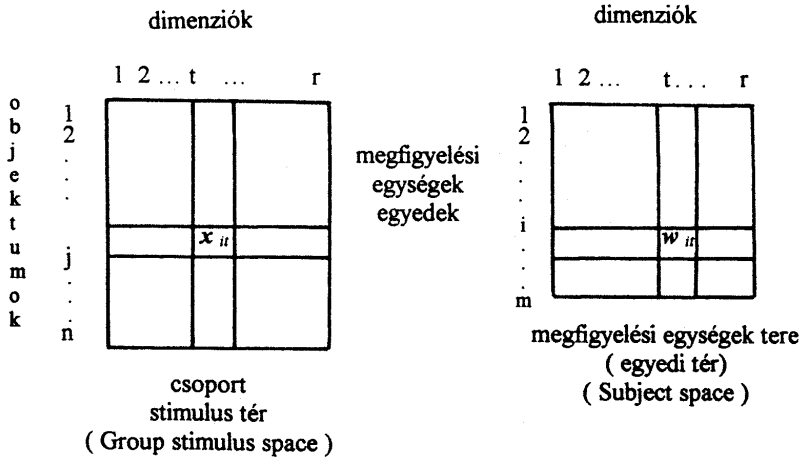
A társadalomtudományokban gyakran előfordul, hogy az n objektum különbözőségi mátrixa nemcsak a megfigyelési egységek (egyedek) egészére adott, hanem az egyedek klasztereire, csoportjaira, vagy legáltalánosabban minden megfigyelési egységre. Ilyen esetben az objektumoknak az eltéréseit kifejező mátrix nem kétutas, hanem háromutas mátrix, vagyis nem adattáblázatunk van, hanem adattömbünk. (A legáltalánosabban használt módszer, az INDSCAL nevét éppen az eltérésmátrix fenti kiterjesztése miatt kapta: INdividual Differences SCAL-ing). Az input adatait mutatja a következő ábra:



14.2. ábra. A háromutas MDS-módszerek inputja

A háromutas MDS-módszerek inputja $m(\geq 2)$ szimmetrikus különbözőségi mátrix. A $\delta_{jk,i}$ jelenti a j-edik és k-adik objektum (stimulus) közötti eltérés mértékét az i-edik egységre vonatkozóan. Az eltérés mérőszáma szimmetrikus, vagyis:

$$\delta_{jk,i} = \delta_{kj,i} \text{ minden } i, j, k\text{-ra.}$$



14.3. ábra. A háromutas (three-way) MDS-módszerek outputja

A háromutas MDS-módszerek két eredménymátrixot adnak. Az egyik az n objektum r dimenzióra (a csoport stimulus térre) vonatkozó koordinátáit tartalmazó $(n \times r)$-típusú mátrix, a másik mátrix m egyed súlyait tartalmazza az r dimenzióra vonatkozóan (egyedi tér). Mindkét mátrix grafikusan ábrázolható. Az egyedek „privát terét” megkaphatjuk,

ha a stimulusok (objektumok) tengelyeit (dimenzióit) az egyed súlyai négyzetgyökével megszorozzuk:

$$y_{jt,i} = \sqrt{w_{it}}\, x_{jt}.$$

14.1.2. A célfüggvény meghatározása

A sokdimenziós skálázás célfüggvényének meghatározásakor a sokdimenziós skálázás alapegyenletéből indulunk ki:

$$f(\delta_{ij}) = d_{ij}.$$

A célfüggvény természetes módon az adatok megfigyelt különbözőségeinek (δ_{ij}) valamilyen előírt függvénye ($f(\delta_{ij})$) és a származtatott térben mért távolságaik (d_{ij}) közötti eltérést kell hogy mérje. A statisztikában az ehhez hasonló eltérések mérésére elfogadott, hogy az eltéréseket négyzetre emeljük és összeadjuk. Ezt követően ajánlatos valamilyen standardizálást alkalmazni, vagyis az összeget elosztani egy skála-tényezővel.

Az így meghatározható célfüggvény általánosan elterjedt, és f-*hatásnak* hívják, S betűvel jelölik (stress of the configuration):

$$S = \sqrt{\frac{\sum_{ij}(d_{ij} - f(\delta_{ij}))^2}{\sum_{ij} d_{ij}^2}}.$$

Minél nagyobb S értéke, annál rosszabb a pontábra és az f-függvény illeszkedése az adatokhoz.

Természetesen az eltérések nagyságának mérésére más mértéket is lehet definiálni. A későbbiekben még kettőt bemutatunk.

A sokdimenziós skálázás feladatát most tehát úgy fogalmazhatjuk meg, hogy adott Δ adathalmazhoz keressük azt a pontábrát ($\widehat{X}$) , amely az optimális f-függvény esetén minimális hatás-értéket ad:

$$S(\Delta, \widehat{X}) = \min_{\substack{\forall x \\ \forall f}} S(\Delta, X, f).$$

14.1.3A sokdimenziós skálázás dimenzióinak értelmezése

A sokdimenziós skálázás pontábrájának értelmezését az ábrázolt pontok irányainak a vizsgálatával szokás végezni. Nyilvánvaló, hogy az alakzat helyzete kapcsolatban lehet a skálázott egyedek jellemzőivel. Ha a sokdimenziós skálázás terének minden irányát nem is, de a koordinátasíkok irányát könnyen tudjuk vizsgálni. A pontábra egyedeinek jellemzőit kapcsolatba hozhatjuk az egyes koordinátatengelyekkel. Erre módszert a korreláció- és regresszióelemzés ad. Az egyes tengelyek és a jellemzők páronkénti korrelációs együtthatója segíthet a tengelyek (dimenziók) elnevezésében hasonlóan ahhoz, ahogyan a faktorelemzésnél a faktorokat értelmezzük. A regresszióelemzést használhatjuk annak a hipotézisnek a tesztelésére, hogy a pontábra adott elhelyezkedésével egy változóhalmaz kapcsolatban van. A regresszióelemzés során keressük az ábra koordinátáinak olyan kombinációját, amely jól becsüli, jól magyarázza a változókat. Ennek jóságát a többszörös korreláció méri.

14.1.4. Az MDS-modellek alkalmazásai

A skálázás elméletét és módszereit bemutató első munkát számos könyv és tanulmány követte. Időrendi sorrendet tartva meg kell említenünk *Coombs* Adatelméletét (1964), amelyben megtalálható a nemmetrikus sokdimenziós skálázás kezdeti leírása mellett az egyéni preferenciákat „feltáró" modell is. Ez a könyv elsősorban az MDS-modellek elméleti és geometriai megalapozásával foglalkozik. 1970 és 1973 között három olyan tanulmány is megjelent, amelyek a módszereknek a piackutatásban való alkalmazhatóságát mutatják be. Természetesen jól használhatják az üzleti és pénzügyi élet más területén dolgozók is az algoritmusok összehasonlítására és a közülük való választásra.

Shepard,R omney és Nerlove kétkötetes munkája (1972) a sokdimenziós skálázás elméletének és alkalmazásainak átfogó bemutatását adja. Az első kötetben a skálázó modellek és módszerek csoportosítása után a nemmetrikus elemzés, az egyéni különbségek vizsgálata és még számos más elméleti kérdés szerepel. A második kötet az alkalmazási lehetőségek széles körét mutatja be. Találunk példát antropológiai, szemantikai vizsgálatokra, valamint a nemzetek közötti különbözőségek elemzésére is. Az említett könyveken kívül számos cikk jelent meg különböző folyóiratokban. A cikkek egy része a skálázás elméleti kérdéseit boncolgatja, más részük gyakorlati problémákat feszeget. *Shepardnak* (1962) két olyan írása jelent meg, amelyben a közelségek elemzését kísérli meg ismeretlen távolságfüggvény segítségével. *Kruskal* (1964) pedig az illeszkedés jóságát helyezi vizsgálatai középpontjába, és bevezeti a legkisebb négyzetes monoton regressziót a távolságok és a hasonlóságok közötti kapcsolat becslésére.

Guttman 1968-ban új számítási módszert dolgoz ki a legkisebb dimenziójú tér megtalálására.

Jelentős eredményt publikál *Carroll* és *Chang* 1970-ben az egyéni különbségek skálázása terén. A sokdimenziós skálázás általánosításaként bevezetik az INDSCAL-nak nevezett módszert, amelyet Eckart-Young általánosított eljárásával oldanak meg.

A módszertani és gyakorlati alkalmazási témákról írt cikkek is egymás után jelennek meg a társadalomtudományi, matematikai és statisztikai folyóiratokban.

Az egyéni különbözőségek skálázásával számos szerző foglalkozik. *Tucker* és *Messick* (1963) cikkében klasztereket elemez sokdimenziós skálázással.

McGee (1968) a nemmetrikus egyéni különbségeket vizsgálja. *Kruskal* és *Carroll* az illeszkedés jóságának különböző mértékeit határozza meg, és elméleti eredményeket közöl a hatás-függvényről.

Wish (1970) a nemzetek közötti hasonlóságot elemzi INDSCAL-modell segítségével.

Wish,D eutsch és *Biener* (1970) az INDSCAL alkalmazásával kimutatták, hogy a nemzetek közötti hasonlóság megítélésénél a dimenziók súlyozásának egyéni különbségei szisztematikusan függnek az egyén politikai beállítottságától, valamint attól, hogy mennyit tud az egyén az adott nemzetről, milyen országból származik és a neme is befolyásolja döntését.

Carroll (1971) az egyéni különbségeknek az észlelésben és a preferenciákban játszott szerepét tárgyalja.

Tucker (1972) olyan módszert dolgoz ki az egyéni különbségek elemzésére, amely az INDSCAL általánosítása, de hiányzik belőle az INDSCAL dimenzionális egyediség tulajdonsága.

Carroll és *Wish* (1973) a háromutas sokdimenziós skálázás modelljeit és módszereit bemutató cikkében részletesen bemutatja az IDIOSCAL-nak nevezett modellt, amelyben

az euklideszi távolság általánosítása szerepel, és speciális esetként magába foglalja az INDSCAL-t is.

A módszertani kérdések között olyanok szerepelnek, mint a távolságok súlyozása (*McGee*, 1966), metrikus, azaz számszerű információk származtatása nemmetrikus adatokból (*Shepard*, 1966) és az illeszkedés jóságának vizsgálata.

A gyakorlati alkalmazások között megtaláljuk a Morse jelekhez hasonló jelzések észlelését elemző modellt (*Wish*, 1967), valamint *Carroll* és *Chang* (1972) cikkét, amelyben az 1960-as amerikai elnökválasztást elemzik sokdimenziós skálázással.

Megjelentek olyan cikkek is, amelyek az alakzat pontosságát, a megfelelő dimenziók meghatározását, valamint a véletlen hibát vizsgálják pl. Monte Carlo módszerrel (*Klahr*, 1969., *Stenson* és *Knoll* 1971).

Young (1970) a nemmetrikus skálázással nyerhető metrikus információt határozza meg. A tényleges és a megállapított távolságok közötti korrelációs együtthatót a „metrikus meghatározottság mértékének" tekinti, és a kapott alakzat pontosságát becsli ennek segítségével.

14.2. A MINISSA-modell
(Michigan-Israel-Nijmegen-Integrated-Small-Space-Analysis)

A MINISSA-modellt J. E. Lingoes (University of Michigan) és Edward E. Roskam (University of Nijmegen) fejlesztette ki 1968-ban.

A MINISSA-modell kétutas adatmátrix elemzését végzi el hasonlósági vagy különbözőségi mértékek alapján. A MINISSA tekinthető a „smallest space" analízis alapprogramjának. A MINISSA-modellben adott n objektum (vagy stimulus) elemei közötti hasonlósági vagy különbözőségi mértékeket tartalmazó mátrix. Az algoritmus feladata megtalálni az n pont koordinátáit az r-dimenziós output térben úgy, hogy a pontok közötti távolságok sorrendisége megegyezzen a különbözőségek sorrendjével. A pontokat vagy stimulusokat jelölje j és $k = 1, \ldots, n$ index.

A pontok koordinátáit x_{jt} $(t = 1, \ldots, r)$ jelöli. Két pont közötti távolság (d_{jk}) a következő:

$$d_{jk} = \left\{ \sum_{t=1}^{r} | x_{jt} - x_{kt} |^p \right\}^{1/p} \tag{14.1}$$

– ha $p = 2$, akkor az ismert euklideszi távolságot kapjuk,

– ha $p = 1$, akkor d_{jk} az ún. city-block metrikával egyenlő.

A pontok különbözőségét jelölje δ_{jk} szimbólum. Ez lehet akár egyszerű rangszám, ami 1-től $(n(n-1)/2)$-ig futhat, vagy bármely más alkalmas mérőszám.

Az x_{jt} koordinátákkal adott pontok konfigurációjának nyomatékát (stress of the configuration) Lingoes és Roskam a következőképpen definiálta:

$$S = \sqrt{\frac{\sum_{jk}\{d_{jk} - f(\delta_{jk})\}^2}{\sum_{jk} d_{jk}^2}}, \tag{14.2}$$

ahol $f(\delta_{jk})$ a δ_{jk} különbözőségi mérték monoton függvénye, vagyis bármely $\delta_{jk} > \delta_{i\ell}$ esetén $f(\delta_{jk}) > f(\delta_{i\ell})$.

Guttman és Lingoes az r-dimenzióra vetített pontok közötti távolságok és a pontok különbözőségei között négyzetes eltérést mért (raw stress):

$$\phi_0 = S_0 = \sum_{jk}(d_{jk} - f(\delta_{jk}))^2. \tag{14.3}$$

A MINISSA-modellben az $f(\delta_{jk})$ függvényt kétféle módon határozzák meg:

a) a Kruskal-féle monoton regressziós eljárással $(\widehat{d}_{ij})$

b) a Guttman-féle rang-kép eljárással (d^*_{ij}).

A MINISSA-modell eljárása a jól választott kezdeti konfigurációból kiindulva keresi a pontoknak azt a konfigurációját, amely mellett S értéke minimális.

Az eredetileg Kruskal (1964) által definiált „nyomaték" (stress) alternatív megfogalmazása a Guttman-féle K együttható (coefficient of alienation). A Guttman-féle K együttható képlete:

$$K = \sqrt{1 - \frac{\{\sum_{jk} d_{jk} f(\delta_{jk})\}^2}{\sum_{jk} d^2_{jk} \sum_{jk}(f(\delta_{jk}))^2}}\,. \tag{14.4}$$

Bizonyítható (lásd Roskam (1969) és Lingoes és Roskam (1971, 1973)) az

a) $\sum_{jk} d_{jk}(d_{jk} - \widehat{d}_{jk}) = 0$ és

b) $\sum_{jk} d^2_{jk} = \sum_{jk}(d^*_{jk})^2$ azonosságok felhasználásával, hogy a Guttman-féle K együttható és az S együttható között a következő relációk érvényesek:

a) ha $f(\delta_{jk}) = \widehat{d}_{jk}$ akkor $K = S$

b) ha $f(\delta_{jk}) = d^*_{jk}$ akkor $K = S\left(1 - \left(\frac{1}{2}S\right)^2\right)$.

A fentiek alapján állítható, hogy azonos megoldásra jutunk, akár K, akár S minimalizálását végezzük el. Bár a megoldás szempontjából indifferens, S minimalizálásánál a Kruskal-féle monoton regressziós eljárást $(\widehat{d})$ alkalmazzák, K minimalizálásánál pedig d^*-ot, a Guttman-féle rang-kép eljárást.

14.2.1. Iteratív eljárás

A MINISSA-eljárás S vagy K minimalizálását iterációval végzi. Kiindul egy kezdeti konfigurációból, az x koordináták egy kezdeti halmazából (X), amiből a d távolságokat kiszámítja. Ezután megkeresi az $f(\delta)$ értékeket a monoton regressziós módszerrel vagy a rang-kép módszerrel. Ez az eljárás második fázisa, aminek a végén számítjuk S értékét. Ezután az eljárás olyan x koordinátákat keres, amelyik jobb illeszkedést ad $f(\delta)$ értékéhez. Az iteráció második lépésében az új konfigurációnak megfelelő távolságértékek alapján újraszámítjuk S értékét. Az eljárás egészen addig folytatódik, amíg vagy S értéke elég kicsi lesz, vagy stacionáriussá nem válik.

Az iteráció első lépésében a kezdeti konfigurációt a következőképpen határozzuk meg:

A kezdeti pontok a **C** mátrix főkomponensei, ahol a **C** mátrix elemei a következők:

$$c_{ij} = \begin{cases} 1 + \sum_k \rho_{ik}/\ell & i = j \\ 1 - \rho_{ij}/\ell & i \neq j \end{cases}$$

ahol

$$\ell = n(n-1)/2$$

ρ_{ij} a δ_{ij} különbözőségek rangszáma.

A kezdeti pontok meghatározásához csak az adatok sorrendiségét használjuk fel.

Az iteráció következő lépéseiben a „legmeredekebb lejtő" („the steepest descent") módszert használjuk, ahol a pontok koordinátáit a gradiens módszerrel változtatjuk.

$$x_{jt}^{(p+1)} = x_{jt}^{(p)} - \alpha_p \left\{ \frac{\partial S}{\partial x_{jt}} \right\}^{(p)}$$

ahol α_p optimálisan választott lépéshossz,

– $\frac{\partial S}{\partial x_{jt}}$ az S parciális deriváltja,

– p az iteráció ciklus-száma.

A MINISSA-program a legmeredekebb lejtő módszerének két változatát tartalmazza. Az egyiket „hard squeeze"-nek nevezik Guttman nyomán, és az S (stress of configuration) függvényt minimalizálja, a másik eljárás a „soft squeeze" , amelynek során az S_0 (raw stress) minimumát keressük. A gyakorlati számítások tanúsága szerint a soft squeeze eljárás egyszerűbb, gyorsabb, de még kielégítő eredményt ad a különbözőségi értékek (δ_{jk}) rang-képei $f(\delta_{jk})$ alapján, a hard squeeze eljárás jobb, bár némileg komplikáltabb, mivel a $f(\delta)$ monoton regressziós értékeit számítja.

14.2.2. A monoton regressziós $\widehat{d}$ érték számítása

A Kruskal-féle eljárás szerint keressük azon $\widehat{d}$ értékek halmazát, amely monoton függvénye az adatoknak (a δ különbözőségeknek) és minimalizálja a távolságoktól mért eltérés négyzetösszeget

$$\sum_{jk} (d_{jk} - \widehat{d}_{jk})^2.$$

Az algoritmus a következő:

Rendezzük a különbözőségeket monoton növekvő sorrendbe. Így az első (j, k) indexpárhoz a legkisebb különbözőségi érték tartozik, az utolsó indexpárhoz a legnagyobb különbözőség tartozik. E sorrendnek megfelelően rendezzük az iteráció során kapott konfiguráció távolságait (d_{jk}) . Ha az illeszkedés tökéletes, akkor a távolságok ebben a sorrendben már monoton növekedőek. Vegyük az első d értéket (amelyik a legkisebb különbözőségi indexhez tartozik) és amíg nem találunk nála nagyobb d értéket, helyettesítsük az első d-t és az őt követő nálánál nem nagyobb értékeket az átlagukkal. Ezután a „nagyobb értéktől" kezdjük a vizsgálódást, és keressük a következő nála

nagyobb d értéket és a közbeeső értékeket és a kiinduló értéket helyettesítjük az átlagukkal. Egészen addig folytatjuk az eljárást, amíg az átlagokra is igaz lesz, hogy monoton nemcsökkenő sorozatot alkotnak. Ezek az átlagos értékek adják a $\widehat{d}$ értékeket.

14.2.3. A Guttman-féle rang-kép eljárás d^* értékének számítása

A rang-kép eljárás a távolságok egyszerű permutációjával határozza meg a d^* értékeket.

Ez a következő rendezést jelenti.

Legyen δ_p a különbözőségek nagyság szerinti sorrendjében a p-edik elem (a legkisebb elem esetén $p = 1$). Rendezzük nagyság szerinti sorba a távolságokat (d). Legyen d_p a távolságok sorrendjében a p-edik elem. Ekkor d_p lesz a δ_p különbözőségi érték rangképe. Más szavakkal, ha δ_{jk} rang-száma p, és a d_{ip} távolság a távolság rangsorában a p-edik elem, akkor a d^*_{jk} egyenlő d_{ip} értékével. A d^* értékek általában nem minimalizálják a $\sum_{jk}(d_{jk} - d^*_{jk})^2$ kifejezést adott d_{ij} értékek esetén. Ezért a MINISSA-eljárásban a d^* értékeket csak az iteráció első lépésekor számítjuk, a további számításokat már a regressziós eljárással ($\widehat{d}$ meghatározásával) végezzük. A monoton regressziós $\widehat{d}$ értékének és a Guttman-féle rang-kép eljárás d^* értékének számítását illusztrálja a következő két táblázat (az adatok forrása: Inter-University Research Councils Series, Report No 32. May, 1977).

14.1/a. táblázat.

Különbözőségek δ	Távolságok d	Átlagolás			$\widehat{d}$	d^*
1	12	10	10	10	10	6
2	8	10	10	10	10	8
3	10	10	10	10	10	9
4	11	11	11	11	11	10
5	18	18	12	12	11,75	11
6	6	6	12	12	11,75	12
7	14	14	14	11,5	11,75	14
8	9	9	9	11,5	11,75	17
9	17	17	17	17	17	18
⋮						
stb.						

14.1/b. táblázat.

Hasonlóságok δ		Távolságok d	A d értékek rangszámai		$\hat{d}$		d^*	
csoportosított	nem csoportosított		csoportosított	nem csoportosított	első	második	első	második
1	3	2,0	3	7	4,25	4,83	5,0	6,67
1	1	8,0	1	1	8,0	4,83	8,0	6,67
1	2	3,0	2	6	4,25	4,83	7,0	6,67
2	6	4,0	2	4	4,0	4,83	3,0	3,33
2	5	5,0	1	3	4,25	4,83	3,0	3,33
2	4	7,0	1	2	4,25	4,83	3,0	3,33
3	8	2,0	3	8	2,0	2,0	2,0	1,67
3	9	1,0	3	9	1,0	2,0	1,0	1,67
3	7	3,0	2	5	3,0	2,0	2,0	1,67

14.1. táblázat. Az $f(\delta)$ értékek illesztésének bemutatása

14.2.4. Normalizálás

A MINISSA-eljárás a koordináták X mátrixát minden lépésben normalizálja:

a) a koordináták átlaga minden tengelyen nulla

$$\sum_j x_{jt} = 0 \quad (t = 1, \ldots, r)$$

b) a koordináták négyzetösszege n

$$\sum_{jt} x_{jt}^2 = n.$$

Az a) és b) feltételek teljesülése esetén euklideszi távolságot számítva a távolságok négyzetösszege n^2

$$\sum_{jk} d_{jk}^2 = \sum_{jkt}(x_{jt} - x_{kt})^2 = \sum_{jkt}(x_{jt}^2 - 2x_{jt}x_{kt} + x_{kt}^2) =$$

$$= n\sum_{jt} x_{jt}^2 - 2\sum_t(\sum_j x_{jt})(\sum_k x_{kt}) + n\sum_{kt} x_{kt}^2 = 2n^2$$

Mivel a számításokban figyelembe vettük d_{jk}-t és d_{kj} -t is, a négyzetösszeg egyenlő n^2-tel.

c) A tengelyeket ezután rotáljuk úgy, hogy

$$\sum_k x_{kt}x_{k\ell} = 0 \quad \forall t \neq \ell$$

Ez azt jelenti, hogy a koordináták a különböző dimenziókban korrelálatlanok. Nincs rotálás akkor, ha nem euklideszi távolságot használtunk. Az ilyen esetekben a tengelyek korreláltak lesznek.

14.2.5A nyomaték (stress) S értékének értelmezése

Nincs rigorózus szabály arra vonatkozóan, hogy a kiindulásul rendelkezésünkre álló és a MINISSA-eljárás eredményeként kapott konfiguráció eltérését mérő S nyomaték (stress of configuration) vagy a Guttman-féle K együttható (coefficient of alienation) amelynek értéke magasnak vagy alacsonynak nevezhető. A gyakorlati számítások alapján a következő kategóriákat alakítottuk ki.

Az illeszkedés jósága:

stress	S	$<$	0,05:	kiváló
0,05	$\leq S <$		0,10:	jó
0,10	$\leq S <$		0,15:	közepes
0,15	$\leq S <$		0,20:	elfogadható
0,20	$\leq S$		:	gyenge

A Guttman-féle K együttható (coefficient of alienation) értéke általában kb. 1,4-szer nagyobb, mint S értéke. A MINISSA-eljárás során K értékét számítjuk, amikor az $f(\delta)$-t a Guttman-féle rang-kép (d^*) eljárással határozzuk meg, és S értékét minimalizáljuk, ha $f(\delta)$-t monoton regresszió módszerével határozzuk meg. Ugyanazon adathalmaz esetén azt várjuk, hogy a dimenziók számának növelésével jobb illeszkedéshez, alacsonyabb S értékhez jutunk. Ez azonban nincs feltétlenül így. Ennek oka lehet az, hogy az egyes megoldásoknál nem kapunk konvergens eredményt a megengedett iteráció számán belül, vagy hogy lokális optimumot kapunk kisebb dimenziószámnál. Spence, I. és Graef, J. (1973) kidolgoztak egy eljárást a dimenzió számának a becslésére. Az M-SPACE eljárás azt a Monte Carlo adathalmazt keresi meg néhány adott dimenzióra, amely a legjobban illeszkedik a nemmetrikus sokdimenziós skálázás algoritmusa alapján számított S értékhez.

14.2.6. Példák a MINISSA-eljárásra (I. Országok fejlettség szerinti értékrendjének skálázása, II. A társadalom értékrendjének vizsgálata)

Országok fejlettség szerinti skálázása

Az országok közötti gazdasági fejlettségi szint összehasonlítására számos módszert alkalmaztak, illetve dolgoztak már ki. Ezek közül egyesek a GNP, GDP vagy nemzeti jövedelem nemzeti valutában kifejezett értékét számítják át valamilyen valutára (a valutaátváltási kulcsok alapján), míg mások a naturális gazdasági mutatók segítségével végezték az összehasonlítást. A gazdasági mutatószámok összegyűjtésénél az utóbbi időben a kibocsátás (flow) jellegű mutatók mellett majdnem ugyanolyan mértékben szerepeltetnek állomány (stock) jellegű mutatókat is. A gazdaság széles rétegeit átfogó mutatókból részben megpróbálnak szintetikus mutatót számítani, és az összehasonlítást e szerint elvégezni, részben az országok közötti összehasonlítást a struktúra alapján végzik, megtartva ezzel az országok szerkezeti sajátosságait, a gazdasági mutatók információtartalmát.

A következőkben bemutatásra kerülő példa az utóbbi megközelítéshez sorolható. A gazdasági szerkezetre vonatkozó mutatók mellé azonban társadalmi indikátorokat is bevettünk, mivel az országok fejlettségét önmagában gazdasági mutatók alapján megítélni nem tartjuk helyesnek. A kiválasztott mutatók a következők (a gazdasági mutatókat a Fejlettségi szintek, arányok, szerkezetek, Országos Tervhivatal Tervgazdasági Intézet

(1977) kiadványából, a társadalmi indikátorokat (Hudson és Taylor, World Handbook és Political and Social Indicators, Yale University Press) kiadványából vettük:

Ipari mutatók

1. Évi nyersacélfogyasztás kg/fő
2. Energiahordozó-fogyasztás szénegyenértékben kifejezve kg/fő
3. Évi termelői célú villamosenergia-fogyasztás kWó/fő
4. Nyerskőolaj évi fogyasztása kg/fő
5. Évi cementfogyasztás kg/fő
6. Évi alumíniumfogyasztás kg/fő
7. Évi cinkfogyasztás kg/fő
8. Évi ólomfogyasztás kg/fő
9. Évi rézfogyasztás kg/fő
10. Évi összes műanyagfajták, regenerált cellulóz és műgyanták fogyasztása kg/fő
11. Évi csomagolópapír-fogyasztás kg/fő

Mezőgazdasági mutatók:

12. Mezőgazdaságban használt, állományban lévő traktorok száma db/ezer fő mezőgazdasági dolgozó
13. Évi összes műtrágya-felhasználás hatóanyagban, kg/mezőgazdasági dolgozó
14. Egy mezőgazdasági dolgozó által eltartott lakosok száma
15. A mezőgazdaságban foglalkoztatottak aránya az összes foglalkoztatotthoz viszonyítva

Élelmiszerfogyasztási mutatók:

16. Évi gabonafogyasztás hántolt rizzsel együtt, lisztben kifejezve kg/fő
17. Napi állatifehérje-fogyasztás gr/fő
18. Napi tej- és tejtermék-fogyasztás kalória/fő
19. Évi cukorfogyasztás kg/fő
20. Évi kávé-, tea- és kakaófogyasztás kg/fő

Öltözködés (ruházkodás) színvonalát tükröző mutatók:

21. Évi hagyományosszál-fogyasztás (gyapot, gyapjú, selyem, len stb.) kg/fő
22. Mesterségesen előállított szintetikus és cellulóz alapú szálak évi fogyasztása kg/fő

Lakáshelyzetet és felszereltséget tükröző mutatók:

23. Egy lakószobára jutó személyek száma
24. Vízvezetékes lakások aránya az összes lakások százalékában
25. WC helyiséggel felszerelt lakások aránya az összes lakások százalékában
26. Évi háztartási villamosenergia-fogyasztás kWó/fő.

Közlekedési mutatók:

27. Diesel és elektromos mozdonyok összes mozdonyállományon belüli százalékos részaránya
28. Vasúton, közúton, belső víziúton, hajón és csővezetéken szállított összes árutonna km/fő
29. Ezer lakosra jutó autópályák és főútvonalak km-ben, km/ezer lakos
30. Ezer lakosra jutó regisztrált személygépkocsik száma
31. Regisztrált teherkocsik, autóbuszok és félvontatók (közúti vontatók) állománya, db/ezer fő

Egészségügyi mutatók:

32. Ezer élveszületett csecsemőre jutó egy éven aluli halottak száma
33. Összes egészségügyi dolgozó/lakos
34. Gyógyintézeti ágyak száma/ezer fő

Szolgáltatási mutatók:

35. A tercier szektorban foglalkoztatottak összes foglalkoztatottakhoz viszonyított aránya százalékban
36. Egy kereskedelmi dolgozóra jutó lakosok száma

Kulturális mutatók:

37. Háztartási és speciális vékonypapír-fogyasztás kg/fő
38. Nyomdai felhasználásra szánt papír (újságpapír kivételével) fogyasztása kg/fő
39. Évi újságpapír-fogyasztás kg/fő
40. Belső levélforgalom (küldött vagy kapott belföldi levelek) évi mennyisége db/fő
41. Telefonkészülékek állománya db/ezer fő
42. Rádió vevőkészülékek állománya db/ezer fő
43. Televíziókészülékek állománya db/ezer fő

Társadalmi ndikátorok:

44. Az oktatásra költött pénz a GNP százalékában
45. Sajtószabadság
46. Urbanizáció (100.000 vagy nagyobb népességszámú városokban lakó népesség aránya)
47. Felsőiskolába járó tanulók száma (millió lakos)
48. Jövedelemegyenlőtlenség, Gini index
49. Tudományos kapacitás (Hozzájárulás a világ tudományos eredményeihez)
50. Nemzetközi szervezetek tagságszáma
51. Diplomáciai képviseletek száma

A vizsgálatban szereplő országok

A vizsgálatba bevont országok számát és listáját a rendelkezésre álló naturális gazdasági mutatók határozták meg. Az országok a következők (a sorszámokat használjuk az ország jelzésére a stimulus térben).

1. Ausztria
2. Belgium-Luxemburg
3. Dánia
4. Egyesült Királyság
5. Finnország
6. Franciaország
7. Görögország
8. Hollandia
9. Irország
10. Olaszország
11. Norvégia
12. NSZK
13. Portugália
14. Spanyolország
15. Svájc
16. Svédország
17. Törökország
18. India
19. Izrael
20. Japán
21. Dél-Afrika
22. Egyiptom
23. Kanada
24. USA
25. Argentína
26. Brazília
27. Chile
28. Mexikó
29. Peru
30. Ausztrália
31. Új-Zéland
32. Bulgária
33. Csehszlovákia
34. Lengyelország
35. Magyarország
36. NDK
37. Románia
38. Jugoszlávia

Az országok gazdasági-társadalmi fejlettségét mérő 51 mutató alapján kiszámítottuk a vizsgált 38 ország között a páronkénti euklideszi távolságokat. Az 51 dimenziós gazdasági-társadalmi térben mért távolságmátrix adta a MINISSA eljárás inputját.

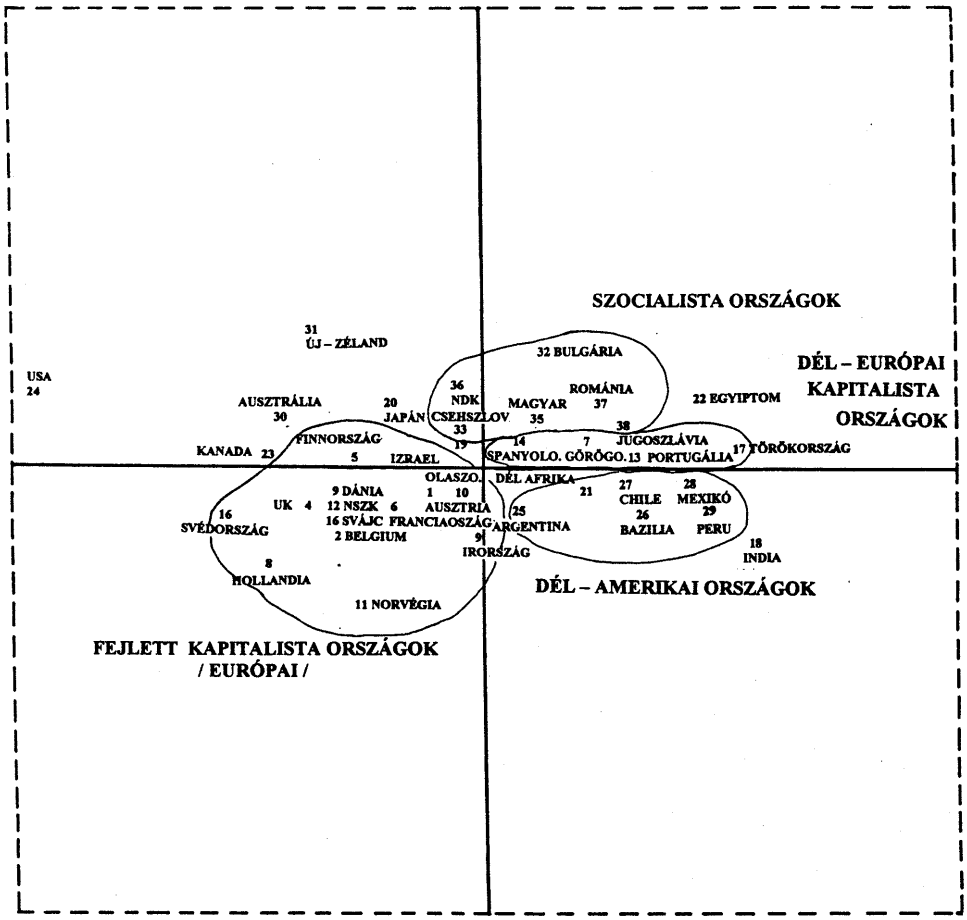
A két-, ill. egydimenziós megoldás koordinátáit a következő táblázat tartalmazza (a MINISSA-eljárás eredménye):

Országok sorszáma	Kétdimenziós megoldás koordinátái		Országok sorszáma	Egydimenziós koordináták
	1	2		1
1	−0,2465	−0,0472	1	−0,2044
2	−0,6847	−0,5213	2	−0,7440
3	−0,7223	−0,0379	3	−0,6790
4	−0,8995	−0,1762	4	−0,8319
5	−0,5551	0,1959	5	−0,4841
6	−0,4198	−0,1904	6	−0,4039
7	−0,5258	−0,0313	7	0,5356
8	−1,0755	−0,5407	8	−1,1818
9	0,0607	−0,3151	9	0,0864
10	0,0189	−0,1016	10	0,0095
11	−0,6643	−0,8207	11	−0,9172
12	−0,7268	−0,2013	12	−0,6612
13	0,7985	0,0050	13	0,8109
14	0,4515	−0,0290	14	0,4550
15	−0,6799	−0,3323	15	−0,6461
16	−1,3608	−0,1637	16	−1,4213
17	1,4717	−0,0295	17	1,5252
18	1,6241	−0,3271	18	1,7219
19	0,0468	−0,0273	19	0,0566
20	−0,3704	0,3871	20	−0,2974
21	0,6752	−0,0988	21	0,6879
22	1,3763	0,3382	22	1,4201
23	−1,1630	0,1821	23	−1,2173
24	−3,0548	0,2826	24	−3,3521
25	0,3731	−0,1970	25	0,3577
26	0,9154	−0,2571	26	0,9762
27	0,8242	−0,1109	27	0,8663
28	1,0706	−0,1224	28	1,0901
29	1,1870	−0,2664	29	1,2409
30	−1,0434	0,3048	30	−1,0797
31	−0,8407	0,7614	31	−1,0434
32	0,5319	0,7400	32	0,6430
33	0,0722	0,2381	33	0,1067
34	0,4323	0,2441	34	0,4405
35	0,4580	0,1720	35	0,4815
36	0,0243	0,5888	36	0,0244
37	0,7560	0,3406	37	0,7917
38	0,8129	0,1642	38	0,8364
Átlag	0,0000	−0,0000	Átlag	−0,0000
Szórás	0,9433	0,3321	Szórás	1,0000

14.2. táblázat. Országok gazdasági-társadalmi fejlettségének MINISSA-koordinátái

A MINISSA-eljárás eredményeként kapott ábra (14.4. számú ábra) azt mutatja, hogy a 38 ország gazdasági-társadalmi fejlettségét lényegében az első dimenzión lehet mérni. Kiugróan távol van az USA a többi országtól. A másik póluson India szerepel, mint legfejletlenebb ország. Megjegyezzük, hogy a kapott eredmények jó egyezést mutatnak Ehrlich Éva vizsgálati eredményeivel, valamint Bánkövi György által konstruált inverz módszerrel számított korrigált GDP/fő adatokkal.

A második tengely segítségével az országok társadalmi és földrajzi csoportjai különülnek el. Ezen a tengelyen azonban a szóródás lényegesen kisebb, ami két okból sem meglepő. *Egyrészt* az országok közötti eltéréseket a gazdasági és társadalmi mutatókból az euklideszi távolsággal mértük, és ez a távolság mérték egyenlő súllyal veszi az egyes változóknál mért eltéréseket, vagyis a változókat egymástól függetlennek tekinti, ami esetünkben nem teljesül. *Másrészt* a társadalmi indikátorok száma a gazdasági fejlettséget mutató változók számához képest aránytalanul alacsony. A bemutatott példa adathalmazát a további vizsgálatainkban kiegészítjük még más társadalmi indikátorokkal, és a MINISSA-eljárás inputjának az országok közötti Mahalanobis-féle távolságot számítjuk, amely a változók közötti korrelációval súlyozza az egyes koordináta eltéréseket.



14.4. ábra. Országok gazdasági-társadalmi fejlettségének MINISSA-pontábrája

A társadalom értékrendjének vizsgálata

A társadalom értékválasztásának struktúráját az 1978-as országos Életmód, Életminőség és Értékrendszer vizsgálat adatai alapján a klaszterelemzés, faktorelemzés és a sokdimenziós skálázás sokváltozós módszerekkel vizsgáltuk. (Hankiss Elemér, Manchin Róbert és Füstös László: Életminőség és Értékrendszer vizsgálata, MTA Szociológiai Kutatóintézet). Most a MINISSA-eljárással kapott eredményeket mutatjuk be. Az alapadatokat a következő kérdésre adott válaszok adták:

– Mik az élet legfőbb értékei, legfőbb javai?

– Mik a legfontosabb dolgok az életben?

– Mik azok, amelyek hiányát a leginkább megsínyli az ember?

A következő kártyákon egy-egy olyan érték látható, amelyeket az emberek általában fontosnak szoktak tartani. Kérem, válassza ki közülük azt a tízet, amelyet Ön a legfontosabbnak tart, a legtöbbre értékel; amelyek véleménye szerint a legfontosabbak ahhoz, hogy az ember boldog és megelégedett lehessen.

1. Barátok, társak, az emberek bizalma, szeretete
2. Elismerés, megbecsülés
3. Örömmel végzett munka, alkotómunka, élethivatás
4. Gyerekek, gyerekáldás
5. Tiszta lelkiismeret
6. Jó munkahelyi légkör
7. Családi boldogság, szeretteink egészsége, boldogulása
8. Szerelem
9. Egészség, erő, munkabírás, erős idegrendszer
10. Kényelmes, egészséges, jól felszerelt lakás, családi ház
11. Hatalom, vezető pozíció, társadalmi rang
12. Rend, nyugalom, béke
13. A művészi élmény, a zene, az irodalom, a szépség
14. A feszültségtől, konfliktustól, gondoktól mentes élet
15. Tanulási lehetőség, diploma, jó szakma
16. Függetlenség, szabadság, az, hogy az ember maga dönthessen sorsáról
17. Egy eszme, amiben hinni lehet
18. A haza, a szülőföld, a nemzeti hagyományok
19. Sport, mozgás, hobby
20. Létbiztonság, biztonság
21. Belső harmónia, lelki béke
22. Anyagi jólét, anyagi gondoktól mentes élet
23. Az, hogy az ember megtalálja a helyét a társadalomban
24. Társadalmi igazságosság, törvényesség, demokrácia, emberi jogok tiszteletben tartása
25. Lehetőség arra, hogy az ember komoly és kitartó munkával előbbre jusson, s vigye valamire
26. Sok szabadidő
27. Önbizalom, jó fellépés
28. Jó külső megjelenés, jó testalkat, szép arcvonások
29. Jó értelmi képességek, okosság, szaktudás
30. Az a tudat, hogy az ember hasznos és fontos munkát végez
31. A megfelelő származás, az, hogy az ember jó esélyekkel induljon az életbe
32. Az eredmény, a siker

A 32 alapérték között korrelációkat számítottunk, amelyek abszolút értékeit feleltettük meg az értékek páronkénti választásának hasonlósági mérőszámaként. Az így kapott hasonlósági mátrix volt a kiindulási mátrixa a MINISSA-eljárásnak, amelynek eredményét a következőkben bemutatjuk. A kétdimenziós térben a MINISSA-eljárás által kapott konfiguráció hibája (stress of the configuration) nagyobb, mint ami az irodalom ajánlása alapján elfogadható. Ez a példa azonban több nagyságrenddel bonyolultabb azoknál a feladatoknál, amelyek alapján a nyomaték (stress) értékére a kategóriákat meghatározták. Gondoljunk csak arra, hogy az 1462 dimenziós mintaterünkben mért 32 pontot vetítettük le kétdimenziós térbe úgy, hogy a hiba 0,243, vagyis 24%-os, és ez nem mondható semmiképpen sem rossznak. A biztonság kedvéért azonban a MINISSA-eljárást több változatban is lefuttattuk más hasonlósági mutatók alkalmazásával.

Az alkalmazott mutatók:

– korrelációs együttható átalakítása $(r + 1)/2$ hasonlósági mértékké
– kontingenciaegyüttható
– Yule-féle mutató
– Gamma-mutató

A korrelációs együttható abszolút értékéből számított eredményeket a fenti négy mutató alapján kapottak egyértelműen megerősítették, vagyis az emberi alapértékek belső struktúrája által kifeszített érték-tér nem függ a páronkénti hasonlóságokat mérő alkalmazott mutatóktól. Másik két sokváltozós statisztikai módszer, a klaszter- és faktorelemzés eredménye alapján is azt állíthatjuk, hogy a MINISSA-eljárás pontábrája összességében kitűnően tükrözi az emberi alapértékek belső struktúráját Magyarországon 1978-ban.

Az iterációk száma = 16
S_0 (row stress) $\widehat{d}$ esetén = 60,243
S (stress) $\widehat{d}$ esetén = 0,243

S_0 (raw stress) d^* esetén = 74,463
K (coefficient of alienation) d^* esetén = 0,267

A MINISSA ábrájába (a stimulus térbe) berajzoltuk a klaszterelemzéssel és a faktorelemzéssel kapott eredményeket is. A klaszterelemzést a MINISSA-eljáráshoz hasonlóan a korrelációs együtthatók abszolút értéke alapján végeztük a teljes lánc módszerrel. A kapott eredmények alapján az emberi alapértékek három nagy csoportját és azon belül mindegyikben még két alcsoportot, összesen 6 típust különíthetünk el.

Az alapértékek típusai:

A1 – haza, eszme, törvényesség, függetlenség
A2 – alkotó munka, hasznos munka, hely a társadalomban, emberek bizalma, jó munkahelyi légkör, elismerés
B1 – jó külső megjelenés, megfelelő származás, hatalom
B2 – szerelem, művészi élmény, sport, sok szabadidő, önbizalom, siker, tanulási lehetőség, jó értelmi képesség
C1 – egészség, gondoktól mentes élet, létbiztonság, anyagi jólét, lakás
C2 – családi boldogság, tiszta lelkiismeret, rend, béke, belső harmónia, gyerek

Az alapértékek fenti hat típusának megfeleltethető az alapértékekre végzett faktorelemzés első hat faktora is. A típusok a három sokváltozós módszer alapján határozottan elkülöníthetők, az értékstruktúrának hat lényeges elemét határozzák meg. Anélkül, hogy most részletesen kitérnénk rá, megjegyezzük, hogy az ábrából kiolvasható a klaszter-

Az alapértékek sorszámai	Dimenziók	
	1	2
1	−0,4539	−0,7900
2	0,1494	−0,8149
3	−0,6575	−0,6115
4	0,7818	−0,2547
5	0,7112	−0,7134
6	−0,6591	−1,1635
7	1,2424	−0,9683
8	−0,1382	0,1014
9	1,2748	−0,1629
10	1,3600	0,0909
11	−0,4574	1,4030
12	0,9941	−0,5724
13	−0,7765	0,6442
14	0,9309	0,6847
15	−0,2665	0,0646
16	−0,8799	−0,1018
17	−1,0970	0,1671
18	−1,0316	0,1402
19	−0,7656	0,4744
20	0,5599	0,0666
21	0,1804	0,1099
22	1,2474	0,5223
23	−0,8377	−0,8385
24	−1,0752	−0,4763
25	0,0721	−0,3810
26	−0,0699	0,9261
27	−0,1936	0,4105
28	0,2887	1,1404
29	−0,3014	0,0412
30	−0,1785	−0,9157
31	0,0550	1,2187
32	−0,0085	0,5588
Átlag	−0,0000	−0,0000
Szórás	0,7423	0,6701

14.3. táblázat. Az alapértékek terének koordinátái a MINISSA-eljárás kétdimenziós megoldásában

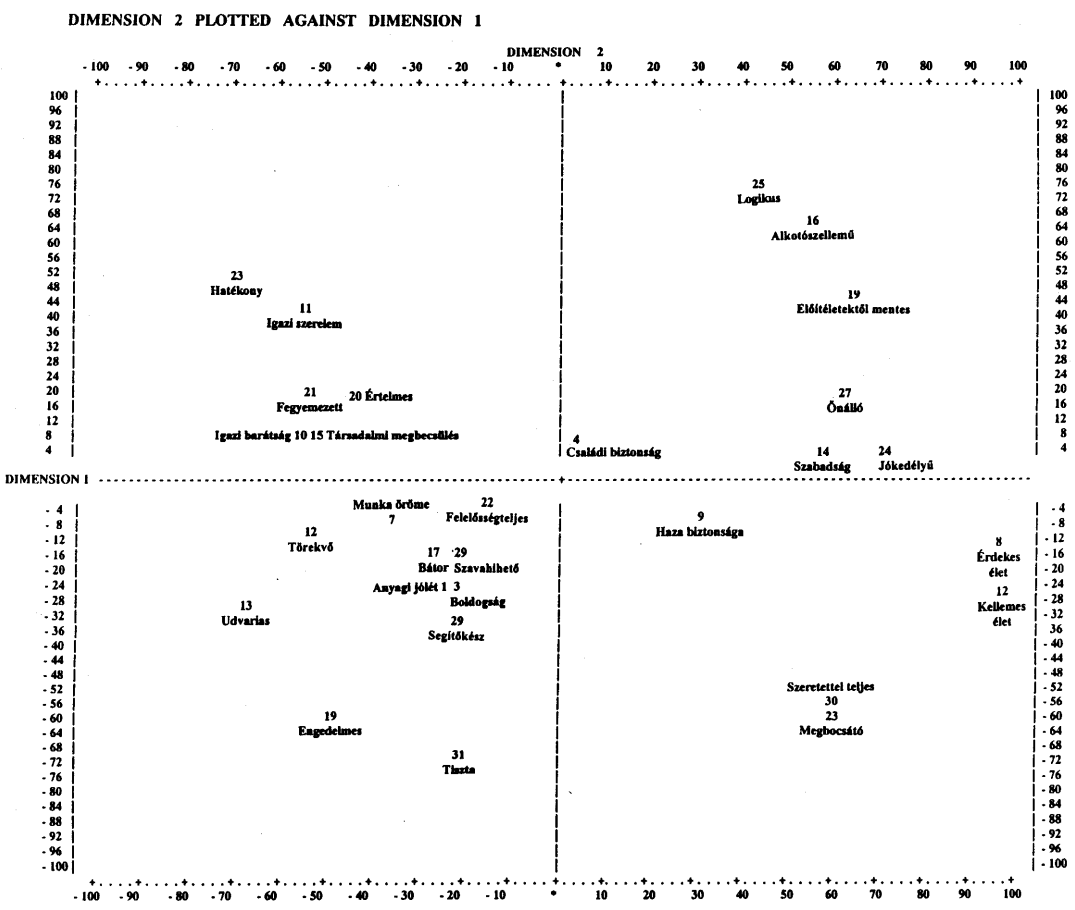
és faktorelemzés kapcsolata a MINISSA-eljárással. Az emberi alapértékek MINISSA-terének konfigurációját a vizsgálat más jellemzőinek alapján összefüggésbe hozhatjuk társadalmi változókkal. Ezt mutatja a következő, 14.4. táblázat.

A MINISSA első dimenzióján az emberi alapértékek különbözősége az anyagi jólét, kényelmes lakás értéktől a haza, eszme értékig terjed, és ezt a különbözőséget a korrelációk tanulsága szerint leginkább az idősebb, falun lakó, alacsony iskolai végzettségű betanított munkás, mezőgazdasági munkás, alacsony jövedelmű nő, társadalmi jellemzők magyarázzák. A másik dimenzió, amelyik a hatalom, siker értéktől a jó munkahelyi

légkör, hasznos munka tudata értékig terjed, inkább a nagyvárosban lakó, delelő korú, magas iskolai végzettségű, vezető, értelmiségi foglalkozású és magas jövedelmű jellemző magyarázza ellenkező előjellel.

	Tengelyek			
	1	2	3	R
Lakóhely				
Falu	0,42	−0,80	0,22	0,93
Kisváros	0,30	−0,83	0,23	0,91
Nagyváros	0,28	−0,83	0,13	0,88
Budapest	0,34	−0,84	0,03	0,90
Életkor				
20–29 évesek	0,24	−0,77	0,31	0,86
30–39 évesek	0,34	−0,80	0,25	0,90
40–49 évesek	0,32	−0,83	0,20	0,91
50–59 évesek	0,37	−0,82	0,11	0,91
60–69 évesek	0,51	−0,75	−0,04	0,91
Iskola				
4 osztálynál kevesebb	0,57	−0,71	0,17	0,93
5–8 osztály	0,50	−0,76	0,19	0,93
9–11 osztály	0,31	−0,81	0,23	0,90
12 osztály	0,23	−0,83	0,11	0,87
13–15 osztály	0,26	−0,83	0,16	0,88
16–17 osztály	−0,03	−0,83	0,20	0,85
18 osztály	−0,01	−0,89	0,04	0,89
Nem				
Férfi	0,26	−0,84	0,18	0,90
Nő	0,46	0,79	0,17	0,93
Foglalkozás				
0 Vezető	−0,01	−0,87	0,10	0,88
1 Felsőszintű szakember	0,04	−0,91	0,14	0,92
2 Egyéb szellemi	0,24	−0,86	0,21	0,92
3 Szakmunkás	0,22	−0,79	0,24	0,85
4 Betanított munkás	0,34	−0,77	0,30	0,90
5 Egyéb fizikai dolgozó	0,53	−0,67	0,20	0,88
6 Mezőgazdasági dolgozó	0,43	−0,76	0,23	0,90
7 Önálló iparos	0,42	−0,69	0,15	0,83
8 Egyéni gazdák	0,29	−0,83	0,09	0,89
9 Nyugdíjas, htb, GYES	0,51	0,77	0,02	0,93
Jövedelem				
1500 Ft vagy alatta	0,50	−0,74	0,16	0,91
1501-3000 Ft között	0,42	−0,79	0,15	0,91
3001-5000 Ft között	0,27	−0,84	0,23	0,91
5001 Ft felett	0,25	−0,85	0,19	0,91

14.4. táblázat. Az alapértékek MINISSA-koordinátái és a vizsgálat pár fő változója közötti korreláció



14.5. ábra. Értékek a társadalmi csoportok értékpreferencia-terében

14.3. Az MRSCAL-modell (MetRic SCALing)

Az MRSCAL a metrikus sokdimenziós skálázás (MDS) klasszikus eljárása. Az MRSCAL eredeti módszerét Richardson (1938) dolgozta ki, ezt módosította Young és Householder (1938). Az MRSCAL programot Torgerson (Theory and Methods of Scaling (1958)) alapján E. E. Roskam (1972) fejlesztette ki, mint a MINISSA-program metrikus párját.

Az MRSCAL-eljárás abból a feltevésből indul ki, hogy az n pontnak (stimulus, objektum...) létezik egy olyan minimális dimenziószámú tere, amelyben a pontok közötti távolságok (d_{ij}) az eredeti térben mért különbözőségeknek (δ_{ij}) lineáris vagy logaritmikus transzformációjával állíthatók elő:

$$d_{ij} = f(\delta_{ij}) \quad \text{vagy} \quad d_{ij} = f(\log(\delta_{ij})).$$

Az MRSCAL célja megtalálni az n pont (objektum) koordinátáinak (**x**) azt a becslését az adott dimenziószámú térben, amelyben a számított távolságok (d_{ij}) lineáris függvényei a mintatérben mért különbözőségeknek (δ_{ij}).

14.3.1A z additív konstans probléma

Az additív konstans problémát eredetileg Torgerson (1958) úgy fogalmazta meg, mint megtalálni azt a konstanst (c), amely megfigyelt viszonylagos távolságokat (különbözőségeket, δ_{ij}) alakítja át abszolút távolságokká (d_{ij}) olyan módon, hogy az eredményül kapott euklideszi tér dimenziószáma minimális legyen.

Ez a következőképpen fejezhető ki:

$$d_{ij} = \delta_{ij} + c \tag{14.5}$$

ahol

– δ_{ij} az i-edik és j-edik objektum (stimulusok) különbözősége a megfigyelési térben

– d_{ij} az i-edik és j-edik objektum távolsága a származtatott térben

– $d_{ij} = \left[\sum_{t}^{r} \left(x_{it} - x_{jt} \right)^2 \right]^{\frac{1}{2}}$

– c az additív konstans

– $i = 1, 2, \ldots, n$

– $j = 1, 2, \ldots, n$

– n az objektumok (stimulusok) száma.

Az additív konstans problémát Messick és Abelson (1956) elemezte először részletesebben. Vizsgálták a konstans hatását a sokdimenziós struktúrára. Eljárásukban az objektumok távolságainak négyzeteiből kiszámították az objektumoknak a tér tengelyeire eső vetületei (koordinátái) skaláris szorzatait, és azt a **B** mátrixba rendezték.

$\mathbf{B} = \{b_{ij}\}$ általános eleme:

$$b_{ij} = \frac{1}{2} \left[\frac{1}{n} \sum_{j}^{r} d_{ij}^2 + \frac{1}{n} \sum_{i}^{n} d_{ij}^2 - d_{ij}^2 - \frac{1}{n^2} \sum_{i}^{n} \sum_{j}^{n} d_{ij}^2 \right] . \tag{14.6}$$

A **B** mátrix diagonális elemei, az objektumok önmaguktól mért távolságai feltételezés szerint egyenlőek nullával:

$$d_{ii} = \delta_{ii} = 0. \tag{14.7}$$

A konstans c-t a különbözőséghez $i \neq j$ esetben adjuk hozzá. Ezért az (14.5) és (14.7) egyenleteket egy egyenletben a következőképpen írhatjuk:

$$d_{ij} = \delta_{ij} + c(1 - \varepsilon_{ij}) , \tag{14.8}$$

ahol

$$\varepsilon_{ij} = \begin{cases} 0 & \text{ha } i \neq j \\ 1 & \text{ha } i = j. \end{cases}$$

A (14.8) egyenlet behelyettesítve a (14.6) egyenletbe már olyan formulát kapunk, amelyből látszik a c konstans hatása a konfigurációra.

Először a távolságok négyzete ((14.5) négyzete) a (14.8) egyenlet alapján:

$$d_{ij}^2 = \delta_{ij}^2 + 2c\delta_{ij} + c^2(1 - \varepsilon_{ij}). \tag{14.9}$$

Ezt behelyettesítve (14.6)-ba kapjuk (a levezetést lásd Messich és Abelson (1956) (3. old.):

$$b_{ij} = \frac{1}{2}\left[\frac{1}{n}\sum_j^n \delta_{ij}^2 + \frac{1}{n}\sum_i^n \delta_{ij}^2 - \delta_{ij}^2 - \frac{1}{n^2}\sum_i^n\sum_j^n \delta_{ij}^2 + \right. \tag{14.10}$$

$$\left. + 2c\left(\frac{1}{n}\sum_j^n \delta_{ij} + \frac{1}{n}\sum_i^n \delta_{ij} - \delta_{ij} - \frac{1}{n^2}\sum_i\sum_j \delta_{ij}\right) + c^2\left(\varepsilon_{ij} - \frac{1}{n}\right)\right].$$

Mátrix jelölésekkel:

$$\mathbf{B} = \mathbf{A} + c\mathbf{E} + \frac{1}{2}c^2\mathbf{H} \tag{14.11}$$

ahol

$$a_{ij} = \frac{1}{2}\left(\frac{1}{n}\sum_j^n \delta_{ij}^2 + \frac{1}{n}\sum_i^n \delta_{ij}^2 - \delta_{ij}^2 - \frac{1}{n^2}\sum_i^n\sum_j^n \delta_{ij}^2\right)$$

$$e_{ij} = \left(\frac{1}{n}\sum_j \delta_{ij} + \frac{1}{n}\sum_i \delta_{ij} - \delta_{ij} - \frac{1}{n^2}\sum_i^n\sum_j^n \delta_{ij}\right)$$

$$h_{ij} = -\frac{1}{n} \quad (i \neq j) \text{ és } h_{ii} = \left(1 - \frac{1}{n}\right).$$

A $\mathbf{B}$ mátrix diagonális elemei:

$$b_{ii} = \frac{1}{n}\sum_j \delta_{ij}^2 - \frac{1}{2n^2}\sum_i\sum_j \delta_{ij}^2 + c\left(\frac{2}{n}\sum_j \delta_{ij} - \frac{1}{n^2}\sum_i\sum_j \delta_{ij}\right) + \tag{14.12}$$

$$+ \frac{1}{2}c^2\left(1 - \frac{1}{n}\right).$$

Ha a $\mathbf{B}$ mátrix legnagyobb sajátértéke β_1 és a hozzátartozó sajátvektor $\mathbf{x}_1$, akkor:

$$\mathbf{B}\,\mathbf{x}_1 = \beta\mathbf{x}_1. \tag{14.13}$$

Messich és Abelson a nagy sajátértékeket összeadta:

$$\sum_i^r \beta_i = \mathbf{x}_1'\,\mathbf{B}\,\mathbf{x}_1 + \ldots + \mathbf{x}_r'\,\mathbf{B}\,\mathbf{x}_r \tag{14.14}$$

és a (14.14)-be beírta a (14.11) egyenletet, majd egyszerűsítésekkel olyan egyenlethez jutott, amelyből már a c értékét kiszámíthatta:

$$\sum_i^r \beta_i = \sum_i^r \mathbf{x}_i'\,\mathbf{A}\,\mathbf{x}_i + c\sum_i^r \mathbf{x}_i'\,\mathbf{E}\,\mathbf{x}_i + \frac{1}{2}c^2\sum_i^r \mathbf{x}_i'\,\mathbf{H}\mathbf{x}_i. \tag{14.15}$$

A **H** mátrix speciális szerkezete miatt a (14.11) egyenlet a következőképpen írható:

$$\sum_{i}^{r} \beta_i = \sum_{i}^{r} \mathbf{x}_i' \mathbf{A}\, \mathbf{x}_i + c \sum_{i}^{r} \mathbf{x}_i' \mathbf{E}\, \mathbf{x}_i + \frac{1}{2} r c^2. \tag{14.16}$$

Messich és Abelson a (14.12) egyenlet megoldásával kapta meg a konstans értékét azzal a feltételezéssel, hogy a maradék $(n - r)$ sajátérték nulla.

Messich és Abelson megoldásának a problémáját az a feltételezés adja, hogy az „igazi" megoldásban a sajátértékeknek van egy minimális száma, amelyek pozitívak, a többi gyök pedig nulla. A gyakorlatban ez a feltételezés nem nagyon teljesül. Egy másik probléma az, ha nagy abszolút értékű negatív gyök merül fel. Ilyen esetben a kutató bajba kerül, mivel ezt nem tudja interpretálni, hacsak nem tételezi fel, hogy ez a hibahatás, és így nem értelmezi.

Cooper (1972) a fenti problémák miatt az additív konstans új megoldását kereste. Abból indult ki, hogy a metrikus skálázásnál van egy hibatag (e):

$$d_{ij} = \delta_{ij} + c + e_{ij}. \tag{14.17}$$

Cooper kereste azt a konstanst, amelyik a hibatag függvényét minimalizálja:

$$G = \frac{1}{2} \sum_{i}^{n} \sum_{j \neq i}^{n} e_{ij}^2 \longrightarrow \min. \tag{14.18}$$

A G függvény kifejtve:

$$G = \frac{1}{2} \sum_{i}^{n} \sum_{j \neq i}^{n} \left((\delta_{ij} + c)^2 + \sum_{k}^{r} (x_{ik} - x_{jk})^2 - \right. \tag{14.19}$$

$$\left. -2(\delta_{ij} + c) \left[\sum_{k}^{r} (x_{ik} - x_{jk})^2 \right]^{\frac{1}{2}} \right)$$

A G függvény c szerinti parciális deriváltja:

$$\frac{\partial G}{\partial c} = \sum_{i}^{n} \sum_{j}^{n} \delta_{ij} + n(n-1)c - \sum_{i}^{n} \sum_{j}^{n} \left[\sum_{k}^{r} (x_{ik} - x_{jk})^2 \right]^{\frac{1}{2}} = 0. \tag{14.20}$$

Ha a különbözőségeket úgy adjuk meg, hogy átlaguk nulla, a (14.20) egyenletből c értékét a következőképpen határozzuk meg:

$$c = \frac{1}{n(n-1)} \sum_{i}^{n} \sum_{j}^{n} \left[\sum_{k}^{r} (x_{ik} - x_{jk})^2 \right]^{\frac{1}{2}}. \tag{14.21}$$

A (G) hibafüggvénynek az objektumok vetületei x_{ik} szerinti parciális deriváltjai, ha figyelembe vesszük, hogy a távolságok invariánsak a tér origójának változtatására, a következő egyenlettel fejezhető ki (lásd részletesebben Cooper [1972]):

$$\frac{\partial G}{\partial x_{j^*k^*}} = \tag{14.22}$$

$$= 2 \left[n x_{j^*k^*} - \sum_{i \neq j^*}^{n} (\delta_{ij^*} + c)(x_{j^*k^*} - x_{ik^*}) \left(\sum_{k}^{r} (x_{j^*k^*} - x_{ik^*})^2 \right)^{-\frac{1}{2}} \right].$$

Cooper a megoldást a Fletcher–Powell iterációs eljárással végezte. (Megjegyezzük, hogy a Fletcher-Powell eljárás hatékonyságát a fenti sokváltozós függvény minimalizálására Gruvaeus és Jöreskog (1970) megvizsgálta, és azt hatékonynak találta.)

A Fletcher–Powell-eljárás felhasználásával egy programot is készítettek az additív konstans és az objektumok koordinátái megkeresésére COSCAL néven. A COSCAL-t FORTRAN-IV. nyelven írták IBM OS 360/91 számítógépre. A program lehetőségként megengedi kezdeti konfiguráció megadását. Különben az additív konstansnak olyan becslését keresi, amelyhez olyan abszolút távolságok tartoznak, hogy a legkisebb távolságok is nagyobbak nullánál. Az abszolút távolságokból a (14.6) egyenlettel számolt skaláris szorzat mátrix első r főkomponensét felelteti meg az r-dimenziós tér kezdeti konfigurációjának. Az illesztés jóságát a következő mutatóval méri:

$$FIT = 1 - \frac{\sum_{i<j}^{n} e_{ij}^2}{\sum_{i<j}^{n} (\delta_{ij} - \delta_i)^2},$$

(ahol δ_i a különbözőségek átlaga).

14.3.2A z MRSCAL-modell

Az MRSCAL-modellben keressük az n pontnak azt a konfigurációját az adott r dimenziószámú térben, amelyre a következő összefüggés igaz:

$$d_{ij} = f(\delta_{ij})$$

ahol

– δ_{ij} : az i és j pont között a megfigyelési (minta-) térben mért különbözőség

– d_{ij} : az r-dimenziós származtatott (redukált) térben mért távolság (Minkowszki metrika)

$$d_{ij} = \left\{ \sum_{k}^{r} | x_{ik} - x_{jk} |^p \right\}^{\frac{1}{p}}$$

– f: a δ_{ij}-k megengedett függvénye.

Az MRSCAL-modellben az „f” függvény lineáris ($d = a\delta + b$), de lehetőség van a különbözőségek logaritmikus transzformációjára ($d = a(\log \delta) + b$).

Az eljárás az illeszkedés jóságát mérő K (coefficient of alienation) együtthatót minimalizálja:

$$K = \sqrt{1 - \frac{\left\{ \sum_{ij} d_{ij} f(\delta_{ij}) \right\}^2}{\sum_{ij} d_{ij}^2 \sum_{ij} (f(\delta_{ij}))^2}}. \tag{14.23}$$

A monotonitási együtthatót a K együtthatóból a következőképpen számítja:

$$MU = \sqrt{1 - K^2}.$$

Az MRSCAL a K együttható minimalizálását iterációval végzi. *Az iteráció két fázisból áll.* Az iteráció *első fázisában* keressük a pontoknak azt a konfigurációját $(\mathbf{x})^{(s+1)}$, amelyben a d_{ij} távolságok legjobban illeszkednek a $f(\delta_{ij})^s$ értékhez, amelyet az előző iteráció második fázisban számítottunk. A pontok koordinátáit a „*legmeredekebb lejtő*”

(the steepest descent) módszerrel változtatjuk:

$$x_{jt}^{(s+1)} = x_{jt}^{(s)} - \alpha_s \left\{ \frac{\partial K}{\partial x_{jt}} \right\}^{(s)} . \tag{14.24}$$

(Az α_s az optimálisan választott lépéshossz.)

Az iteráció *második fázisában* keressük a δ_{ij} adatoknak azt a függvényét (f), amely a legjobban illeszkedik az első fázisban kapott $d_{ij}^{(s)}$ értékekhez. A második fázisban a legkisebb négyzetek módszere értelmében legjobban illeszkedő regressziós függvényt határozzuk meg. A kezdő lépésben megadhatunk egy kezdeti konfigurációt, vagy a programra bízhatjuk annak előállítását. Ekkor a kezdeti pontok a **C** mátrix főkomponensei. A **C** mátrix elemei:

$$\{c_{ij}\} = \Delta_{ij} \sum_k \frac{\delta_{ik}^0}{a} - \frac{\delta_{ij}^0}{a}$$

ahol

- $\Delta_{ij} = 1$, ha $i = j$ és 0 különben,
- δ_{ik}^0 a δ_{ik} bármely megengedhető transzformációjának maximuma,
- „a” pedig a következő függvénynek:

$$\sum_{ij} \left(\delta_{ij} - a\widehat{\delta}_{ij}^{(s-1)} - b \right)^2$$

a legkisebb négyzetek módszere értelmében legjobb regressziós becslése.

14.3.3. Példa az MRSCAL-eljárásra (A társadalom értékrendjének vizsgálata)

Milton Rokeach az amerikai nemzeti mintán (1409 felnőtt, húsz év feletti lakos) 36 eszköz- és célérték struktúráját vizsgálta 1967-68-ban. A 18 cél- és 18 eszközérték vizsgálatát az Életmód, Életminőség és Értékrendszer vizsgálat (Hankiss E., Manchin R., Füstös L. 1978) megismételte a magyar országos reprezentatív mintán (808 fő).

A kérdés így hangzott:

> „Kérem, rendezze ezeket az értékeket sorrendbe, aszerint, hogy mint irányelvek, milyen fontos szerepet játszanak az Ön életében. Tanulmányozza gondosan a kártyákat, azután válassza ki közülük a legfontosabbat. Tegyük itt le az asztalra, legfölül.
>
> Most válassza ki a sorrendben következő legfontosabbat. Tegyük a másik alá.
>
> És így tovább.
>
> Kérem, szánjon időt a munkára és mérlegelje minden egyes érték fontosságát. Mikor elkészült, lapozza végig még egyszer a kártyákat s ellenőrizze, hogy helyes-e a sorrend. S ha megvan, akkor engedje meg, hogy leírjam a sorrendet.”
>
> Ezután 18 cél- és 18 eszközértéket kellett külön-külön sorbarendezni.
>
> Mérési értékként tehát rangszámokat kaptunk.

Az értékek struktúráját Milton Rokeach faktorelemzéssel vizsgálta. A könyvében megtalálható korrelációs mátrix alapján, valamint a különböző társadalmi-demográfiai

kategóriák értékválasztásainak közölt mediánjai alapján elvégeztük adatai másodlagos feldolgozását, és ezzel párhuzamosan végeztük el ugyanezen elemzéseket a magyar mintán is. A kapott eredményeket összefoglalóan közöljük.

A 18 cél- és 18 eszközérték a következő:

1. Anyagi jólét
   (jómód, bőség)
2. Béke
   (háborútól és konfliktusoktól mentes világ)
3. Boldogság
   (megelégedettség)
4. Bölcsesség
   (életbölcsesség)
5. Családi biztonság
   (szeretteinkről való gondoskodás)
6. Belső harmónia
   (belső feszültségektől mentes élet)
7. Egyenlőség
   (testvériség, mindenki számára azonos lehetőség)
8. Az elvégzett munka öröme
   (teljesítmény, hasznosság)
9. Érdekes, változatos élet
   (élményekben gazdag, aktív élet)
10. A haza biztonsága
    (külső támadásokkal szembeni védettség)
11. Igazi barátság
    (szoros emberi kapcsolat)
12. Igazi szerelem
    (meghitt testi és lelki kapcsolat)
13. Kellemes, élvezetes élet
    (örömök, sok szabadidő)
14. Emberi önérzet
    (öntudat, önbecsülés)
15. Szabadság
    (függetlenség, választás lehetősége)
16. A szépség világa
    (a természet és a műalkotás szépsége)
17. Társadalmi megbecsülés
    (elismerés, tisztelet)
18. Üdvözülés
    (megváltás, örök élet)
19. Alkotó szellem
    (újító, eredeti gondolkodású)
20. Bátor, gerinces
    (kiáll a nézeteiért)
21. Előítéletektől mentes
    (elfogulatlan, nyílt gondolkodású)
22. Engedelmes
    (kötelességtudó, tisztelettudó)
23. Értelmes
    (gondolkodó, intelligens)

24. Fegyelmezett
(önuralommal rendelkező)
25. Felelősségteljes
(megbízható, felelősségtudó)
26. Hatékony
(hozzáértő, szakszerű)
27. Jó kedélyű
(vidám, könnyű szívű)
28. Logikus gondolkodású
(racionális, ésszerű)
29. Megbocsátó
(nem bosszúálló)
30. Önálló
(független, erős egyéniség)
31. Segítőkész
(mások jólétéért dolgozik)
32. Szavahihető
(becsületes, őszinte)
33. Szeretetteljes
(ragaszkodó, gyöngéd)
34. Tiszta
(rendes, ápolt)
35. Törekvő
(szorgalmas, vinni akarja valamire)
36. Udvarias
(jómodorú, jól nevelt)

A 18 cél- és 18 eszközérték terében 25 társadalmi csoportot vizsgáltunk.

A társadalmi csoportok a magyar mintában

Település

1. Falu
2. Kisváros
3. Nagyváros
4. Budapest

Életkor

5. 20–29 éves
6. 30–39 éves
7. 40–49 éves
8. 50–59 éves
9. 60–69 éves

Iskolai végzettség (években)

10. 4 év vagy kevesebb
11. 5–8 év
12. 9–11 év
13. 12 év
14. 13–15 év
15. 16–17 év
16. 18 év vagy több

Jövedelem (személyes jövedelem)

17. 1000 Ft vagy kevesebb
18. 1001–2000 Ft
19. 2001–2500 Ft
20. 2501–3000 Ft
21. 3001–4000 Ft
22. 4001–6000 Ft
23. 6001 Ft felett

Nem

24. férfi
25. nő

Társadalmi csoportok az amerikai mintában

Nem

1. férfi
2. nő

Jövedelem

3. 2000 $ vagy kevesebb
4. 2–3999 $
5. 4–5999 $
6. 6–7999 $
7. 8–9999 $
8. 10–14999 $
9. 15000 $ vagy több

Iskolai végzettség

10. 4 év vagy kevesebb
11. 5–8 év
12. 9–11 év
13. 12 év
14. 13–15 év
15. 16–17 év
16. 18 év vagy több

Életkor

17. 20–29 éves
18. 30–39 éves
19. 40–49 éves
20. 50–59 éves
21. 60–69 éves
22. 70 éves vagy idősebb

Az MRSCAL-módszer eredményeit négy táblázatban és négy ábrán mutatjuk be. Először a magyar minta, utána az amerikai minta megoldásait vesszük sorra. Az értékek páronkénti hasonlóságait a korrelációs mátrixból képeztük az $(r + 1)/2$ képlet szerint. A társadalmi csoportok eltéréseit az értékválasztásaik mediánjaiból számítottuk euklideszi távolság alapján (δ_{ij}):

$$\delta_{ij}^2 = \sum_{k=1}^{36} (y_{ik} - y_{jk})^2.$$

Az eredmények értelmezésével ezen a helyen részletesen nem foglalkozunk.

	1	2		1	2
1	−0,9131	−0,2820	19	1,0181	−0,6308
2	0,8194	0,8667	20	1,0002	0,3435
3	−1,0115	0,0066	21	1,0043	−0,1805
4	0,2208	−0,1240	22	−0,8662	0,7708
5	−0,5858	0,0994	23	0,7898	−0,7337
6	0,2107	−0,9124	24	0,2916	0,9247
7	0,5452	0,9287	25	0,9746	0,0613
8	0,1960	−0,5836	26	0,6188	−0,8448
9	−0,4678	−1,0339	27	−1,0385	−0,5411
10	0,8359	0,8935	28	0,9810	−0,6120
11	−0,2428	−0,8825	29	−1,1393	0,3197
12	−0,5757	−0,8447	30	1,0454	−0,3841
13	−0,9986	−0,5431	31	−0,5108	0,5711
14	0,5405	−0,1816	32	0,0369	0,7824
15	0,9406	0,7091	33	−1,1788	−0,1710
16	−0,3153	−0,5478	34	−0,9759	0,6589
17	0,6661	0,4732	35	−0,3849	0,6589
18	−0,5705	0,8752	36	−0,9603	0,6753

14.5. táblázat. Rokeach-értékek a magyar társadalomban (Lineáris modell) Multidimensional Scaling Solution, Coefficient of Monotonicity MU = 0,9609 Coefficient of Alienation = 0,2768

	1	2
1	−0,9627	−1,0604
2	−0,0059	0,0829
3	0,8786	−0,7327
4	0,7371	0,0311
5	0,4027	−0,9868
6	−0,0518	0,4777
7	−0,1150	0,0265
8	−0,3518	0,0556
9	−0,7169	0,5714
10	−1,1676	0,9574
11	−1,0290	0,0468
12	−0,1379	−0,0667
13	0,6138	−0,4184
14	0,8348	−0,3256
15	1,3289	0,9298
16	2,3358	0,0636
17	−1,3376	−0,0106
18	−1,1557	−0,2551
19	−0,8206	0,3213
20	−0,3757	−0,4223
21	0,0688	0,1204
22	0,4659	0,0198
23	0,8500	0,5814
24	0,2440	0,0473
25	−0,5322	−0,0544

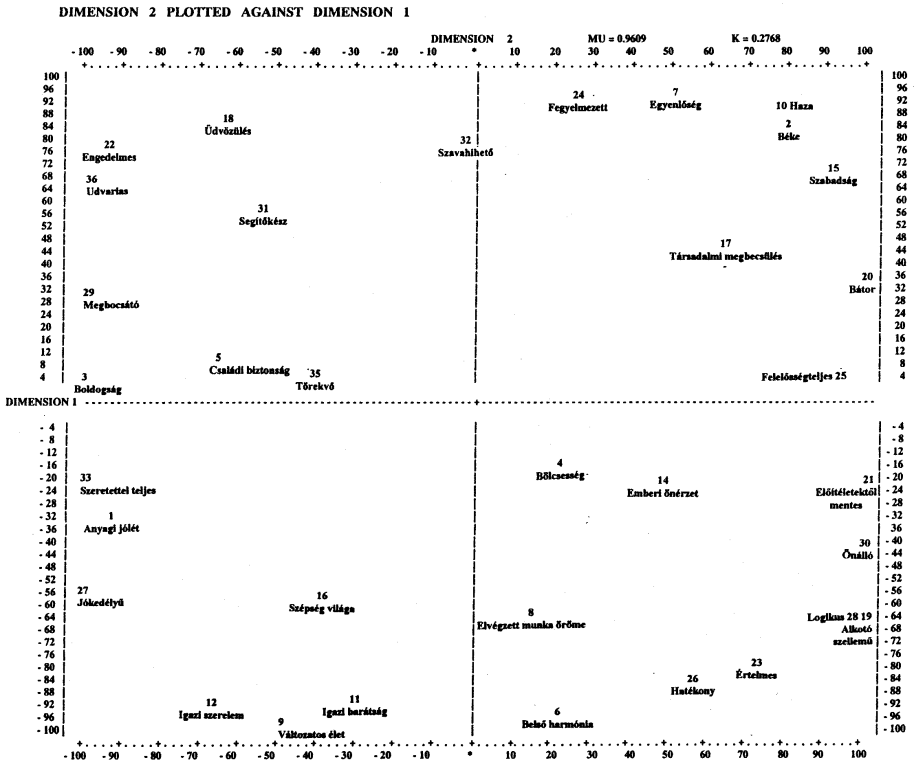
14.6. táblázat. Társadalmi csoportok a Rokeach-értékek terében Magyarországon (Lineáris modell) Multidimensional Scaling Solution, Coefficient of Monotonicity MU = 0,9888 Coefficient of Alienation = 0,1493

	1	2		1	2
1	0,2871	−1,1555	19	−0,3103	−1,0660
2	−0,5093	−1,0321	20	−0,7541	−0,5509
3	−1,1392	−0,0178	21	−0,7829	−0,7992
4	0,4276	0,9712	22	0,7989	−0,8259
5	−0,1980	−0,4039	23	0,9633	−0,6808
6	−0,0486	1,1385	24	−0,6487	−0,1286
7	0,2580	0,1989	25	1,0455	0,5695
8	−0,4594	0,6540	26	0,8716	0,5071
9	0,6370	−0,8507	27	0,6296	0,7088
10	−0,4341	0,9405	28	−1,0895	−0,3782
11	0,2599	−0,3973	29	−1,0376	−0,1680
12	0,1888	0,9774	30	−1,1119	0,3252
13	0,3544	−1,0847	31	−1,0163	0,4638
14	0,8614	0,8630	32	1,1047	−0,1860
15	−0,4649	0,2200	33	1,0486	0,1572
16	−0,1947	−0,8320	34	0,9393	−0,228
17	0,8147	−0,2747	35	−0,0464	0,4674
18	−0,8245	0,7747	36	−0,4198	0,9181

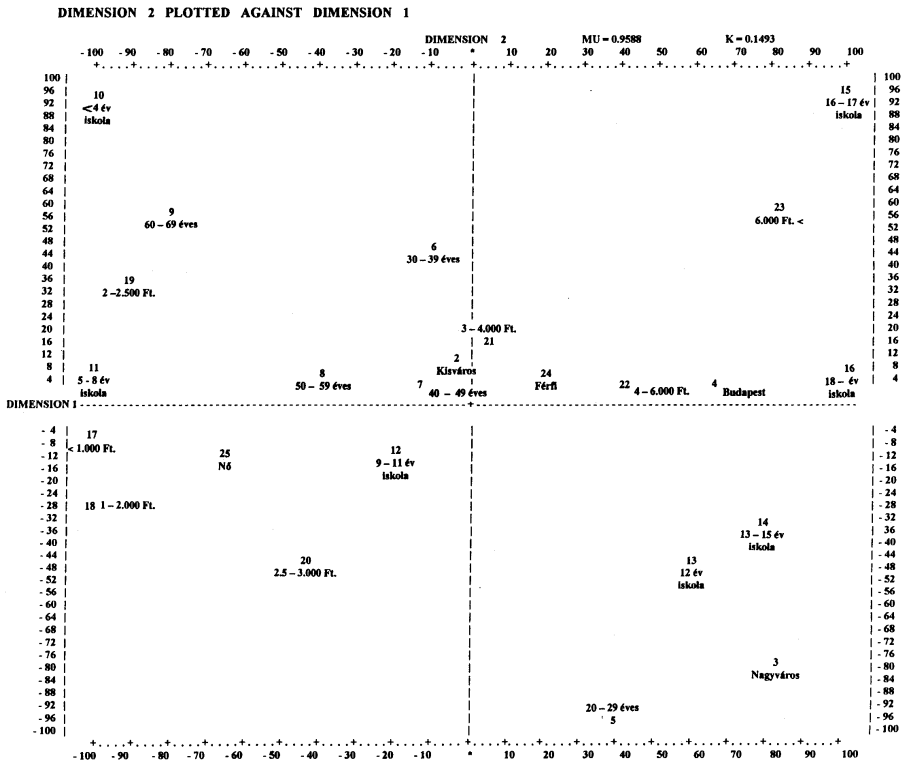
14.7. táblázat. Rokeach-értékek az amerikai társadalomban (Lineáris modell) Multidimensional Scaling Solution, Coefficient of Monotonicity MU = 0,9519 Coefficient of Alienation = 0,3064

	1	2
1	0,0848	−0,2448
2	−0,3194	0,2270
3	−1,0537	−0,5575
4	−0,6355	−0,3168
5	−0,2267	0,1267
6	−0,1476	0,3196
7	0,1023	0,3777
8	0,4794	0,0144
9	1,5411	−0,1896
10	−1,6931	0,5810
11	−0,9847	−0,2753
12	−0,4762	−0,0817
13	−0,0033	0,2511
14	0,7840	0,3431
15	1,6656	0,6164
16	2,0980	−0,9835
17	−0,0650	0,8579
18	0,4122	0,5080
19	−0,0918	0,0117
20	−0,3118	−0,1637
21	−0,4285	−0,5707
22	−0,7299	−0,8510

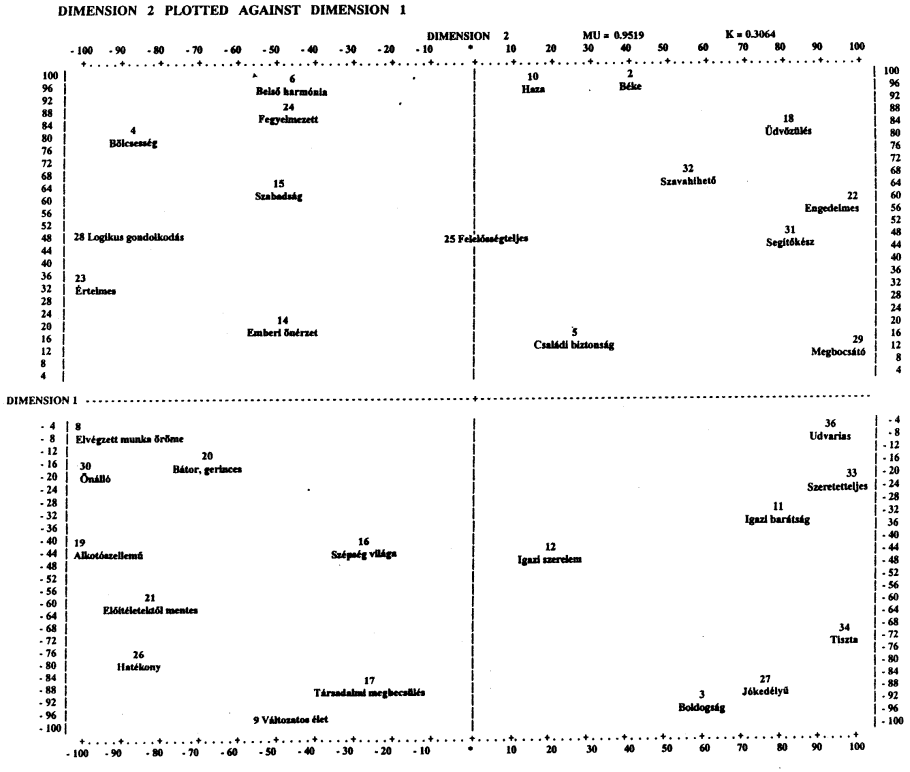
14.8. táblázat. Társadalmi csoportok a Rokeach-értékek terében Amerikában (Lineáris modell) Coefficient of Monotonicity MU = 0,9970 Coefficient of Alienation = 0,0768



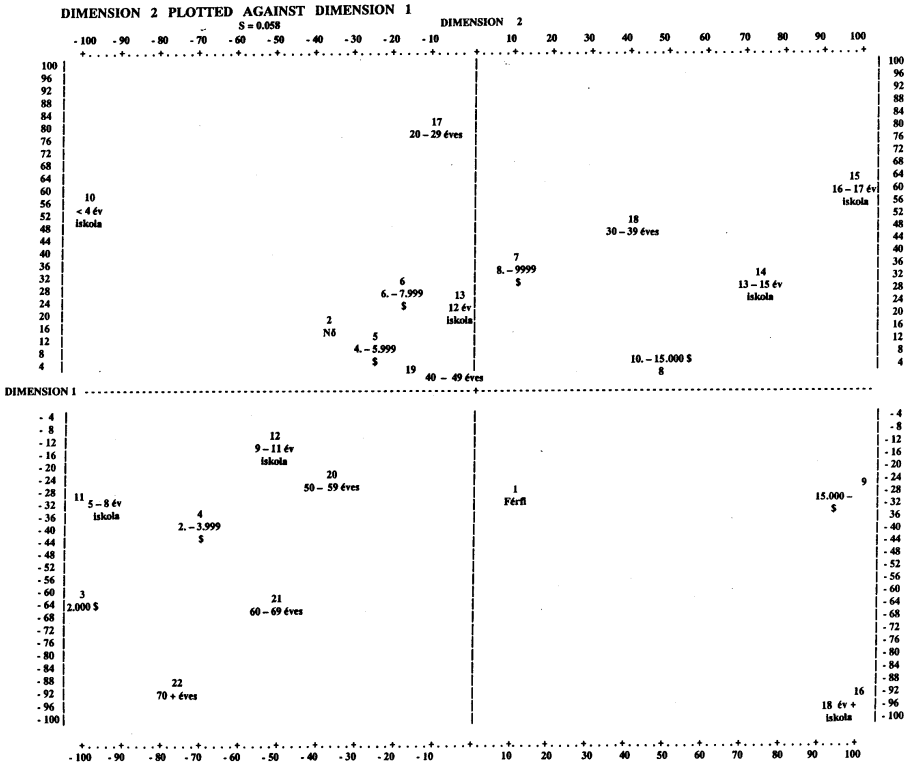
14.6. ábra. Rokeach-értékek a magyar társadalomban (Lineáris modell)



14.7. ábra. Társadalmi csoportok a Rokeach-értékek terében Magyarországon (Lineáris modell)



14.8. ábra. Rokeach-értékek az amerikai társadalomban (Lineáris modell)



14.9. ábra. Társadalmi csoportok a Rokeach-értékek terében az USA-ban (Lineáris modell)

14.4. A MINIRSA-modell (MINI-Rectangular [smallest] Space Analysis)

A MINIRSA-modell kétdimenziós (kétutas) adatmátrixok belső elemzését végzi el az euklideszi távolság modell szerint. Az eljárás monoton transzformációt használ, ezért nem metrikus módszer.

Legyen Q az objektumok, P pedig a személyek halmaza.

A P halmazhoz tartozó n személy mindegyike a Q objektumhalmaz m elemét valamilyen közös tulajdonságuk alapján sorbarendezi. Más szóval az n személyre rendelkezésünkre áll az m objektumra vonatkozó preferenciarendezés. A MINIRSA-modell lényege, hogy az objektumokat és a személyeket az r-dimenziós tér pontjaiként ábrázolja úgy, hogy a személyeknek az objektumoktól mért távolságai d_{ij} (ahol $i = 1, 2, \ldots, n$ és $j = 1, 2, \ldots, m$) megfeleljenek a személyek preferenciarendezéseinek. A megfelelés jóságát egy hibafüggvénnyel mérjük, amit az eljárás során minimalizálunk.

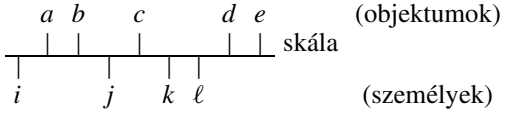
A MINIRSA-eljárás ugyanazon az elven épül fel, mint a MINISSA-módszer, kivéve, hogy itt a hibafüggvényt kicsit eltérő formában adjuk meg:

$$S^2 = \frac{1}{n}\sum_i \frac{\sum_j \left(d_{ij} - \widehat{d}_{ij}\right)^2}{\sum_j \left(d_{ij} - \overline{d}\right)^2}, \quad \text{ahol } \overline{d}_i = \frac{\sum_j d_{ij}}{m}.$$

Az eljárás ugyan a két módszernél hasonló, de ne tévesszük össze a két módszert: emlékeztetőül, a MINISSA-eljárás az objektumok között mért különbözőségek (vagy hasonlóságok) mátrixa alapján jelöli ki az objektumokhoz az r-dimenziós térben a pontokat úgy, hogy a közöttük mért távolságok monoton függvényei legyenek az objektumok között, a mintatérben mért távolságoknak. A MINIRSA-eljárás során a rangértékeket tartalmazó adatmátrix soraihoz és oszlopaihoz (a személyekhez és objektumokhoz) szimultán jelölünk ki pontokat az r-dimenziós térben úgy, hogy a személyeknek az objektumoktól mért távolságai és a személyeknek az objektumokra vontkozó rangsorai között legyen monoton függvénykapcsolat.

14.4.1. Rangsorolások sokdimenziós vetítése

Tételezzük fel, hogy m-számú objektumot valamilyen közös tulajdonság alapján összehasonlíthatunk, és hogy az objektumoknak a közös egydimenziós skálán léteznek skálaértékei. Ezen az egydimenziós skálán minden megfigyelési egységet (vizsgálati személyt, a továbbiakban egyszerűen személyt) egy ún. maximum preferencia ponttal reprezentálunk. Feltételezzük, hogy a személy az objektumokat annál jobban preferálja, minél közelebb vannak az őt reprezentáló maximum preferencia ponthoz. Ezért ezeket a pontokat ideális pontoknak nevezzük. A személyek preferencia-rangsorai így az objektumok és a személyek ideális pontjai között mért távolságok monoton függvényei. A következő ábra öt objektumot (a, b, c, d, e) és négy személyt (i, j, k, ℓ) ábrázol az egydimenziós skálán:

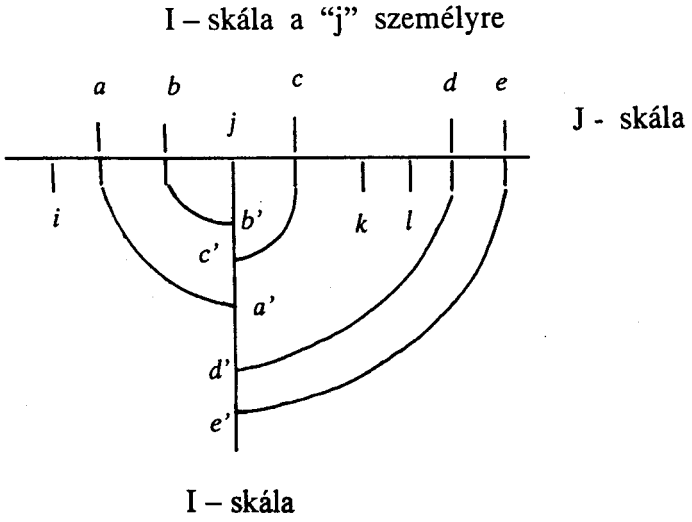


14.10.a. ábra. Az egydimenziós J-skála

Ezt az egydimenziós skálát J-skálának (joint scale) nevezzük. A J-skála alapján a személy preferencia-rangsorának megfelelő I-skálát vetítéssel kaphatjuk meg. A vetítést a j személyre a Coombs-féle „unfolding" technikával a következőképpen végezzük el:

Ha a J-skála alapján a példában szereplő másik három személy ideális pontjához hasonlóan elvégezzük a vetítést, a következő rangsorokat kapjuk:

Személy	preferencia-rangsor
i	$a\ b\ c\ d\ e$
j	$b\ c\ a\ d\ e$
k	$c\ d\ b\ e\ a$
ℓ	$d\ c\ e\ b\ a$



14.10.b. ábra. I-skála a „j"-személyre

Nyilvánvaló, hogy adott J-skála esetén nem kapunk meg minden lehetséges preferencia-sorrendet. Például, nem lehet olyan személy, amelynek preferencia-sorrendje *e a b c d*, ha az objektumok skála-értékei olyanok, hogy *b*, *c* és *d* skálaértékei az *a* és *e* értékei közé esnek (mint a fenti példában). Általában

$$p^i(j,k) \text{ jelölje a } \geq_e \text{ relációt adott } i\text{-re.}$$

A $p^i(j,k)$ értelmezése például: *j* gyakrabban következik be mint *k* egy adott *i* feltétel mellett, vagy *i* szituációban *j* reagálás (válasz) elfogadhatóbb, mint *k*.

Az $\geq_e$ szimbólumot empirikus relációnak nevezzük, (megkülönböztetve az aritmetikai $\geq$ relációtól), utalva arra, hogy itt a dolgok közötti, és nem a számok közötti relációról van szó. A $\geq_e$ szimbólum gyenge rendezési relációt fejez ki, amelyre fennállnak a következő tulajdonságok:

a) $p^i(j,j)$ (reflexivitás)
b) ha $p^i(j,k)$ és $p^i(k,j)$, akkor *j* ekvivalens *k*-val (antiszimmetria)
c) ha $p^i(j,k)$ és $p^i(k,h)$, akkor $p^i(j,h)$ (tranzitivitás)

Könnyű belátni, hogy az egydimenziós skála esetén nem lehetséges az összes elképzelhető rangsort a fenti vetítéses módszerrel előállítani. Ha *m* objektumunk van, akkor az *m* objektumot $m! = m(m-1)\ldots 1$ féleképpen tudjuk sorbarendezni. Ezzel szemben az egydimenziós skála esetén ennél lényegesen kevesebb rangsort tudunk előállítani:

$$\binom{m}{2} + 1 = \frac{1}{2}m(m-1) + 1.$$

Például, ha 5 objektumunk van, akkor a lehetséges 120 sorrendből csak 11-et kapunk meg.

Olyan esetben, amikor egy „lehetetlen" rangsort találunk, az egydimenziós vetítés helyett a sokdimenziós vetítést kell alkalmaznunk. Ebben az esetben az objektumokat és a személyeket (ideális pontokat) a többdimenziós térben reprezentáljuk. A preferenciarendezés itt is az ideális pont és az objektumok között mért távolságok rangsorolását

jelenti, de itt a pontok, amelyek között a távolságokat mérjük, a sokdimenziós térben vannak. Például a kétdimenziós térben 5 objektum lehetséges 120 sorrendjéből 46 sorrendet tudunk megadni. Látszik, hogy a dimenziószám emelésével az objektumok és személyek olyan konfigurációját kaphatjuk, amely alapján az összes lehetséges preferenciarendezést egyre nagyobb mértékben kaphatjuk vissza. Könnyen beláthatjuk, hogy m objektum esetén az $(m-1)$ dimenziós térben már minden lehetséges sorrend szóba jöhet (ezért az $m-1$ dimenziós megoldás triviális). Általában az előforduló és lehetséges sorrendek száma korlátozott, sőt bizonyos rendezések ki is zárhatják egymást.

A gyakorlatban sokszor nem is cél, hogy minden előforduló sorrendet reprodukáljunk a megoldásban, hanem inkább úgy fogalmazhatnánk, hogy a legjobb megoldásnak azt fogadjuk el, amelyik a lehető legtöbb rangsort reprodukál a lehetséges legkisebb dimenziószámú térben.

14.4.2. A MINIRSA-módszer

Az objektumokat jelölje $j, k, \ldots, m$, a személyeket (megfigyelési egységeket stb.) pedig $i, \ldots, n$.

Minden személynek az m objektumra van egy preferencia-rangsora. A $p^i(j,k)$ azt jelenti, hogy az i személy a j objektumot preferálja (előnybe helyezi) k-val szemben. Azt, hogy az i személy a j objektumot hogyan preferálja, a sokdimenziós térben a d_{ij} távolsággal mérjük. Legyenek a személyek ideális pontjainak koordinátái az $\mathbf{X}$ mátrixban, az objektumok konfigurációja pedig az $\mathbf{Y}$ mátrixban.

Az r-dimenziós térben az ideális pontok és objektumok közötti távolságot az általános távolság függvénnyel mérjük:

$$d_{ij} = \left\{ \sum_t \mid x_{it} - y_{jt} \mid^p \right\}^{1/p} \qquad t = 1, \ldots, r \tag{14.25}$$

ahol

– x_{it} az i-edik ideális pont (személy) t-edik dimenzióra vonatkozó koordinátája

– y_{jt} a j-edik objektum t-edik dimenzióra vonatkozó koordinátája

– d_{ij} az i és j pontok közötti távolság

– ha $p = 2$, akkor az ismert euklideszi távolságot kapjuk.

Valamely $\mathbf{X}, \mathbf{Y}$ konfigurációra kiszámítjuk a távolságokat. Most bevezetjük a $\widehat{d}_{ij}$ becsléseket a következő kritérium szerint

$$\widehat{d}_{ij} < d_{ij} \longleftrightarrow p^i(j,k). \tag{14.26}$$

A $\widehat{d}_{ij}$ értékeket úgy kell megválasztani, hogy kielégítsék a fenti kritériumot, és hogy a d_{ij} és $\widehat{d}_{ij}$ közötti különbség lehetőleg kicsi legyen. A becsült és tényleges távolságok közötti eltérést a következő függvénnyel mérjük:

$$\text{stress}_2 = S = \sqrt{\frac{1}{n} \sum_i \frac{\sum_j \left(d_{ij} - \widehat{d}_{ij}\right)^2}{\sum_j \left(d_{ij} - \overline{d}_i\right)^2}} \tag{14.27}$$

ahol

$$\overline{d}_i = \frac{\sum_j d_{ij}}{m}.$$

Az S értéke 0 és 1 közé esik. Az eljárás során a konfigurációt úgy próbáljuk változtatni, hogy az S értéke csökkenjen. Ha az S értéke nulla, vagy közelítőleg nulla, akkor a kapott konfiguráció megfelel vagy közelítőleg megfelel a következő kritériumnak:

$$d_{ij} \leq d_{ik} \longleftrightarrow p^i(j,k). \tag{14.28}$$

Más szóval minden személyre (i-re) a kapott konfigurációban a d_{ij} távolságok és a preferencia rangsorok p^i között (majdnem) teljes hasonlóság van.

A konfiguráció változtatását a „legmeredekebb lejtő" módszerével végezzük, ahol a pontok koordinátáit a gradiens módszerrel változtatjuk:

$$\begin{aligned} x_{it}^{\ell+1} &= x_{it}^{(\ell)} - \alpha_\ell g_{it} \\ y_{jt}^{\ell+1} &= y_{jt}^{(\ell)} - \alpha_\ell h_{jt}, \end{aligned} \tag{14.29}$$

ahol g_{it} és h_{jt} az x_{it} és y_{jt}-re vonatkozó gradiensek

– α_ℓ lépéshossz,

– ℓ az iteráció ciklusszáma.

A gradienseket az S függvény parciális deriváltjaival állítjuk elő.

$$\frac{\partial S}{\partial d} = \sum_{ij} \left[\frac{S_i \left(d_{ij} - \widehat{d}_{ij}\right)}{nSS_i^*} - \frac{S_i - \left(d_{ij} - \overline{d}_i\right)}{nST_i^*} \right] \tag{14.30}$$

ahol

$$\begin{aligned} S_i^* &= \sum_j \left(d_{ij} - \widehat{d}_{ij}\right)^2 \\ T_i^* &= \sum_j \left(d_{ij} - \overline{d}_i\right)^2 \\ S_i &= S_i^* / T_i^*. \end{aligned}$$

Ha a (14.30)-ban a szögletes zárójelben álló kifejezés Q_{ij}-vel jelöljük, a gradienseket a következő egyenletek adják:

$$g_{it} = \frac{\partial S}{\partial x_{it}} = \sum_j Q_{ij} \frac{|\, x_{it} - y_{jt} \,|^{p-1}}{d_{ij}^{p-1}} \operatorname{sign}(x_{it} - y_{jt}) \tag{14.31}$$

és

$$h_{jt} = \frac{\partial S}{\partial y_{jt}} = \sum_i Q_{ij} \frac{|\, x_{it} - y_{jt} \,|^{p-1}}{d_{ij}^{p-1}} \operatorname{sign}(x_{it} - y_{jt}). \tag{14.32}$$

A MINIRSA-eljárás számításának menete a fentiek alapján a következő:

1. Ha a felhasználó nem biztosít kezdeti konfigurációt, akkor a számítógépes program maga generál egy kezdeti konfigurációt, majd a személyek ideális pontjait elhelyezzük a két legpreferáltabb objektumaik közé.
2. Az objektumok és személyek konfigurációját normalizáljuk úgy, hogy az origót az összes pont centroidjába helyezzük, és hogy az összes koordináta négyzetösszege egyenlő legyen $(n + m)r$-rel.

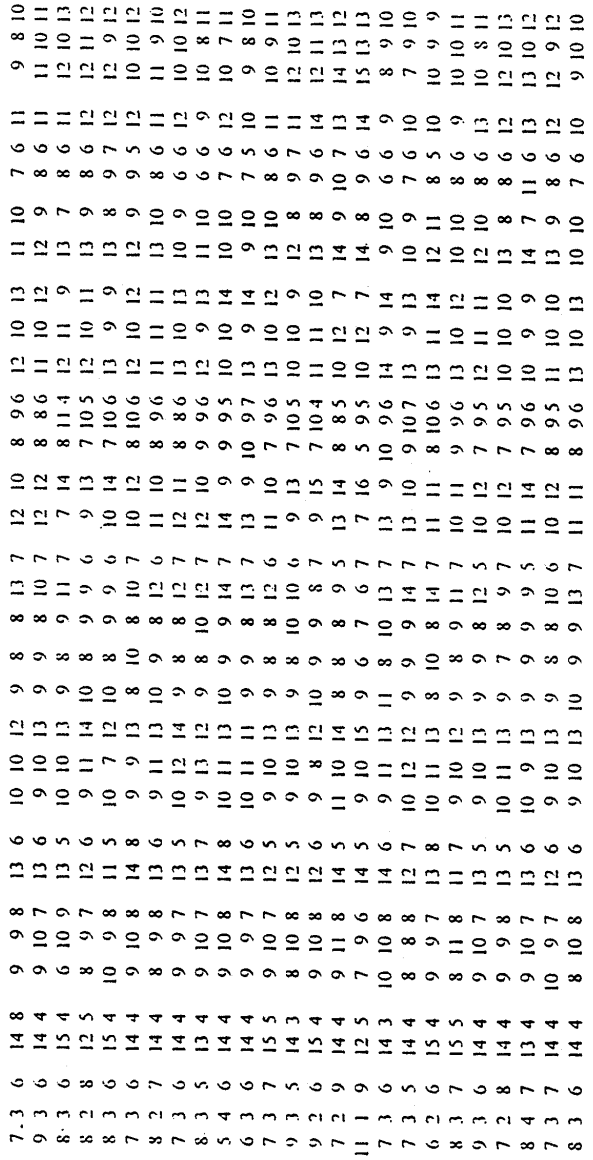
3. A d_{ij} távolságokat kiszámítjuk a (14.25) képlet szerint.
4. A $\widehat{d}_{ij}$ értékek illesztését Kruskal (1964) eljárásával végezzük. Ha a $\widehat{d}_{ij}$ nem elégíti ki a (14.28) kritériumot, a $\widehat{d}_{ij}$ értékeket a d_{ij} értékek átlagával tesszük egyenlővé.
5. Az S_i^*, T_i^* és S_i értékek és az S_2 (stress) számítása.
6. Ha az S_2 elérte a feltételezett minimumot, a számítás véget ért.
7. A gradienseket a (14.31) és (14.32) szerint számítjuk, majd a gradiensekkel a (14.29) szerint módosítjuk a koordinátákat.
   A lépéshosszt is Kruskal eljárása szerint számítjuk.
8. Az iterációt a 2. ponttól folytatjuk.
9. A végső megoldást rotáljuk úgy, hogy a koordináták tengelyei megegyezzenek a konfiguráció főkomponenseivel, (a konfiguráció sajátvektoraival). A dimenziók így a hozzájárulásuk mértékében rendezettek lesznek.

Megjegyzés: A „legmeredekebb lejtő" módszere nem biztosítja hogy az eljárás végén eljutunk a globális minimum értékéhez. Ezért a program lokális minimum esetén automatikusan egy teljesen különböző konfigurációból újraindítja a számítást. A megoldásként végül azt a konfigurációt fogadjuk el, amely mellett a célfüggvény értéke a legkisebb.

14.4.3. Példa a MINIRSA-eljárásra (Társadalmi csoportok az értékek terében)

A sokdimenziós skálázás különböző módszereinek ismertetésekor eddig is az Értékszociológiai Műhely Életmód, Életminőség és Értékrendszer vizsgálatának 1987-80-as évek között gyűjtött adataiból a Rokeach értéktesztet választottuk bemutató példának. Eddig a példák a Műhelyben folyó kutatómunka elemzéseiben valóságos modellek voltak. Most azonban a példa csak az eljárás bemutatása kedvéért készül. A MINIRSA-eljárás úgy látszik nem olyan robusztus, mint az előző eljárások, ezért a próbálkozások után még mindig nem jutottunk elemezhető eredményhez. Mivel most csak a módszer bemutatása volt a célunk, a következőkben közölt eredménytáblázatok, ill. ábrák körülbelül arra használhatók, hogy az adatok milyen változtatása (elhagyása) esetén juthatnánk elemezhető modellhez. De vegyük az elejéről. Kiindulásul 26 társadalmi csoportra álltak rendelkezésre a Rokeach teszt 36 értékének egész számra kerekített értékei, mediánjai (amiket a preferencia-sorrendekből számítottunk).

Ez a 26×36-os mátrix volt az első MINIRSA-modell inputja. Az értékek közül az Üdvözülés (18. sorszámú) értéket minden társadalmi csoportban nagyon hátra sorolták, ezért ez az érték a többi értéket a térben „összenyomta". Az eredmény így értelmezhetetlen volt. Ezt az értéket kihagytuk a következő futásokból. Ugyanezért és ugyanígy kellett tenni „A szépség világa" értékkel és a 70 évnél idősebbek társadalmi csoportjával. A közölt eredményekből pedig az olvasható ki, hogy a társadalmi csoportok a „Bölcsesség" értéket átlagosan hátrasorolják, és hogy ennek az értéknek a megítélésénél nincs társadalmi konszenzus, így ezt az értéket az adatmátrixból ki kell hagyni, és a számításokat újra el kell végezni. A Bölcsesség érték elhagyás előtti és utáni eredményeket is közöljük, hogy a módszer érzékenysége szemléletesen is áttekinthető legyen. (A Bölcsesség a 4. sorszámú érték volt, ezért az első ábrában az értékek sorszáma ezzel a csúsztatással érvényes.)



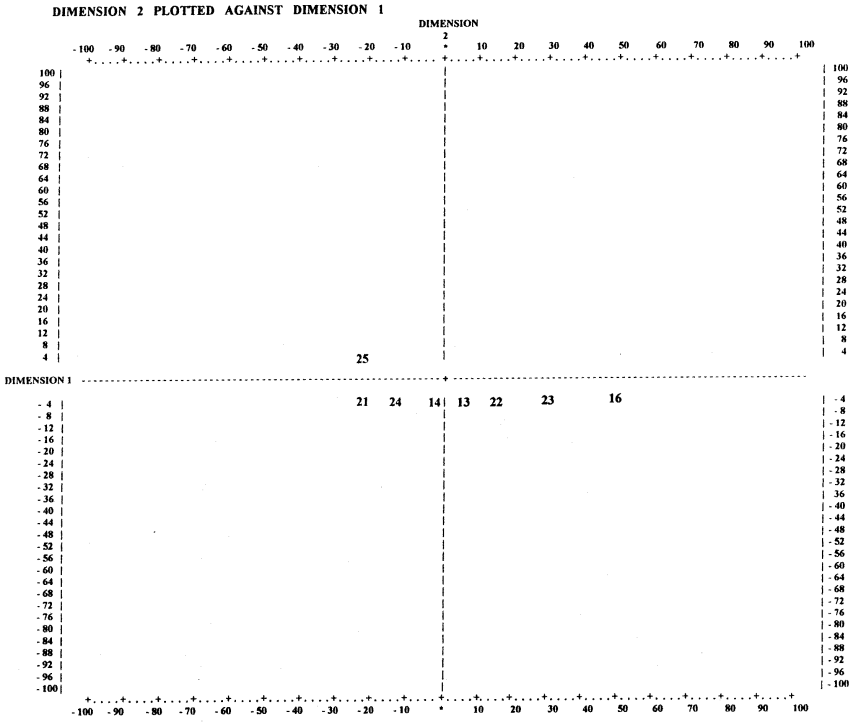
14.9.a. táblázat. MINIRSA input. Társadalmi csoportok az értékek terében, Rokeach-értékek kerekített értékei a 18. és 16. érték elhagyásával

	1	2		1	2
1.	−0,6195	−0,0470	1.	0,1933	−0,3148
2.	−0,6028	−0,0431	2.	0,2229	−0,0620
3.	−0,5684	−0,0052	3.	0,2178	0,0724
4.	1,3389	−0,1234	4.	4,0057	0,5541
5.	−0,5992	−0,0273	5.	0,2260	−0,0495
6.	−0,5964	−0,0486	6.	0,2157	0,2374
7.	−0,6136	−0,0492	7.	0,1793	−0,4168
8.	−0,6141	−0,0339	8.	0,2052	−0,2908
9.	−0,6159	−0,0307	9.	−1,2939	0,5655
10.	−0,6182	−0,0446	10.	0,2269	−0,0633
11.	−0,6220	−0,0421	11.	0,2106	0,2602
12.	−0,6116	−0,0371	12.	0,1761	−0,4282
13.	0,3975	−0,0462	13.	−1,3378	0,5121
14.	0,0054	−0,0535	14.	0,1881	−0,4011
15.	−0,5908	−0,0409	15.	−0,2231	−1,5453
16.	2,1330	−0,1372	16.	0,1810	−0,3888
17.	−0,6235	−0,0282	17.	0,2293	0,2654
18.	−0,6212	−0,0361	18.	0,2223	−0,1550
19.	−0,6204	−0,0495	19.	0,2219	0,2848
20.	−0,6143	−0,0267	20.	−1,3872	−0,4798
21.	−0,6053	−0,0479	21.	0,2264	−0,2426
22.	0,8002	−0,1031	22.	0,1771	−0,4119
23.	1,3391	−0,1475	23.	0,2243	−0,0869
24.	−0,2098	−0,0667	24.	0,1790	−0,4776
25.	−0,6130	−0,0328	25.	0,2080	0,2917
			26.	0,2215	0,2989
			27.	−1,3559	0,4636
			28.	0,2263	0,2082
			29.	0,2010	0,2283
			30.	0,2221	−0,1118
			31.	−1,2907	0,5445
			32.	−1,3761	0,3909
			33.	0,1644	−0,4338
			34.	−1,3599	−0,5255
ÁTLAG:	−0,2066	−0,0539	ÁTLAG:	0,0000	0,0000
SZÓRÁS:	0,7642	0,0348	SZÓRÁS:	0,9364	0,3510

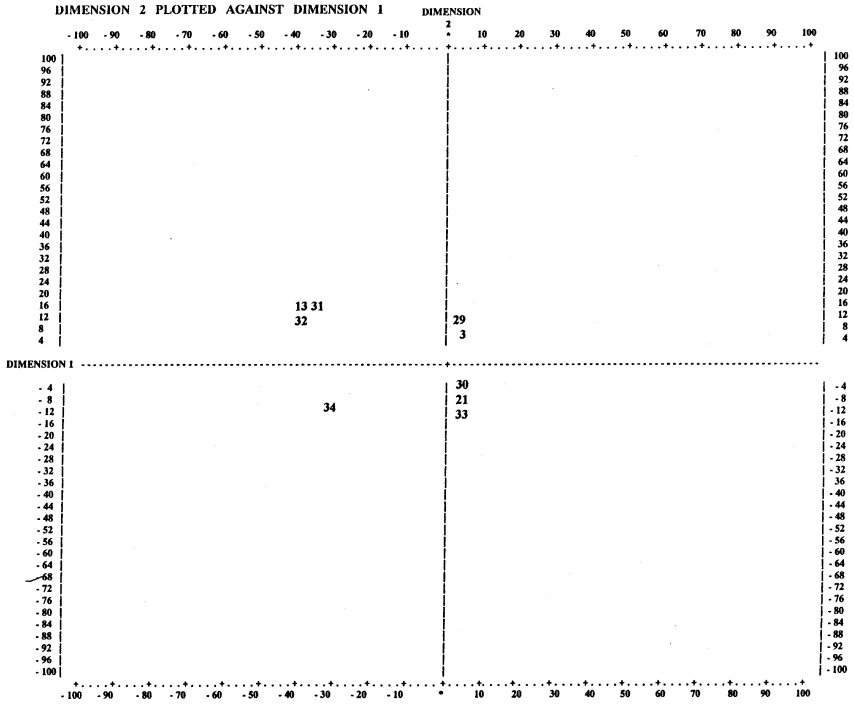
14.9.b. táblázat. MINIRSA output

	1	2		1	2
1.	0,0740	−0,2133	1.	−0,3394	−0,3819
2.	−0,0991	0,1339	2.	0,1279	0,1156
3.	0,2237	0,3264	3.	−0,2679	−0,2979
4.	0,1838	0,3818	4.	0,1298	0,1229
5.	0,2264	0,2949	5.	−0,7570	0,2292
6.	0,0714	0,1630	6.	0,7658	0,4322
7.	0,0708	−0,0160	7.	0,5432	−0,9765
8.	0,1610	−0,1130	8.	0,7690	−0,8643
9.	0,1703	−0,2002	9.	0,5563	−0,0452
10.	0,0087	−0,2100	10.	−0,8299	0,1735
11.	0,1485	−0,2841	11.	−0,7968	0,6887
12.	0,1511	0,0640	12.	1,6081	−0,3608
13.	0,2204	0,3820	13.	−0,7980	0,2479
14.	0,1339	0,4990	14.	1,0331	0,0479
15.	0,1559	0,5336	15.	−0,7513	0,1917
16.	0,1798	0,5505	16.	0,9492	1,0846
17.	−0,1622	−0,3328	17.	−0,3686	−0,1527
18.	0,1394	−0,3204	18.	1,1480	0,7749
19.	0,0435	−0,1468	19.	−0,6819	−0,9818
20.	0,1975	−0,1472	20.	−0,6765	0,3319
21.	0,0606	0,1750	21.	−0,8075	0,2087
22.	0,2003	0,3501	22.	−0,2093	−0,0367
23.	0,1410	0,5032	23.	−0,9624	0,8205
24.	0,1774	0,1984	24.	1,2041	0,0513
25.	0,1279	−0,1420	25.	0,7632	1,2469
			26.	1,0000	−1,0099
			27.	1,1020	0,2810
			28.	−0,2545	−0,4884
			29.	−0,2470	−0,2055
			30.	1,0008	−0,0139
			31.	−0,2665	−1,0955
			32.	−0,8061	−0,1453
			33.	−1,0532	−0,4277
			34.	−1,3599	−0,5255
ÁTLAG:	0,1428	0,0952	ÁTLAG:	−0,0000	−0,0000
SZÓRÁS:	0,0590	0,2859	SZÓRÁS:	0,8369	0,5474

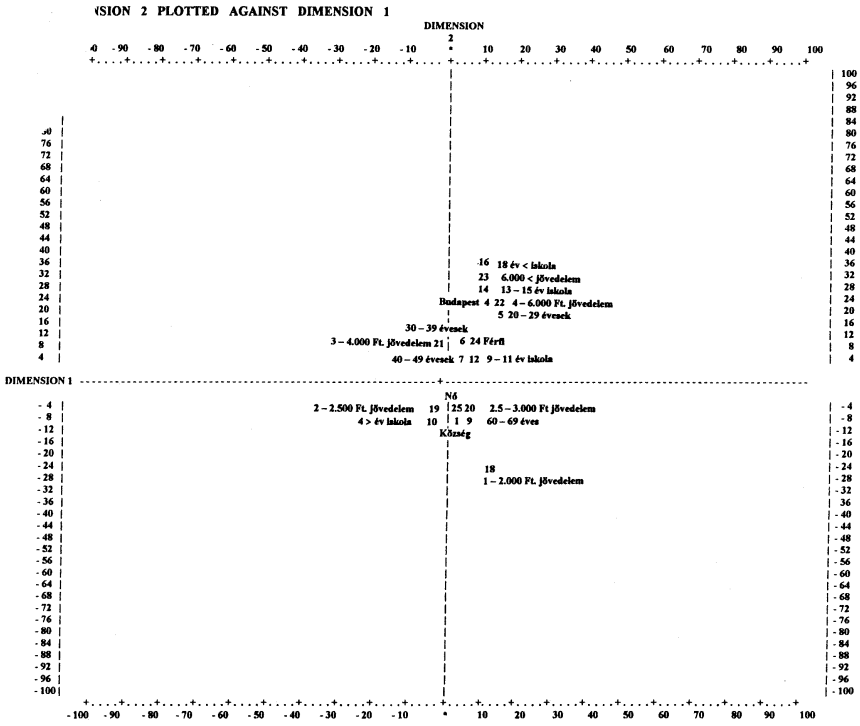
14.9.c. táblázat. MINIRSA output



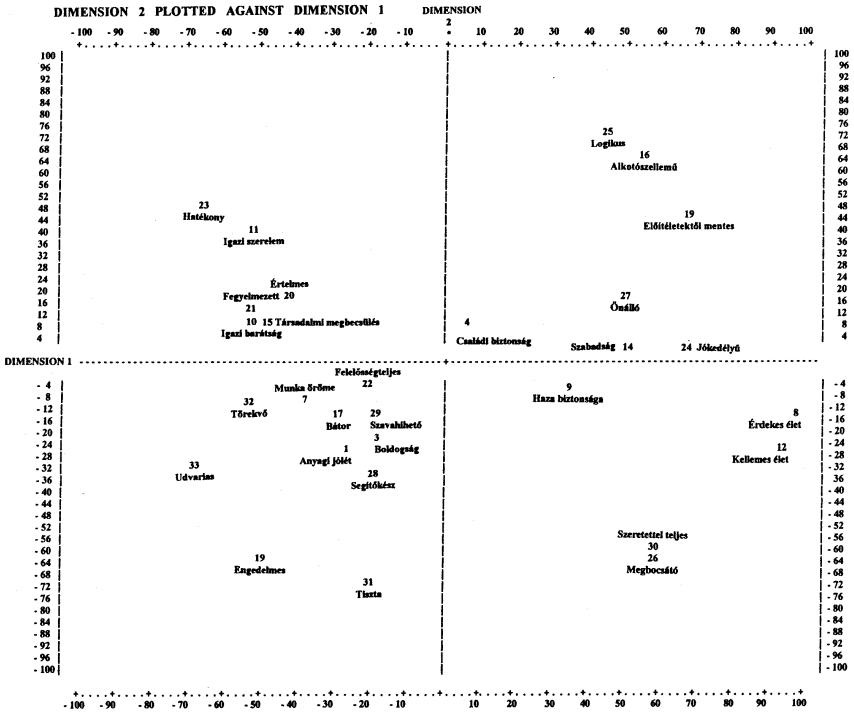
14.11. ábra. Társadalmi csoportok az értékek terében



14.12. ábra. Értékek a társadalmi csoportok terében



14.13. ábra. Társadalmi csoportok az értékpreferencia terében



14.14. ábra. Értékek a társadalmi csoportok értékpreferencia terében (MINIRSA-modell)

14.5. Az INDSCAL-modell (INdividuaD ifferences SCALing)

Az INDSCAL-modellt Carroll és Chang (1970) fejlesztette ki az egyéni különbségek skálázására.

Az INDSCAL-modellben – más sokdimenziós skálázó eljárásokhoz hasonlóan – feltételezzük, hogy a stimulusok különbözősége (vagy hasonlósága) a származtatott pontok közötti (euklideszi) távolság monoton függvénye: Az INDSCAL-modell abban különbözik a többi skálázó eljárástól, hogy meghatározza a dimenziók relatív fontosságát (súlyát) minden egyed számára. Az egyedekre vonatkozó dimenziósúlyok azt jelzik, hogy mennyire kell megnyújtani az egyes tengelyeket (dimenziókat) ahhoz, hogy a stimulusok közötti távolságok maximálisan korreláljanak a stimulusok közötti hasonlóságokkal.

Az i-edik egyed számára a j-edik és a k-adik stimulus különbözősége az INDSCAL-modell feltétele értelmében a távolságok monoton függvénye:

$$\delta_{jk,i} \approx F(d_{jk,i}),$$

ahol $d_{jk,i}$ a j-edik stimulus és k-adik stimulus közötti távolság az i-edik egyedre vonatkozóan

$$d_{jk,i} = \left(\sum_{t=1}^{r} \left(y_{jt,i} - y_{kt,i} \right)^2 \right)^{\frac{1}{2}}$$

F: monoton függvény.

Az „egyedi" távolságokat megkaphatjuk a „csoport" térkoordinátáiból is, ha a koordinátákat az egyedekre és tengelyekre vonatkozó súlyokkal transzformáljuk:

$$y_{jt,i} = \sqrt{w_{it}} x_{jt}$$

Ennek alapján:

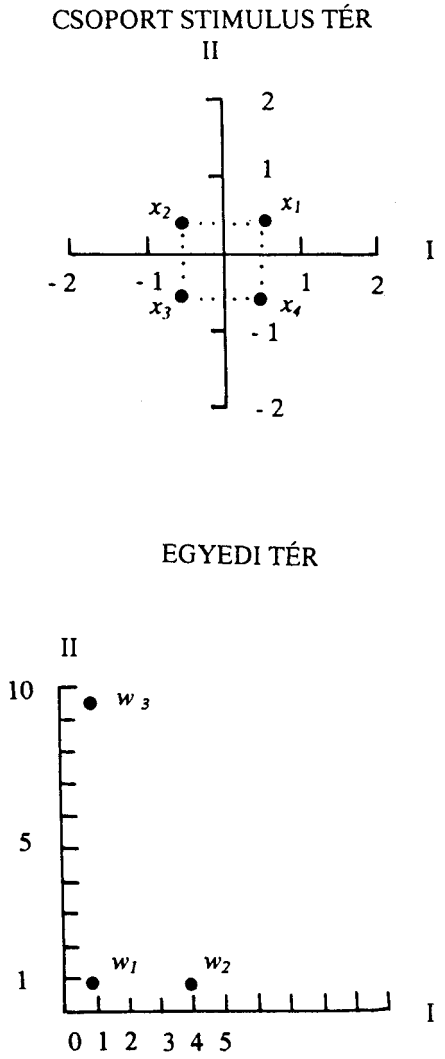
$$d_{jk,i} = \left(\sum_{t=1}^{r} w_{it} \left(x_{jt} - x_{kt} \right)^2 \right)^{\frac{1}{2}}.$$

Az INDSCAL-modell a kétutas MDS-módszerek általánosítását adja azzal, hogy az euklideszi távolság helyett súlyozott euklideszi távolságot számít.

Ha a csoport stimulus térdimenziói normalizáltak, akkor egy egyed valamely dimenzióra vonatkozó súlyának négyzete nem más, mint az egyed hasonlósági adatai varianciájának azon aránya, amennyit a kérdéses tengely magyaráz. Igy a w_{it} az i-edik egyed számára a t-edik dimenzió fontosságát jelöli.

Az INDSCAL-modell geometriai interpretációját egy fiktív példán mutatja a következő ábra. (*Forrás: CarrollD . and M. Wish: Models and Methods for Theree-way Multidimensional scaling, Bell Laboratories, New Jersey*).

A fenti hipotetikus példában a csoport stimulus tér négy objektumot tartalmaz. A megfigyelési egységeknek a csoport stimulus tér tengelyeire vonatkozó súlyait az egyedi térben találhatjuk. Eszerint az 1. sorszámú megfigyelési egység privát tere megegyezik a csoporttérrel, mivel a dimenziókat azonosan 1-gyel súlyozza. A 2. és 3. számú egyed a csoport stimulus tér dimenzióit eltérő módon értékeli, így a 2. sorszámú egyed az első dimenziónak ad nagyobb súlyt, míg a 3. számú egyed a második dimenziót értékeli fontosabbnak, vagyis az x_1 és x_2 stimulus eltávolodik az x_3 és x_4 stimulusoktól a 3. megfigyelési egység értékelése szerint.

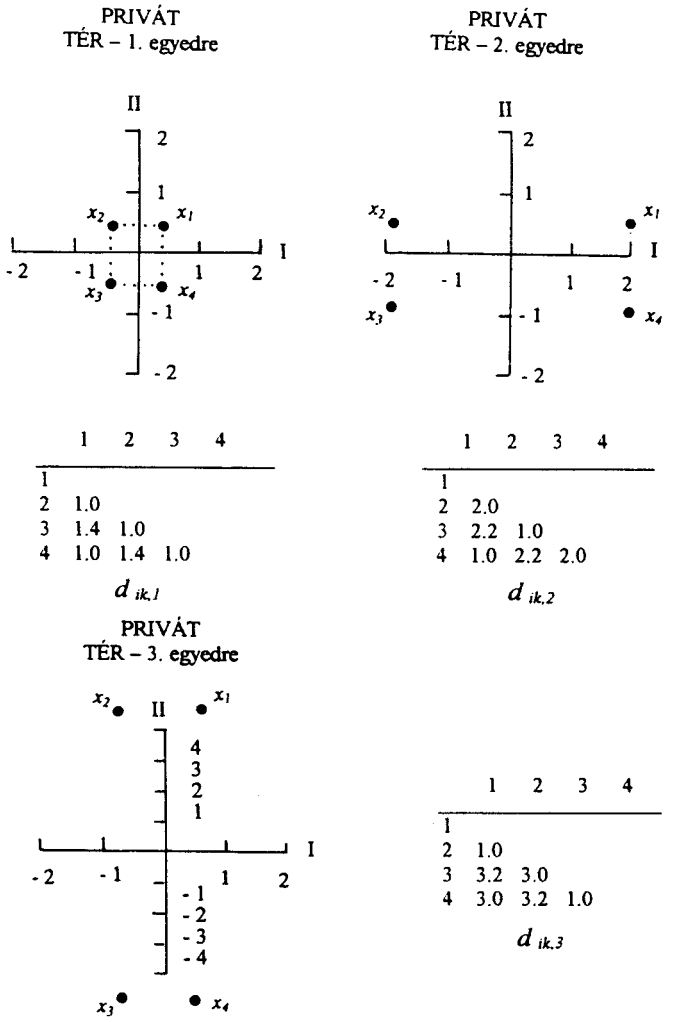


14.15. ábra. Az INDSCAL-modell geometriai interpretációja

A privát tér tehát a csoport stimulus tér dimenzióit újraskálázza az egyéni súlyoknak megfelelően.

14.5.1A z INDSCAL-modell

Az INDSCAL-módszere a megfigyelési egységek különbözőségi mátrixaiból (háromutas vagy háromdimenziós adatmátrixból) kiindulva két eredmény mátrixot határoz meg: „a csoport stimulus teret" és az „egyedi teret" („szubjektum" terét). A „csoport stimulus tér" mátrixa az n stimulus r dimenzióra vonatkozó koordinátáit tartalmazza. (A mátrix általános eleme x_{jt} = a j-edik stimulus t-edik dimenzióra vonatkozó koordinátája



14.16. ábra. Az INDSCAL-modell geometriai interpretációja

a „csoport stimulus térben".) Az „egyedi tér" mátrixa a megfigyelési egységek (egyedek) súlyait tartalmazza a „csoport stimulus tér" dimenzióira vonatkozóan. (A mátrix általános eleme w_{it} = az i-edik egyed t-edik dimenzióra vonatkozó súlya.) A két eredménymátrix az induló háromutas mátrixnak – később definiálandó értelemben – legjobb r-dimenziós közelítését adja.

A következőkben az INDSCAL-modell két output mátrixának számítási menetét mutatjuk be.

A különbözőségek átalakítása becsült távolságokká

A számítás első lépéseként a „klasszikus" metrikus kétutas" MDS-eljárásokhoz hasonlóan a különbözőségeket becsült távolságokká alakítjuk át. (Ha a kiinduló adatok

hasonlóságok, akkor azokat –1-gyel szorozva különbözőségekké alakítjuk.) Ezt az eljárást „additív konstans becslés” néven szokták említeni.

A metrikus feltételek alapján feltehetjük, hogy

$$d_{jk,i} = \delta_{jk,i} + c_i .$$

A c_i-k legkisebb értéke, amely esetén még teljesül a távolságokra a háromszög egyenlőtlenség ($d_{j\ell} \leq d_{jk} + d_{k\ell}$):

$$c_{\min} = \max_{j,k,\ell}(\delta_{j\ell} - \delta_{jk} - \delta_{k\ell}).$$

A $c_{\min}$ értéke a képletből adódóan lehet negatív is.

Ez az „additív konstans” módszer, amelyet Torgerson (1958) írt le mint egy „egydimenizós altér” sémát, a „komparatív távolságokat” abszolút távolságokká alakítja.

A becsült távolságoknak becsült skaláris szorzatokká való átalakítása

Az r-dimenziós tér két pontja (vektora):

$$\mathbf{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jr}) ,$$

és

$$\mathbf{x}_k = (x_{k1}, x_{k2}, \ldots, x_{kr}).$$

A két vektor skaláris szorzata:

$$b_{jk} = \mathbf{x}'_j \mathbf{x}_k = \sum_{t=1}^{r} x_{jt} x_{kt} .$$

Két vektor skaláris szorzata geometriailag úgy értelmezhető, mint a hosszúságaik szorzatával beszorzott két vektor által bezárt szög koszinusza: $b_{jk} = |\mathbf{x}_j| \cdot |\mathbf{x}_k| \cos(x_j, x_k)$.

Az MDS-módszereknél a tér origójának nincs jelentősége, ezért az origót a pontok súlypontjába szokták helyezni. Ebben az esetben $\sum_{j=1}^{n} x_{jt} = 0$ minden $t = 1, 2, \ldots, r$ tengelyre.

Ezt a feltevést azért tehetjük, mert az euklideszi távolságok invariánsak egy konstans hozzáadásával szemben (minden ponthoz ugyanazt a konstans vektort hozzáadva a pontok közötti távolságok nem változnak). A skaláris szorzatokra ez azonban nem igaz. Ezért a b_{jk} vektorok skaláris szorzata jelölésénél mindig feltesszük, hogy az origó a súlypontban helyezkedik el. Torgerson (1958) megmutatta, hogy az euklideszi távolságok átalakíthatók a pontok súlypontjába helyezett origó szerinti vektorok skaláris szorzatává:

$$b_{jk} = -\frac{1}{2}\left(d^2_{jk} - d^2_{.k} - d^2_{j.} + d^2_{..}\right),$$

ahol

$$d^2_{.k} = \frac{1}{n}\sum_j d^2_{jk}$$

$$d^2_{j.} = \frac{1}{n}\sum_k d^2_{jk}$$

$$d^2_{..} = \frac{1}{n}\sum_j \sum_k d^2_{jk} \quad \text{és}$$

$$d^2_{jk} = \sum_{t=1}^{r}(x_{jt} - x_{kt})^2 .$$

A fenti képletet használhatjuk a becsült távolságok átalakítására becsült skaláris szorzatokká. ($\widehat{d}_{jk} = \delta_{jk} + c_{\min}$).

A b_{jk} becsült szorzatokat mátrix alakba írva Torgerson egyenlete a következő:

$$\widehat{\mathbf{B}} = -\frac{1}{2}\widehat{\mathbf{D}},$$

ahol:

$$\widehat{d}^{*2}_{jk} = \widehat{d}^{2}_{jk} - \widehat{d}_{.k} - \widehat{d}_{j.} + \widehat{d}^{2}_{..},$$

vagyis $\widehat{\mathbf{D}}$ mátrix a távolságok kettős centrirozása utáni távolságokat tartalmazza.

Ha a becsült koordinátákat ($\widehat{x}_{jt}$) az $\widehat{\mathbf{X}}$ mátrix tartalmazza, a skaláris szorzat mátrix a következő egyenlettel egyezik meg:

$$\widehat{\mathbf{B}} = \widehat{\mathbf{X}}\,\widehat{\mathbf{X}}'.$$

Ez a mátrix-faktorizáció a főkomponens- vagy faktorelemzésnél, illetve az ezekhez kapcsolódó módszereknél fordul elő, amikor egy szimmetrikus mátrixot két kisebb rangú mátrix szorzatára bontunk.

A CANDECOMP-eljárás

Az INDSCAL-modellben a skaláris szorzat háromdimenziós tömbjét kapjuk azzal, hogy minden egyedre definiáljuk a szorzatokat:

$$b_{jk,i} = \sum_t y_{jt,i}\, y_{kt,i},$$

ahol:

$$y_{jt,i} = \sqrt{w_{it}}\, x_{jt}$$

és ennek alapján:

$$b_{jk,i} = \sum_t w_{it} x_{jt} x_{kt}.$$

Ezt hívjuk az INDSCAL-modell skaláris szorzat formájának, amit a CANDECOMP-(CANonical DECOMPosition of N-way tables) modell speciális esetének tekinthetünk ($N = 3$):

$$z_{ijk} = \sum_t a_{it} b_{jt} c_{kt}.$$

A fenti CANDECOMP-modellbe helyettesítve az INDSCAL változóit (azt nevezzük az INDSCAL-programban INDIFF-modellnek „INdividual DIFFerences Scaling”):

$$z_{ijk} = b_{jk,i}$$
$$a_{it} = w_{it}$$
$$b_{jt} = c_{jt} = x_{jt}.$$

A modell két paraméterhalmazát (b_{jt} és c_{kt}) a legkisebb négyzetek módszerével becsüljük. A becslést az egyenlet újrafogalmazásával kaphatjuk meg. Vezessük be a következő jelöléseket:

$$\widehat{z}^{*}_{is} = \widehat{z}_{ijk}$$
$$\widehat{g}_{st} = \widehat{b}_{jt}\widehat{c}_{kt} \quad \text{ahol } \widehat{b}_{jt} \text{ és } \widehat{c}_{kt} \text{ a } b_{jt} \text{ és } c_{kt}$$

aktuális becsült értékei.

A CANDECOMP-modell ennek alapján:

$$\hat{z}^*_{is} = \sum_t a_{it} g_{st}.$$

Ezzel az egyszerűsítéssel a trilineáris modellt bilineárissá alakítottuk át. Mátrix jelölésekkel:

$$\hat{\mathbf{Z}}^* = \mathbf{A}\,\mathbf{G}'.$$

Az $\mathbf{A}$ legkisebb négyzetek módszere szerinti becslését a következőképpen kapjuk:

$$\hat{\mathbf{A}} = \hat{\mathbf{Z}}^*\,\hat{\mathbf{G}}(\hat{\mathbf{G}}'\,\hat{\mathbf{G}})^{-1}.$$

Iterációs eljárás

Az általános becslési eljárás, amit az INDSCAL-modellben alkalmazunk a következő: (ezt nevezte Wold (1966) NILES-eljárásnak (Nonlinear Iterative LEast Squares)).

Rögzített b és c értékekre a legkisebb négyzetek módszere szerint becsüljük az a értékeit, majd a legkisebb négyzetek módszere szerint becsüljük a b-ket rögzített a és c értékekkel, és az iterációt addig folytatjuk amíg konvergens becslésekhez nem jutunk. Ez az eljárás nem biztosítja minden esetben a globális minimumot, azonban a gyakorlati számítások azt mutatják, hogy „majdnem mindig" globális optimumhoz jutunk.

Carroll és Chang (1970) a fenti metrikus eljárást a Kruskal (1964) -féle monoton regressziós eljárás alkalmazásával kvázi nemmetrikus eljárássá alakította át. Az iteratív eljárás Carroll és Chang módszerében a következő:

Adottak a $\delta_{jk,i}$ különbözőségi adatok minden egyedre. Először a c_i additiv konstans értékét becsüljük és alakítjuk át távolsággá: ${}_{(0)}\hat{d}_{jk,i}$. Ezen „külső" iterációs ciklus az I-edik lépésben a következő:

$$\text{adott} \quad \delta_{jk,i} \quad \text{és} \quad {}_{(I-1)}\hat{d}_{jk,i}.$$

Első fázis:

a) átalakítjuk a becsült távolságokat becsült skaláris szorzatokká:

$${}_{(I-1)}\hat{d}_{jk,i} \longrightarrow {}_{(I-1)}\hat{b}_{jk,i}.$$

b) a $\hat{b}$ értékekre alkalmazzuk a CANDECOMP-eljárást és normalizálunk. Eredményül a következőt kapjuk:

$${}_{(I)}\mathbf{X} = \left\{{}_{(I)}x_{jt}\right\} \quad \text{és} \quad {}_{(I)}\mathbf{W} = \left\{{}_{(I)}w_{it}\right\}.$$

Második fázis:

a) a távolságok kiszámításához minden i, j, k-ra a súlyozott euklideszi távolság formuláját használjuk:

$$d^2_{jk,i} = \sum_{t=1}^{r} w_{it}(x_{jt,i} - x_{kt,i})^2.$$

b) a legkisebb négyzetek monoton regresszióját alkalmazzuk (MFIT-rutin) a távolságok becsléséhez:

$${}_{(I)}\hat{d}_{jk,i} = {}_{(I)}F_i(\delta_{jk,i}) \approx {}_{(I)}d_{jk,i},$$

ahol ${}_{(I)}F_i$ egy monoton nemcsökkenő függvény.

Az iteráció I számának növelésével folytatódik, és visszatér az első fázis elejére egészen addig, amíg az illeszkedés tovább már nem javul. Az illeszkedés jóságát a Kruskal-féle STRESSFORM2 számításával mérjük:

$$S = \sqrt{\sum_i \left[\frac{\sum_j \sum_k \left(d_{jk,i} - \hat{d}_{jk,i}\right)^2}{\sum_j \sum_k \left(d_{jk,i} - \bar{d}_i\right)^2} \right]}.$$

Az S értéke az iterációk sorozatán nem monoton csökkenő, mivel az iteráció két fázisában különböző kritériumot optimalizáltunk (az első fázisban a skaláris szorzatokat, a második fázisban a távolságokat). A Stress-kritériumot Carroll és Chang helyettesítette a következővel:

$$\text{STRAIN} = \sqrt{\sum_i \left[\frac{\sum_j \sum_k \left(b_{jk,i} - \hat{b}_{jk,i}\right)^2}{\sum_j \sum_k b_{jk,i}^2} \right]},$$

ahol

$$b_{jk,i} = \sum_i w_{it} x_{jt} x_{kt} \quad \text{és}$$

$$\hat{b}_{jk,i} = -\frac{1}{2}\left(\hat{d}_{jk,i}^2 - \hat{d}_{.k,i}^2 - \hat{d}_{j.,i}^2 + d_{..}^2\right)$$

$$\hat{d}_{jk,i} = F_i(\delta_{jk,i})$$

Ekkor az iteráció lépéseinek második fázisában a STRAIN-kritériumot optimalizáló F_i értékeket keressük, míg a többi paramétert konstansnak tekintjük. Így a második fázisban is a skaláris szorzatokban kapjuk a legkisebb négyzeteket.

Az INDSCAL-modell egyértékűség tulajdonsága

Az INDSCAL-megoldás egyértékű, ha

1. a csoport stimulus tér dimenziói függetlenek,
2. legkevesebb két egyedre vonatkozó adatmátrixunk van,
3. nincs két dimenzió, amelyre fennáll a súlyok „párhuzamos minta” („parallel pattern”) tulajdonsága.

A harmadik feltételt a következőképpen értelmezzük: egy dimenziópárra (s és t) vonatkozó súlyok akkor és csakis akkor mutatnak párhuzamos mintát, ha minden i és j egyedpárra igaz a következő egyenlőség:

$$w_{is} \cdot w_{jt} = w_{it} \cdot w_{js}.$$

Geometriailag ez azt jelenti, hogy az egyedi tér s-edik és t-edik dimenzióknak megfelelő kétdimenziós alterében az egyedek súlyai egy, az origón átmenő egyenesen fekszenek.

14.5.2. Az IDIOSCAL-modell

A háromutas adatmátrixok elemzésére kidolgozott MDS-modellek az INDSCAL-modellből, mint alapmodellből indultak ki, és elsősorban a távolságok definiálásában ($d_{jk,i}$) különböznek.

A legáltalánosabbnak tekintett háromutas MDS-módszert, az IDIOSCAL-t *(Individual Differences In Orientation SCALing)* Carroll és Chang (1972) fejlesztette ki. A távolságokat az általánosított euklideszi távolsággal határozzák meg:

$$d_{jk,i} = \sqrt{\sum_{t}^{r}\sum_{t'}^{r}(x_{jt} - x_{kt})c_{tt',i}(x_{jt'} - x_{kt'})}$$

ahol $\mathbf{C}_i = \{c_{tt',i}\}$ egy $r \times r$ típusú szimmetrikus definit vagy szemidefinit mátrix.

A modellben a skaláris szorzat a következő:

$$b_{jk,i} = \sum_{t}\sum_{t'} x_{jt}c_{tt',i}x_{kt'}$$

Mátrix jelölésekkel:

$$\mathbf{B}_i = \mathbf{X}\,\mathbf{C}_i\,\mathbf{X}'.$$

A $\mathbf{C}_i$ dekompozíciójára két eljárást javasoltak.

Carroll–Chang-eljárás a $\mathbf{C}_i$ dekompozíciójára

A $\mathbf{C}_i$ Carroll–Chang-féle felbontása:

$$\mathbf{C}_i = \mathbf{T}_i\beta_i\mathbf{T}_i',$$

ahol: $\mathbf{T}_i$ ortogonális és β_i diagonális mátrixok.

Geometriailag ez úgy értelmezhető, mint a csoport stimulus tér ortogonális rotációja, amelyik egy i-edik egyed új koordinátarendszerét adja a β súlyozás után. Egy új jelölés bevezetésével ezt más módon is leírhatjuk:

$$\mathbf{S}_i = \mathbf{T}_i\,\beta_i^{\frac{1}{2}}$$

és így:

$$\mathbf{C}_i = \mathbf{S}_i\,\mathbf{S}_i'.$$

Az i-edik egyed „privát terét” a „csoport tér” lineáris transzformációjával kapjuk. Erre a transzformációra igaz, hogy bármely ortogonális $\mathbf{U}$ mátrixra érvényes, hogy

$$\mathbf{S}_i^* = \mathbf{S}_i\,\mathbf{U}$$

$$\mathbf{S}_i^*\,\mathbf{S}_i^{*\prime} = \mathbf{S}_i\mathbf{U}\,\mathbf{U}'\,\mathbf{S}_i' = \mathbf{S}_i\,\mathbf{S}_i' = \mathbf{C}_i.$$

Tucker és Harshman eljárása $\mathbf{C}_i$ dekompozíciójára

A $\mathbf{C}_i$ Tucker- és Harshman-féle felbontása:

$$\mathbf{C}_i = \mathbf{D}_i\,\mathbf{R}_i\,\mathbf{D}_i$$

ahol $\mathbf{D}_i$ diagonális mátrix, $\mathbf{R}_i$ szimmetrikus mátrix, diagonális elemei egyenlőek 1-gyel.

Tucker és Harshman az $\mathbf{R}_i$ mátrixot korrelációs mátrixként értelmezte (vagy általánosabban a dimenziók közötti hajlásszög koszinuszaiként), a $\mathbf{D}_i$ mátrix diagonális elemei

a dimenziókra vonatkozó koordináták szórásai. Így a $\mathbf{C}_i$ mátrix kovarianciamátrixnak tekinthető. Amennyiben a dimenziók ortogonálisak, az $\mathbf{R}_i$ egységmátrix lesz, és így $\mathbf{C}_i = \mathbf{D}_i^2$ diagonális mátrix. Ha ez minden i-re (egyedre) igaz, akkor speciális esetként az INDSCAL-modellt kapjuk.

14.5.3. Harshman PARAFAC-2 modellje

Harshman (1972) PARAFAC (PARAllel FACtors)-2 modellje az IDIOSCAL-modell egy speciális esete, amelyben a $\mathbf{C}_i$ a következő:

$$\mathbf{C}_i = \mathbf{D}_i \mathbf{R} \mathbf{D}_i .$$

Az $\mathbf{R}$ minden egyedre azonos, de a $\mathbf{D}_i$ súlyokat vagy újraskálázó faktorokat tartalmazó mátrix különböző. Harshman az $\mathbf{R}$-t úgy értelmezte, mint a nem merőleges tengelyek közötti hajlásszögek koszinuszait (korrelációit) tartalmazó mátrixot.

A PARAFAC-2 modell

A csoport stimulus tér koordinátáit a ferdeszögű koordinátarendszerben az $\mathbf{X}$ mátrix tartalmazza. Ezeket a tengelyeket minden egyed különbözőképpen súlyozhatja, ezért

$$\mathbf{X}_i^* = \mathbf{X} \mathbf{D}_i ,$$

ahol $\mathbf{D}_i$ diagonális mátrix.

Az $\mathbf{X}_i$-ben lévő i-edik egyed pontjait egy $\mathbf{T}$ lineáris transzformációval visszük át egy ortogonális koordinátarendszerbe, ahol már számíthatjuk az euklideszi távolságot a pontok különbözőségének mérésére:

$$\mathbf{X}_i^\circ = \mathbf{X}_i^* \mathbf{T} = \mathbf{X} \mathbf{D}_i \mathbf{T},$$

ahol $\mathbf{X}_i^\circ$ a stimulusok pontjait tartalmazza a derékszögű koordináta-rendszerben.

Az $\mathbf{X}_i^\circ$ -vel definiált térben egyszerű euklideszi távolságokat számolva a PARAFAC-2 modellben a $\mathbf{C}_i$ a következő:

$$\mathbf{C}_i = \mathbf{D}_i \mathbf{T} \mathbf{T}' \mathbf{D}_i = \mathbf{D}_i^* \mathbf{R} \mathbf{D}_i^*,$$

ahol $\mathbf{R}$ a $(\mathbf{T}\mathbf{T}')$ egy normalizált változata

$$\mathbf{R} = \mathbf{E} \mathbf{T} \mathbf{T}' \mathbf{E} \quad \text{és} \quad \mathbf{E} = \left\{ \left[diag(\mathbf{T}\mathbf{T}') \right]^{-\frac{1}{2}} \right\},$$

valamint $\mathbf{D}_i = \mathbf{D}_i \mathbf{E}^{-1}$.

14.5.4. Tucker háromszempontú skálázó modellje

Tucker (1972) a háromszempontú (three-mode) faktorelemzés módszerét alkalmazta a sokdimenziós skálázásra. Ebben a modellben $\mathbf{C}_i$ a következő:

$$\mathbf{C}_i = \sum_{s=1}^{S} a_{is} \mathbf{G}_s ,$$

ahol $\mathbf{G}_s$ az s-edik „magmátrix”.

A $\mathbf{C}_i$ általános eleme:

$$c_{tt',i} = \sum_{s}^{S} a_{is} g_{stt'}.$$

Az a_{is} az S-dimenziós egyedi tér egy pontja (az S nem szükségszerűen egyenlő r-rel), és a $g_{stt'}$ a „magmátrix" (belső struktúra-mátrix) egy eleme. Ha a $\mathbf{C}_i$ mátrix előállításában S értékét elég nagynak választjuk, akkor tetszőleges $\mathbf{C}_i$ mátrixot elő tudunk állítani a fenti módon.

A Tucker-féle háromszempontú skálázási módszer néhány jellemzője:

– A tengelyek irányítottsága nem egyértelmű. A stimulus térre és az egyedi térre bármely nemszinguláris lineáris transzformációt alkalmazunk, tudunk találni olyan „magmátrixot", amelyik ugyanazokat az adatokat adja.

– Az egyedi tér és a stimulus tér dimenziószámának nem kell megegyeznie. Az egyedi tér dimenziószáma lehet több is és kevesebb is, mint a stimulus téré.

– Az egyedi tér és a stimulus tér egyedül nem elégséges a struktúra leírására, ehhez a „magmátrixra" is szükség van. A magmátrix értelmezése azonban nem egyértelmű, bár Tucker (1972) javasolt egy módszert, amit „ideális egyed" („ideal individual") koncepciónak nevezett. A megközelítés analóg azzal, amit Tucker és Messick (1963) a „points-of-view" modellben használt.

A három-szempontú skálázó modell megegyezik az INDSCAL-modellel, ha $S = r$, és léteznek a stimulus és egyedi tér lineáris transzformációi, úgy, hogy a „magmátrix" diagonális $r \times r$ típusú mátrixokat tartalmaz.

14.5.5. Többutas MDS-modellek

A többutas MDS-modellek szükségessége akkor merül fel, amikor minden egyedre vonatkozóan két vagy több sorozat adat (különbözőségi mátrix) áll rendelkezésünkre, így a kiinduló adatok négy- vagy többdimenziós adattömböt alkotnak. Például Wish elemzett egy négydimenziós adatmátrixot, ahol a megfigyelési egységek különféle szempontok szerint hasonlították össze a vizsgált nemzeteket.

A többutas INDSCAL-modell

Az INDSCAL-modell négyutas általánosítását skaláris szorzat formában a következő formula fejezi ki:

$$b_{jk,i\ell} = \sum_{t} w_{it} v_{\ell t} x_{jt} x_{kt},$$

ahol $b_{jk,i\ell}$ a j-edik és k-adik stimulus közötti skaláris szorzat az i-edik egyed ℓ-edik szempontja (út) szerint.

Az N-utas általánosítás egyenlete:

$$b_{jk,i_1,i_2,\ldots i_{N-2}} = \sum_{t} w_{i_1 t} w_{i_2 t} \ldots w_{i_{N-2} t} x_{jt} x_{kt},$$

ahol w indexei a különböző szempontokra vonatkoznak.

Az INDSCAL-program jelenlegi verziója az általánosítás $N \leq 7$ esetére alkalmas.

A többszempontú skálázás Tucker-féle modellje

A háromszempontú skálázás négy-dimenziós általánosítása:

$$b_{jk,i\ell} = \sum_s \sum_u \sum_t \sum_{t'} a_{is} a^*_{\ell u} g_{sutt'} x_{jt} x_{kt'},$$

ahol $g_{sutt'} = g_{sut't}$.

A $\mathbf{C}_{i\ell}$ négydimenziós változata:

$$\mathbf{C}_{i\ell} = \sum_{s=1}^{S} \sum_{u=1}^{U} a_{is} a_{\ell u} \mathbf{G}_{su},$$

ahol $\mathbf{C}_{i\ell}$ és $\mathbf{G}_{su}$ szimmetrikus $r \times r$ típusú mátrixok.

Az N-szempontú eset egyenlete:

$$b_{jk,i_1,i_2,\ldots,i_{N-2}} =$$
$$= \sum_{t_1} \sum_{t_2} \cdots \sum_{t_{N-2}} a_{i_1 t_1} \cdot a_{i_2 t_2} \cdot a_{i_{N-2} t_{N-2}} g_{t_1 t_2} \cdots g_{t_{N-1} t_N} \cdot x_{j t_{N-1}} x_{k t_N}.$$

Ezt az általánosítást Tucker (1964, 1972) fejtette ki.

Az eddig tárgyalt modellek mindegyike az euklideszi távolság fogalmát használta. Az általánosítás egyik formája lehet az is, amikor az euklideszi távolság helyett más metrikát használunk, pl. a Minkowszki-metrikát. Bővebben erről lásd Carroll és Wish (1973) művében.

Az általánosítás másik iránya, amely a csoport stimulus tér nemlineáris transzformációját végzi el. Kétféle próbálkozás történt ilyen irányban:

1) Mindegyik koordináta-tengelyre monoton transzformációt hajtunk végre:

$$y_{jt,i} = f_{it}(x_{jt}),$$

ahol f_{it} monoton nemlineáris függvény.

Ezeket a függvényeket úgy szokták értelmezni, mint különféle „pszichofizikai" transzformációkat.

2) a privát tér pontjait a csoport stimulus térből úgy származtatjuk, hogy minden egyed „ideális pontjától" mért távolság monoton növekvő függvénye szerint végezzük el a transzformációt. Ilyen transzformáció a következő:

$$\mathbf{y}_{j,i} = \left[\frac{\mathbf{x} - p_i}{|x_j - p_i|} \right] g_i \left(|\mathbf{x}_j - p_i| \right).$$

Ennek a modellnek egy földrajzi illusztrációját adja Carroll és Wish (1974) egy new-york-i ember példáján keresztül, aki azt gondolja, hogy Los Angeles és San Francisco nagyon közel vannak egymáshoz (mivel mindkettő távoli), és New York és Boston sokkal távolabb van (feltételezhetően azért, mert mind New York, mind Boston közel van az Ő "előnyös pontjához").

14.5.6. Példa az INDSCAL-modellre (I. Különbségek a nemzetek hasonlóságának megítélésében, II. A társadalom értékstruktúrájának vizsgálata)

Különbségek a nemzetek hasonlóságának megítélésében

Myron Wish (1974) két munkatársával (Morton Deutsch és Lois Biener) egy kutatómunka részeként vizsgálta 21 nemzet hasonlóságát, ahogyan azokat az emberek megítélik. Arra voltak kíváncsiak, hogy a fejlett és fejlődő országokból származó egyetemi hallgatók a nemzetek jellemzőiben milyen dimenziói szerint mekkora különbségeket észlelnek az országok között.

Az egyedek

A vizsgálat megfigyelési egységei egyetemi hallgatók, a Columbia Egyetem diákjai közül kerültek ki. Összesen nyolc különböző országból származtak, az egyetem külföldi diákjait reprezentálták. Komplett adathalmazt 75 diáktól kaptak.

Stimulus

A vizsgálatba bevont nemzeteket egy korábbi vizsgálat során elemezték már különböző objektív és szubjektív jellemzők szerint. A kiválasztott 21 nemzet nem tekinthető az összes nemzet véletlen mintájának. Elsősorban a nagy népességszámú vagy méretű országok közül választottak.

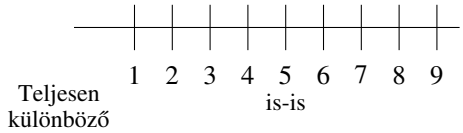
Az országok listája:

Brazília	Franciaország	Lengyelország
Kína	Görögország	Egyesült Államok
Kongó	India	Szovjetunió
Kuba	Indonézia	Dél-Afrika
Egyiptom	Izrael	Spanyolország
Egyesült Királyság	Japán	Nyugat-Németország
Etiópia	Mexikó	Jugoszlávia

A kérdőív

A megfigyelési egységek (diákok) az országok minden lehetséges párosítására (összesen 210 pár) egy kilenc pontú skálán jelölték meg „mindent egybevetve” a hasonlóság vagy különbözőség mértékét.

A hasonlósági skála:



A megkérdezettek minden országot elhelyeztek a következő kétpólusú skálákon.

Skála: 1	9
Ellenszenves	Rokonszenves
Szegény	Gazdag
Agresszív	Békés
Gyenge	Erős, hatalmas
Közösségorientált	Individualista
Nemiparosodott	Iparosodott
Belsőleg megosztott	Belsőleg egységes
Kicsi	Nagy
Hanyatló	Fejlődő
Kicsi hatással van más nemzetek kultúrájára	Nagy hatással van más nemzetek kultúrájára
Kicsi a lehetőség az emberek társadalmi státusának megváltozására	Nagy a lehetőség az emberek társadalmi státusának megváltozására
Instabil	Stabil
Rossz	Jó
Az embereknek kevés joguk van	Az emberek jogainak a köre széles
Írástudatlan népesség	Magasan iskolázott népesség
A népesség elégedetlen	A népesség elégedett

Ezen kívül még két kérdést válaszoltak meg a diákok. Az első az országok politikai hovatartozását tudakolta egy esetleges nagyobb háború esetén, amelyben az USA is részt vesz, a második az USA vietnami háborúját megítélő különböző megállapítások közüli választást kérte.

Az INDSCAL-modell eredményei

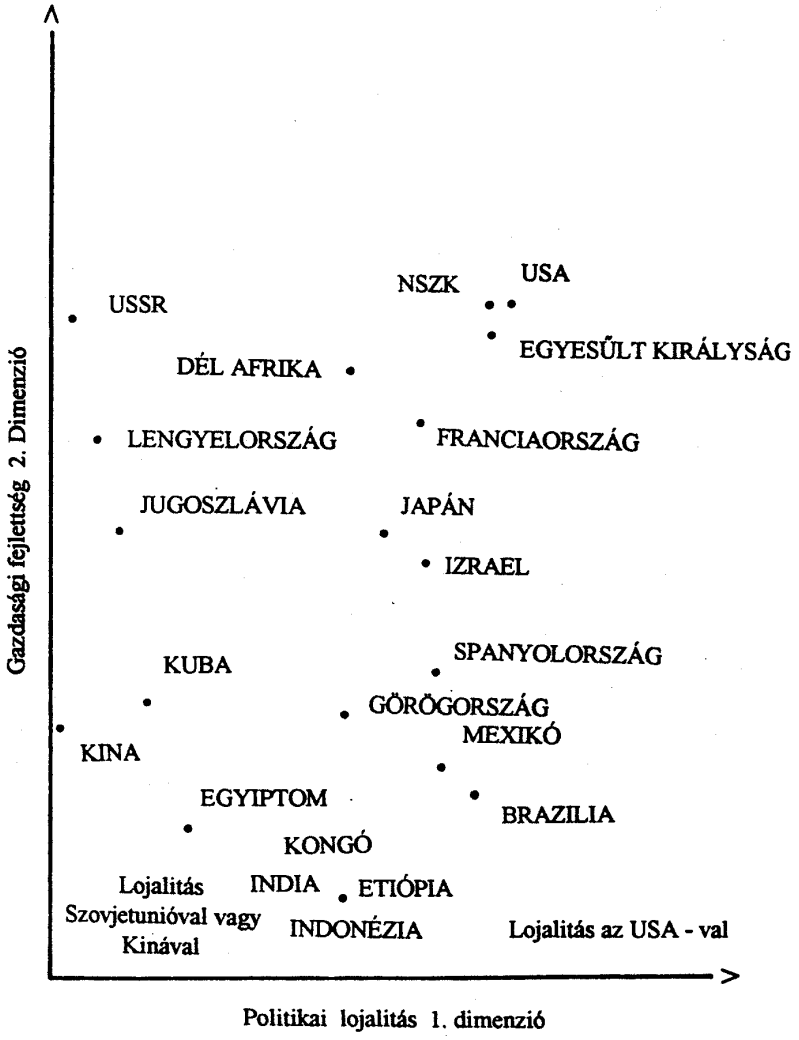
A négydimenziós INDSCAL-modell rotálatlan dimenzióit tartalmazza a következő táblázat.

USA	0,290	0,334	0,172	0,160
Egyesült Királyság	0,263	0,310	0,096	0,192
NSZK	0,261	0,326	0,141	0,204
Franciaország	0,175	0,223	−0,050	0,155
Izrael	0,168	0,108	−0,169	0,212
Japán	0,120	0,115	0,456	0,106
Dél-Afrika	0,069	0,268	−0,283	−0,493
Görögország	0,065	−0,092	−0,197	0,290
Spanyolország	0,189	−0,038	−0,178	−0,026
Brazília	0,212	−0,186	0,016	−0,168
Mexikó	0,196	−0,156	0,016	−0,083
Etiópia	−0,012	−0,296	−0,353	−0,039
India	−0,064	−0,295	0,250	−0,102
Indonézia	−0,052	−0,301	0,268	−0,123
Kongó	−0,086	−0,283	−0,237	−0,442
Egyiptom	−0,215	−0,220	−0,204	−0,076
Kína	−0,403	−0,114	0,339	−0,174
Kuba	−0,283	−0,080	−0,082	−0,212
Jugoszlávia	−0,234	0,051	−0,033	0,226
Lengyelország	−0,308	0,126	−0,131	0,299
Szovjetunió	−0,353	0,201	0,102	0,094

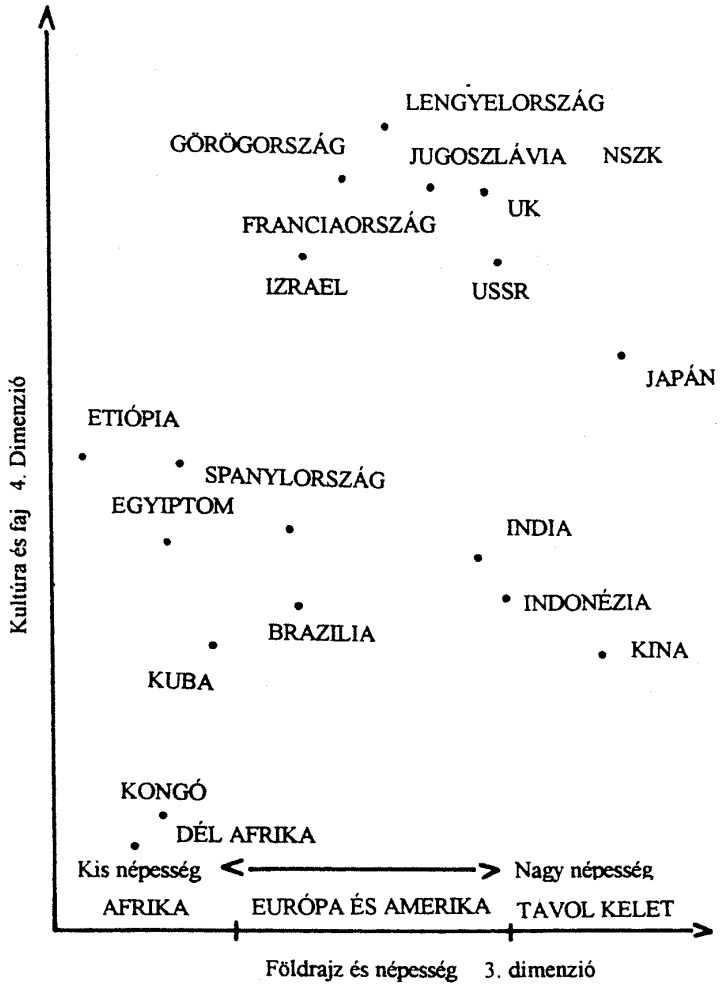
14.10. táblázat. INDSCAL-koordináták 21 országra

Az első dimenziót a politikai lojalitás tengelyének nevezték el, amelynek egyik végén az USA-val lojális országok, másik végén a Szovjetunióval vagy Kínával lojális országok helyezkednek el. A második tengely a gazdasági fejlettség nevet kapta. A harmadik dimenzió a földrajzi elhelyezkedéssel és a népesség különböző jellemzőivel függött össze, ezért a „földrajz és népesség" címet kapta. A negyedik dimenzió elnevezése „kultúra és faj".

Az első és második, valamint a harmadik és negyedik dimenziók szerint a csoport stimulus tér a következő:



14.17. ábra. Nemzetek hasonlósága a csoport stimulus térben
[1. és 2. dimenzió]



14.18. ábra. Nemzetek hasonlósága a csoport stimulus térben
[3. és 4. dimenzió]

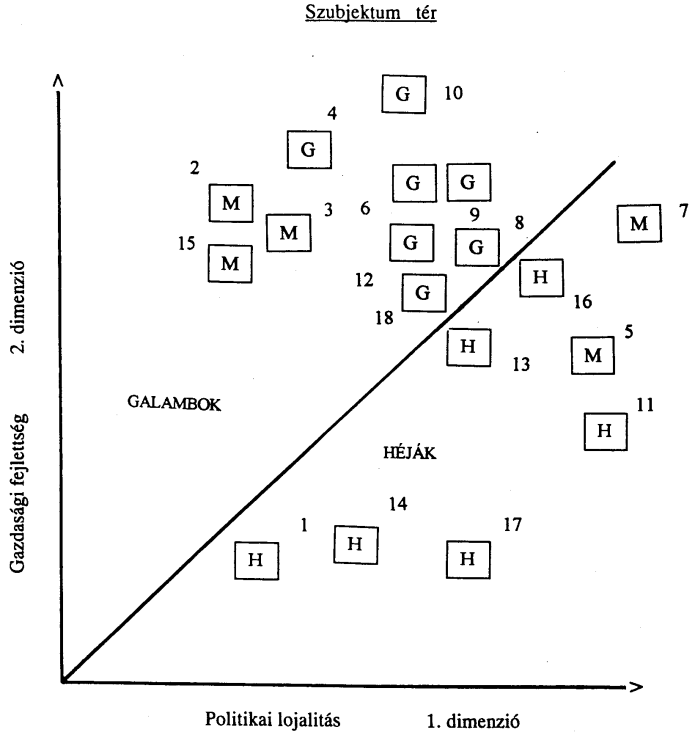
	1. dim. Politikai lojalitás	2. dim. Gazdasági fejlettség	3. dim. Földrajz és népesség	4. dim. Kultúra és faj	Többszörös korreláció
1. Lojalitás az USA-val	0,965	0,379	0,012	0,299	0,974
2. Individualitás	0,818	0,103	−0,124	−0,159	0,876
3. Béke	0,500	−0,043	−0,065	0,201	0,596
4. Emberi jogok	0,740	0,535	0,273	0,499	0,874
5. Rokonszenv	0,606	0,301	0,246	0,700	0,890
6. Jó	0,556	0,482	0,395	0,660	0,861
7. Hasonlóság az ideálishoz	0,498	0,735	0,390	0,736	0,940
8. Mobilitás	0,548	0,676	0,384	0,631	0,885
9. Stabilitás	0,233	0,755	0,318	0,641	0,846
10. Elégedettség	0,350	0,703	0,241	0,708	0,842
11. Belső egység	0,022	0,423	0,116	0,672	0,698
12. Kulturális hatás	0,321	0,546	0,415	0,560	0,738
13. Iskolázottság	0,336	0,890	0,268	0,741	0,973
14. Gazdagság	0,461	0,906	0,255	0,405	0,936
15. Iparosodottság	0,316	0,924	0,407	0,537	0,975
16. Hatalom	0,142	0,748	0,556	0,356	0,885
17. Haladás	−0,000	0,461	0,544	0,434	0,715
18. Nagyság	−0,099	0,056	0,522	−0,203	0,614

14.11. táblázat. Az INDSCAL-dimenziók korrelációi a 18 bipoláris skálával

A korrelációk alapján látható, hogy az egyes tengelyek elnevezése nem is olyan egyértelmű. Az első dimenzió legerősebben korrelál az USA-val való lojalitás változóval, innen kapta a nevét. Ezen kívül még erős korrelációt mutat az individualitás skálával. A második dimenzió a „gazdasági fejlettség"-gel azonosítható a korrelációk alapján. A harmadik dimenzió magas korrelációt (.80) mutat az országok népességének nagyságával. A negyedik dimenzió azokkal a skálákkal korrelál, amelyek az országok belső helyzetét, viszonyait írják le. A vietnami háborúra vonatkozó kérdéssel kapcsolatban a megkérdezetteket a válaszaik alapján három csoportba sorolták (1) „galambok" (G); (2) „mérsékeltek" (M); (3) „héják" (H). A megkérdezettek közül 18 dimenzió súlyait mutatja a következő táblázat, illetve ábra.

Megkérdezettek száma	1. dim.	2. dim.	R
1.	0,23	0,09	0,25
2.	0,22	0,52	0,57
3.	0,26	0,50	0,57
4.	0,26	0,59	0,65
5.	0,60	0,22	0,64
6.	0,36	0,56	0,68
7.	0,62	0,38	0,73
8.	0,50	0,42	0,66
9.	0,38	0,57	0,69
10.	0,33	0,70	0,78
11.	0,63	0,10	0,64
12.	0,47	0,43	0,64
13.	0,45	0,31	0,55
14.	0,41	0,08	0,42
15.	0,25	0,46	0,53
16.	0,52	0,34	0,63
17.	0,55	0,05	0,56
18.	0,53	0,39	0,67

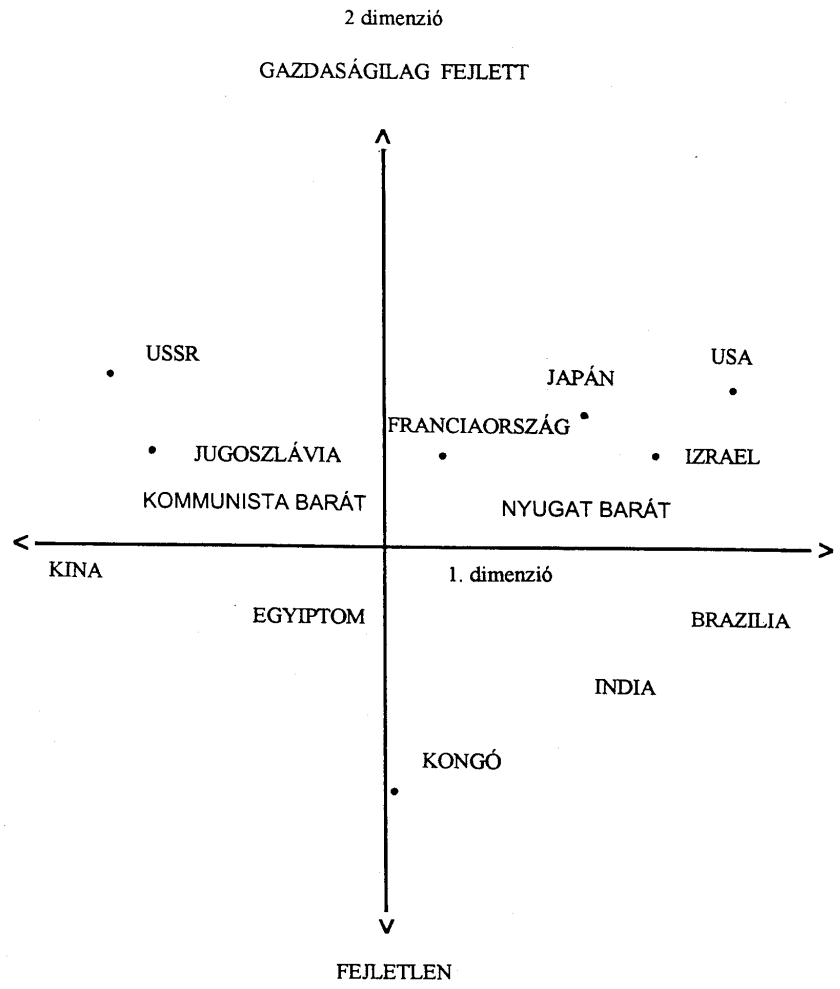
14.12. táblázat. 18 megkérdezett szubjektum terének koordinátái



14.19. ábra. 18 megkérdezett szubjektív tere

Az ábrában G, M, H betűkkel jelöltük a három csoporthoz való tartozást. Mint látható, a héják sokkal nagyobb súlyt adnak a politikai hovatartozásnak, mint a gazdasági fejlettségnek, a galambokra pedig ennek fordítottja igaz.

A következő két ábra a héják és a galambok világnézete közötti különbséget tükrözi.

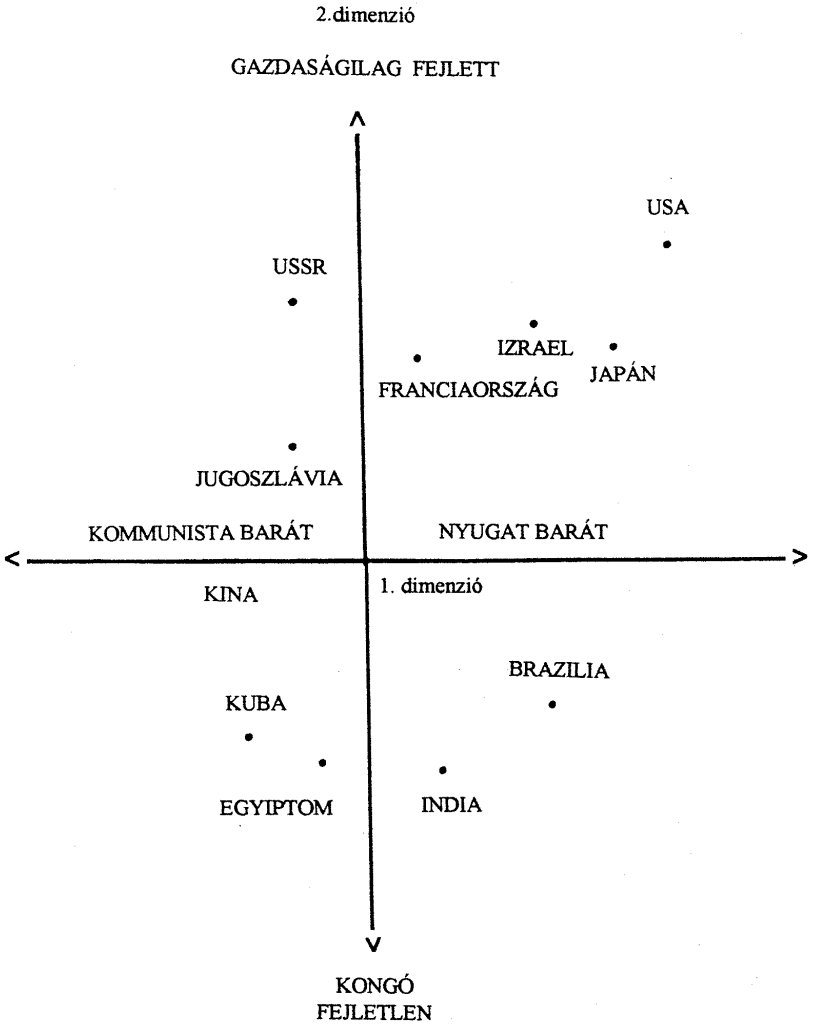


14.20. ábra. Nemzetek hasonlósága egy héja megítélésű személy szerint

A héják csoportjából származó egyén a vízszintes tengelyt megnyújtja, így messze tolja egymástól a kommunista és nem kommunista országokat, és majdnem teljesen elfedi a gazdasági fejlettség dimenziója szerint egyébként meglévő különbségeket. A 4. sorszámú „galamb" egyén terének függőleges nyújtása a gazdasági fejlettség dimenziójának ad nagyobb súlyt.

Ez a példa azt mutatja, hogy az INDSCAL-modell a többi három-utas modellhez hasonlóan hogyan alkalmazkodik az egyének közötti nagy különbséghez.

A társadalom értékstruktúrájának vizsgálata



14.21. ábra. Nemzetek hasonlósága a galamb megítélésű személy szerint

Az Életmód, Életminőség, Értékrendszer vizsgálat 1978-as kérdőíves felmérése 32 emberi alapérték választását is megkérdezte. Az 1500 vizsgálati személyből 13 szempont szerint társadalmi csoportokat képeztünk, és minden csoportban külön megnéztük, hogy az emberi alapértékek milyen típusokba rendeződnek. A klaszterelemzés hierarchikus struktúrájában az értékek különbözőségeit a kapcsolódásuk szintjeivel mértük a dendrogramban. Ezzel az alapértékek különbözőségi mátrixáig jutottunk, amelyek a 13 társadalmi csoportra egy három-utas mátrixot adtak, amely az INDSCAL-eljárás alapadatát képezte.

Társadalmi csoportok

1. Szegény 1 950,-Ft-ig
2. Szegény 2 750,-Ft-ig

3. Gazdag	3 800.-Ft-tól 9 999.-Ft-ig
4. Ateista	nem vallásos, vallásellenes
5. Vallásos	rendszeresen jár templomba
6. Fiatal	30 év alattiak
7. Delelő	40–49 évesek
8. Öreg	60 év felettiek
9. Férfi	
10. Nő	
11. Beteg 1	nincs visszatérő betegsége
12. Beteg 2	van visszatérő betegsége
13. Egészséges	teljesen egészséges

14.13. táblázat.

Az INDSCAL-modell eredményeit a következő táblázatok, illetve ábrák mutatják:

	1. dim.	2. dim.
1. Szegény 1	0,397	0,363
2. Szegény 2	0,171	0,303
3. Gazdag	0,185	0,214
4. Ateista	0,307	0,212
5. Vallásos	0,328	0,434
6. Fiatal	0,205	0,308
7. Delelő	0,179	0,282
8. Öreg	0,448	0,395
9. Férfi	0,305	0,329
10. Nő	0,597	0,307
11. Beteg 1	0,446	0,327
12. Beteg 2	0,199	0,229
13. Egészséges	0,289	0,321

14.14. táblázat. A társadalmi csoportok terének (subject space) koordinátái

Az alapértékek sorszáma	Dimenziók	
	1.	2.
1. emberek bizalma	−0,09900	0,03691
2. elismerés	−0,08844	−0,00089
3. alkotómunka	−0,26144	0,19112
4. gyerek, gyerekáldás	0,20567	0,09832
5. tiszta lelkiismeret	0,24333	0,15116
6. munkahelyi légkör	−0,07092	0,05644
7. családi boldogság	0,21158	0,21603
8. szerelem	−0,05011	0,18359
9. egészség	0,27831	0,12400
10. kényelmes lakás	0,21849	0,12976
11. vezető pozíció	0,02492	−0,23064
12. nyugalom, béke	0,21282	0,13810
13. művészi élmény	−0,01129	−0,30958
14. gondoktól mentes élet	0,23683	0,05769
15. tanulási lehetőség	−0,06339	0,06779
16. függetlenség	−0,20313	0,14870
17. eszmében való hit	−0,18725	−0,05649
18. haza, nemzet	−0,16583	0,06577
19. sport, mozgás	−0,01963	−0,35487
20. létbiztonság	0,25957	0,04522
21. belső harmónia	0,09774	0,01792
22. anyagi jólét	0,29884	0,07453
23. hely a társadalomban	−0,27739	0,19033
24. törvényesség	−0,25410	0,17510
25. kitartó munka	−0,17864	0,18251
26. sok szabadidő	0,03946	−0,30393
27. önbizalom	−0,06729	−0,17859
28. jó megjelenés	0,01561	−0,25961
29. jó értelem	−0,08503	0,03705
30. hasznosság	−0,26390	0,20105
31. származás	0,03179	−0,28379
32. siker	−0,02818	−0,2399

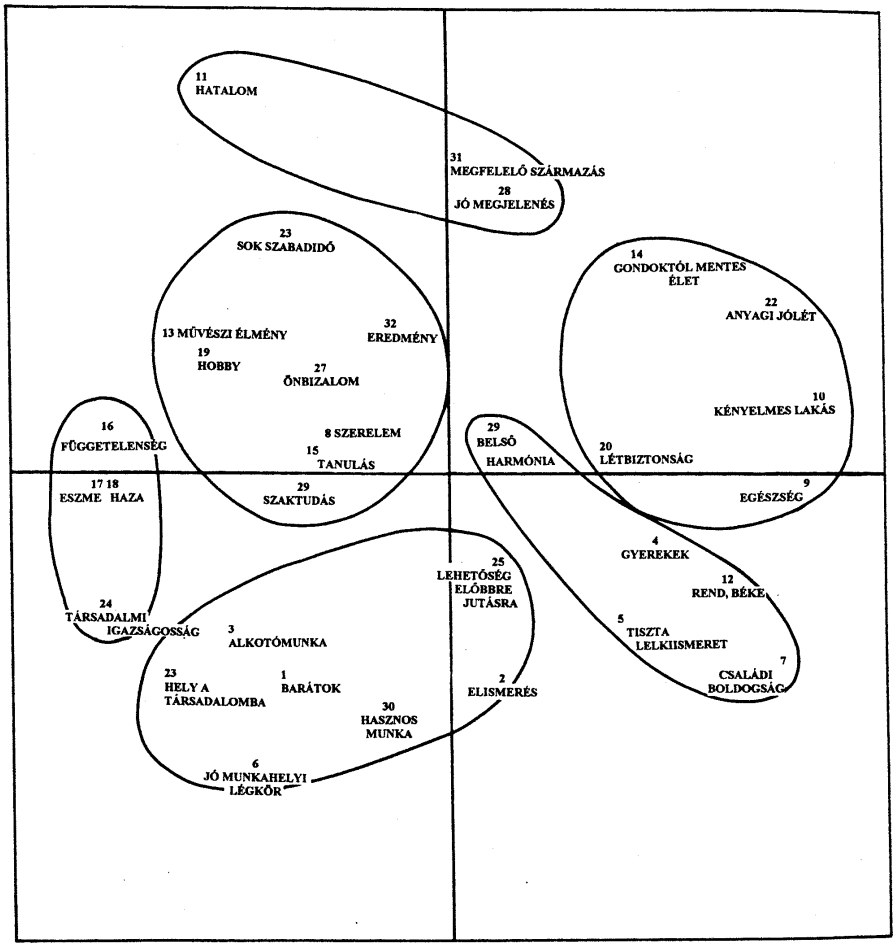
14.15. táblázat. Az alapértékek terének koordinátái (stimulus tér)

Az INDSCAL dimenzióit az egyes társadalmi csoportok eltérő módon súlyozzák. Így a társadalom értékterének dimenzióit „másképpen” megítélő csoportok az INDSCAL eredménye alapján

szegények; nők; delelők; fiatalok; vallásosak; betegek.

Az átlagos értékstruktúrához hasonló súlyozást az öregek; szegények; betegek társadalmi csoportja adja.

A többi csoport a két megítélés közé esik.



14.22. ábra. Az alapértékek tere a klaszter- és faktorelemzéssel kapott értéktípusok berajzolásával

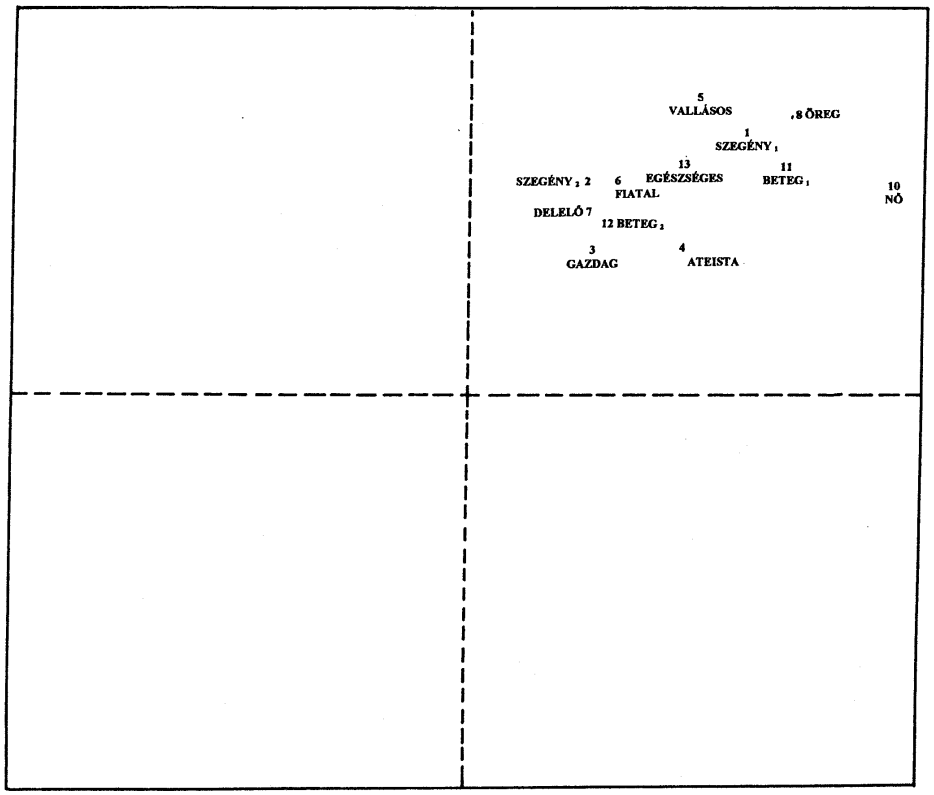
14.6. A PREFMAP-modell (PREFerence MAPing)

A PREFMAP-eljárás egyedeknek vagy ezek kategóriákba sorolt megfigyelési egységeinek keresi az „ideális” pontjait egy *a priori* térben a tér pontjaira vonatkozó preferencia értékek alapján.

A PREFMAP-modell a következő feltételezésekből indul ki:

– rendelkezésünkre állnak n egyednek (embernek, társadalmi csoportnak stb.) m változóra (jellemzőre, stimulusra) vonatkozó megfigyelési értékei, amelyek valamilyen preferencia skálán értelmezettek. Ezeket az adatokat a következő mátrix tartalmazza:

$$\mathbf{S} = \{s_{ij}\} \qquad i = 1, 2, \ldots, n; \quad j = 1, 2, \ldots, m$$



14.23. ábra. A társadalmi csoportok tere (subject space)

Az **S** mátrix i-edik sora az i-edik egyed preferencia értékeit tartalmazza. Feltételezzük, hogy ha

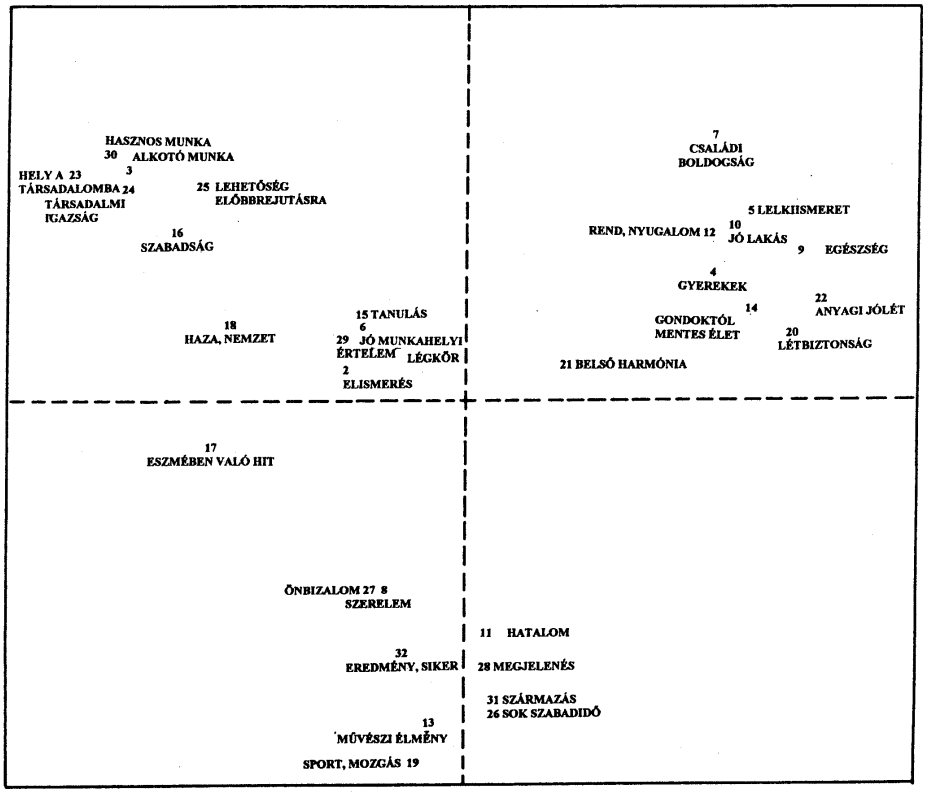
$$s_{ij} > s_{ik}$$

akkor az i-edik egyed a j-edik és k-adik stimulusok közül a j-edik stimulust preferálja.

– rendelkezésünkre áll a stimulusok (objektumok) konfigurációja az r-dimenziós térben. Az r-dimenziós tér koordinátáit *vagy* mérési eredményekből, ill. ezek aggregálásából nyerjük – ekkor a stimulusok megfigyelési vagy mintateréből indulunk ki – *vagy* más sokváltozós módszer eredményeként kapott ún. származtatott terében értelmezzük. Ilyen származtatott tér lehet pl. a faktortér, vagy a MINISSA-eljárás axiológiai tere.

Az adott *a priori* stimulus térbe a PREFMAP-eljárás 4 különböző modellel illeszti a stimulusokra vonatkozó preferenciákkal rendelkező egyedeket.

A következőkben sorra vesszük a modelleket.



14.24. ábra. Alapértékek tere (stimulus tér)

14.6.1. Az általános távolság modell (I. modell)

Ebben a modellben – és a következő két modellben is – feltesszük, hogy a stimulus tér pontjai és az ideális pont közötti távolság négyzete és a preferencia értékek között lineáris összefüggés van.

$$s_{ij} = a_i d_{ij}^2 + b_i + e_{ij}, \tag{14.33}$$

ahol

– a_i és b_i a lineáris függvény együtthatói ($a_i \geq 0$)

– e_{ij} hibatag.

Az I. modellben feltételezzük, hogy minden egyed a stimulus tér tengelyeit a T_i ortogonális transzformációs mátrix szerint transzformálja. Így:

$$\mathbf{x}_j^* = \mathbf{T}_i \, \mathbf{x}_j \tag{14.34}$$

$$\mathbf{y}_i^* = \mathbf{T}_i \, \mathbf{y}_i, \tag{14.35}$$

ahol

– $\mathbf{x}_j$ vektor a j-edik stimulus koordinátáit tartalmazza

– $\mathbf{y}_i$ vektor i-edik egyed ideális pontja.

A stimulusok és az ideális pont között a súlyozott távolságot a transzformált térben számítjuk:

$$d_{ij}^2 = \sum_{t=1}^{r} w_{it} \left(x_{jt}^* - y_{it}^* \right)^2 . \tag{14.36}$$

A (14.36) egyenletet mátrix formában a következőképpen írhatjuk:

$$d_{ij}^2 = \left(\mathbf{x}_j^* - \mathbf{y}_i^* \right)' \mathbf{W}_i \left(\mathbf{x}_j^* - \mathbf{y}_i^* \right),$$

ahol $\mathbf{W}_i$ diagonális mátrix, diagonális elemei a w_{it} súlyok.

Ezt az egyenletet kifejtve a következő formához jutunk:

$$d_{ij}^2 = \left(\mathbf{x}_j^* \right)' \mathbf{W}_i \mathbf{x}_j^* - 2 \left(\mathbf{y}_i^* \right)' \mathbf{W}_i \, \mathbf{x}_j^* + \left(\mathbf{y}_i^* \right)' \mathbf{W}_i \, \mathbf{y}_i^* . \tag{14.37}$$

Behelyettesítve most a (14.37) egyenletbe a (14.34) és (14.35) egyenleteket, a távolságot az eredeti tér koordinátáival számíthatjuk ki:

$$d_{ij}^2 = \mathbf{x}_j' \mathbf{T}_i' \mathbf{W}_i \mathbf{T}_i \mathbf{x}_j - 2\mathbf{y}_i' \mathbf{T}_i' \mathbf{W}_i \mathbf{T}_i \mathbf{x}_j + \mathbf{y}_i' \mathbf{T}_i' \mathbf{W}_i \mathbf{T}_i \mathbf{y}_i . \tag{14.38}$$

Legyen most

$$\mathbf{R}_i^* = \mathbf{T}_i' \mathbf{W}_i \mathbf{T}_i \tag{14.39}$$

és a (14.38) egyenlet jobb oldalának utolsó tagja $\mathbf{y}_i' \mathbf{T}_i' \mathbf{W}_i \mathbf{T}_i \mathbf{y}_i$ legyen c_i^* konstans, mivel nem függ a stimulusok koordinátáitól, $(\mathbf{x}_j)$-től.

Így a (14.38) egyenlet a következő egyszerűsített alakot ölti:

$$d_{ij}^2 = \mathbf{x}_j' \mathbf{R}_i^* \mathbf{x}_j - 2\mathbf{y}_i' \mathbf{R}_i^* \mathbf{x}_j + c_i^* \tag{14.40}$$

A (14.40) egyenletet helyettesítsük a (14.33) alapegyenletünkbe (a hibatagot elhagyjuk):

$$\begin{aligned} s_{ij} &\approx a_i \left[\mathbf{x}_j' \mathbf{R}_i^* \mathbf{x}_j - 2\mathbf{y}_i' \mathbf{R}_i^* \mathbf{x}_j + c_i^* \right] + b_i \\ &= \mathbf{a}_i \mathbf{x}_j' \mathbf{R}_i^* \mathbf{x}_j - 2a_i \mathbf{y}_j' \mathbf{R}_i^* \mathbf{x}_j + a_i c_i^* + b_i . \end{aligned} \tag{14.41}$$

Tovább egyszerűsítsük a jelöléseket:

$$\mathbf{R}_i = a_i \mathbf{R}_i^* \tag{14.42}$$

$$\mathbf{b}_i' = -2a\mathbf{y}_i' \mathbf{R}_i^* = -2\mathbf{y}_i' \mathbf{R}_i \tag{14.43}$$

és

$$c_i = a_i c_i^* + b_i . \tag{14.44}$$

Ennek alapján:

$$s_{ij} \approx \mathbf{x}_j' \mathbf{R}_i \mathbf{x}_j + \mathbf{b}_i' \mathbf{x}_j + c_i . \tag{14.45}$$

A (14.45) egyenlet a stimulus koordinátáinak (x_{jt}) másodfokú függvénye.

Skalároritmetikai jelölésekkel a (14.45) egyenlet:

$$s_{ij} \approx \sum_{t=1}^{r} \sum_{t'=1}^{r} r_{tt_i'} \left(x_{jt} x_{jt'} \right) + \sum_{t=1}^{r} b_{it} x_{jt} + c_i . \tag{14.46}$$

A fenti egyenlet az s_{ij} és x_{jt} változók között egy kvadratikus regressziós függvényt definiál. A másodfokú regressziós egyenlet ismeretlen együtthatóira becslést kaphatunk a legegyszerűbb módon úgy, hogy a magyarázó $x_{j1}, \; x_{j2}, \ldots, \; x_{jr}$ változók mellé

bevezetjük a dummy $x_t \times x_t$ változókat, így a (14.46) függvényt visszavezetjük egy többszörös lineáris regressziós függvényre, amelynek paraméterbecslésére jól ismert eljárások vannak.

A lineáris regressziós probléma megoldásával jutunk a c_i, b_{i1}, b_{i2}, ..., b_{ir}, $r_{11,i}$, $r_{22,i}$, ..., $r_{rr,i}$, $r_{12,i}$, ... együtthatók, vagyis $\mathbf{R}_i$ és $\mathbf{b}_i$ elemeinek becsléséhez. A becslésekből a (14.43) egyenlet alapján az i-edik egyed „ideális" pontját számítjuk ki:

$$\mathbf{y}_i' = -\frac{1}{2}\mathbf{b}_i'\mathbf{R}_i^{-1}. \tag{14.47}$$

Az $\mathbf{R}_i$ mátrix ismeretében a $\mathbf{T}_i$ transzformációs mátrix és a $\mathbf{w}_{it}$ súlyokat tartalmazó diagonális $\mathbf{W}_i$ mátrix pedig a sajátértékeket tartalmazza (a (14.39) egyenlet alapján). A sajátértékek (súlyok) nemnegatívak, ha $\mathbf{R_i}$ pozitív definit vagy szemidefinit. A gyakorlatban ez nem mindig áll fenn, így előfordulhatnak negatív súlyok is, ami komoly értelmezési problémákat jelenthet. Egy esetben könnyen értelmezhetjük a negatív súlyokat. Ha i-edik egyed negatív w_{it} súlya azt jelenti, hogy az ideális pont y_{it} a legkevésbé preferált értéket jelöli, és minél tovább haladunk ezen a tengelyen, annál preferáltabb értéket találunk.

Összefoglalva az I. modellt:

1. Minden egyed a stimulus tér koordinátáit egyedi rotációval transzformálhatja a saját preferencia dimenzióihoz.
2. Minden egyed különbözőképpen súlyozhatja saját dimenzióit.
3. Minden egyedet a stimulus tér legjobban preferált helyén, mint a saját „ideális" pontját ábrázoljuk.

14.6.2. A súlyozott távolság modell (II. modell)

A II. modell annyiban különbözik az I. modelltől, hogy nem engedjük meg az egyedek különböző ortogonális transzformációját, de azt megengedjük, hogy a stimulus tér dimenzióit különbözőképpen súlyozzák. A II. modell speciális esete az I. modellnek, mivel a $\mathbf{T}_i$ transzformációs mátrix minden egyednél azonos egységmátrix $\mathbf{T}_i = \mathbf{I}$.

A kiinduló egyenletünk azonos az I. modellével:

$$s_{ij} = a_i\, d_{ij}^2 + b_i + e_{ij}$$

A távolságot a II. modellben a kiindulási stimulus térben számítjuk:

$$d_{ij} = \sum_{t=1}^{r} w_{it}\left(x_{jt} - y_{it}\right)^2. \tag{14.48}$$

Az I. modell (14.45) egyenlete most a következő:

$$s_{ij} \approx \mathbf{x}_j'\mathbf{W}_i\mathbf{x}_j + \mathbf{b}_i'\mathbf{x}_j + c_i \tag{14.49}$$

vagy más formában:

$$s_{ij} \approx \sum_{t=1}^{r} w_{it}x_{jt}^2 + \sum_{t=1}^{r} b_{it}x_{jt} + c_i \tag{14.50}$$

ahol a (14.43) egyenletnek megfelelő együtthatók:

$$\mathbf{b}_i' = -2\mathbf{y}_i'\,\mathbf{W}_i, \tag{14.51}$$

amiből:

$$\mathbf{y}_i' = -\frac{1}{2}\mathbf{b}_i'\mathbf{W}_i^{-1}. \tag{14.52}$$

A (14.52) egyenlet jobb oldalán $\mathbf{W}_i$ diagonális mátrix, aminek inverze egyszerűen számítható úgy, hogy a w_{it} diagonális elemek reciprokát vesszük.

A (14.50) egyenletben az adott s_{ij} stimulus koordináták ismeretében a másodfokú regresszió módszerével becslést kaphatunk a $c_i,\ b_{i1},\ b_{i2}, \ldots, b_{ir}, w_{i1}, \ldots, w_{ir}$ együtthatókra.

Ezekből azután az ideális pont y_i koordinátáit kiszámíthatjuk:

$$y_{it} = -\frac{1}{2}\ \frac{b_{it}}{w_{it}}$$

Összefoglalva a II. modellt:

1. Minden egyed ideális preferencia pontját ugyanazon preferencia tengelyekhez illesztjük.
2. Minden egyed eltérő súlyokat rendelhet a közös dimenziókhoz.
3. Minden egyed ideális pontját a stimulus térben ábrázoljuk.

14.6.3A nem súlyozott távolság modellek (III. modell)

Ebben a modellben is – mint a másik három PREFMAP-modellben – keressük az egyed ideális pontját a stimulusok által meghatározott térben. Az i-edik egyed ideális pontja (y_i) és a j-edik stimulus közötti távolság négyzete $\left(d_{ij}^2\right)$ és az i-edik egyed j-edik stimulusra vontkozó preferencia értéke között a modell értelmezése szerint lineáris összefüggés van.

$$s_{ij} = a_i\ d_{ij}^2 + b_i + e_{ij}. \tag{14.53}$$

A PREFMAP I., II. és III. modellje között a különbség a távolság d_{ij} eltérő meghatározásában van. A III. modellben a d_{ij} távolságot a következőképpen értelmezzük:

$$d_{ij}^2 = \sum_{t=1}^{r} u_t \left(x_{jt} - y_{it}\right)^2, \tag{14.54}$$

ahol $u_t = \pm 1$.

A III. modell lehetőséget ad arra, hogy egy-egy, vagy akár az összes tengelynek negatív súlya legyen. Azonban ezek a súlyok minden egyedre vonatkozóan azonosak. A preferencia értékeket becslő regressziós egyenlet a III. modellben:

$$s_{ij} \approx a_i'\mathbf{x}_j'\mathbf{U}\mathbf{x}_j + \mathbf{b}_i'\mathbf{x}_j + c_i \tag{14.55}$$

vagy más formában:

$$s_{ij} \approx a_i\left[\sum_{t=1}^{r} u_t x_{jt}^2\right] + \sum_{t=1}^{r} b_{it}x_{jt} + c_i. \tag{14.56}$$

Mivel a szögletes zárójelben lévő kifejezés független i-től, összesen $r+1$ független változónk van. A legkisebb négyzetek módszerével a többszörös regressziós egyenletet megoldva, a $c_i, b_{i1}, b_{i2}, \ldots, b_{ir}, a_i$ együtthatók becsléséhez jutunk. A becsült együttha-

tókból az „ideális pont” koordinátáit a következőképpen számíthatjuk:

$$y_{it} = -\frac{1}{2}\,\frac{b_{it}}{a_i u_t}. \tag{14.57}$$

Az a_i együttható előjelétől függően az egyed ideális pontja lehet maximum vagy minimum megoldás.

Ha az egyik dimenziónak a súlya negatív $u_t = -1$ és a többi súly pozitív, akkor az ideális pontokról azt mondjuk, hogy nyeregpontok.

Összefoglalva a III. modellt:

1. Nem engedjük meg a tengelyek rotálását.
2. Nem engedjük meg a tengelyek különböző súlyozását, de megengedjük, hogy egyes dimenzióknak negatív súlya legyen.
3. Minden egyed ideális pontját a stimulus térben ábrázoljuk.

14.6.4. A vektor modell (IV. modell)

A vektor modellben az egyedek preferencia értékeit a stimulusok koordinátáinak lineáris függvényével becsüljük:

$$s_{ij} \approx \mathbf{b}_i'\mathbf{x}_j + c_i. \tag{14.58}$$

Más formában:

$$s_{ij} \approx \sum_{t=1}^{r} b_{it}x_{jt} + c_i. \tag{14.59}$$

A IV. modell regressziós függvénye csak lineáris tagokat tartalmaz. A (14.58), ill. (14.59) egyenlet speciális esete az általános modellnek. Legyen $\mathbf{b}_i' = a_i\mathbf{y}_i'$, a (14.58) egyenlet a következő lesz:

$$s_{ij} \approx a_i\mathbf{y}_i'\mathbf{x}_j + c_i. \tag{14.60}$$

Ez pedig a (14.55) egyenlet speciális esete, melyben a kvadratikus tag együtthatója nullával egyenlő.

A b_{it} súlyokat a többszörös lineáris regresszió módszerével becsüljük.

A becslésekből azután kiszámíthatjuk az „ideális pont” koordinátáit:

$$y_{it} = \frac{b_{it}}{\sqrt{\sum_{t=1}^{r} b_{it}^2}} \tag{14.61}$$

ami a regressziós együtthatók normalizált alakja. Mivel

$$\sum_{t=1}^{r} y_{it}^2 = 1,$$

minden egyed a neki megfelelő y_{it} koordinátákkal ábrázolva az origó körüli egységsugarú kör kerületén helyezkedik el.

14.6.5. A nem metrikus PREFMAP-modell

Egészen idáig feltételeztük, hogy a preferencia értékeket intervallummérési szinten mértük. Kerestük a stimulus pontok és az ideális pont távolságának négyzete és a preferencia értékek között a legkisebb négyzetek módszere értelmében legjobb lineáris függvényt ($s_{ij} = F_i\left(d_{ij}^2\right)$).

A gyakorlati vizsgálatokban a preferencia értékek nagyon sokszor csak mint rangszámok értelmezhetők. Kiinduló egyenletünket a következővel helyettesítjük:

$$\widetilde{s}_{ij} = d_{ij}^2 + e_{ij} \tag{14.62}$$

ahol $\widetilde{s}_{ij} = M_i^{(\ )}\left(s_{ij}\right)$ és M_i monoton nemcsökkenő függvény.

A nem metrikus PREFMAP-modellt iterációs eljárással oldjuk meg.

1. lépés: A preferencia értékek $\left(s_{ij}\right)$ becslését először regressziós egyenlettel (másodfokú vagy lineáris) számítjuk ki. Ez a lépés megegyezik a metrikus eljárással. A becsült értékeket $\widetilde{s}_{ij}^{(1)}$-vel jelöljük.

2. lépés: A Kruskal-féle monoton regressziós eljárással az i-edik egyed $M_i^{(1)}$ monoton függvényét becsüljük, amellyel az eredeti s_{ij} értékekből a legjobban becsüljük a $\widetilde{s}_{ij}^{(1)}$-ket:

$$\widehat{s}_{ij}^{(1)} = M_i^{(1)}\left(s_{ij}\right).$$

3. lépés: Az s_{ij}-ket helyettesítjük az $\widetilde{s}_{ij}^{(1)}$ értékekkel és kiszámítjuk az $\widetilde{s}_{ij}^{(2)}$ új becsléseket.

4. lépés: Az új $\widetilde{s}_{ij}$ értékek alapján az új monoton függvényt $M_i^{(2)}$, és az $\widehat{s}_{ij}$ új értékeit $\widehat{s}_{ij}^{(2)}$ számítjuk.

5. lépés: Az előző lépéseket addig folytatjuk, amíg az eljárás konvergens nem lesz, amíg a monoton függvényben vagy a regressziós együtthatókban nincs további változás. A programban a CRITERION-paraméterrel lehet beállítani az $\widehat{s}_{ij}$ értékek egymásutáni eltérései négyzetösszegére azt az értéket, amelynél kisebb változás esetén az eljárás befejeződik.

14.6.6A z illeszkedés jóságának vizsgálata

A PREFMAP-modellek az egyedek preferencia értékeit lineáris (vagy másodfokú) regressziós függvénnyel becsülik. A becslés jóságának vizsgálatához a regressziónál is alkalmazott többszörös korrelációs együtthatót használjuk, amely nem más, mint az eredeti preferencia értékek és a többszörös regressziós függvény becsült értékei között számított zéró rendű korrelációs együttható.

A többszörös korreláció szignifikanciáját F hányadossal vizsgálhatjuk.

Az F hányados ismert formulája:

$$F = \frac{R^2/(k-1)}{\left(1-R^2\right)/(n-k)}, \tag{14.63}$$

ahol: R a többszörös korrelációs együttható; n a stimulus pontok száma; k a becsült regressziós együtthatók száma; $(k-1)$ és $(n-k)$ a szabadságfokok.

A PREFMAP-eljárás négy modellje különböző feltételeket tartalmaz az egyedek stimulus térbe illesztéséről. Ezek közül a legáltalánosabb az I. modell, és a legegyszerűbb a IV. modell, közöttük pedig az általánosítás szerinti hierarchia van. Általában a komplexebb modellben a többszörös korreláció értékei magasabbak, mint a kevésbé általános modellben. Kérdés, hogy az illeszkedés javulása szignifikánsnak tekinthető-e. Ennek megválaszolására az F hányadost használhatjuk. Az F hányados értéke a modellek között:

$$F_{ab} = \frac{\left(R_a^2 - R_b^2\right) / (k_a - k_b)}{\left(1 - R_a^2\right) / (n - k_b)}, \tag{14.64}$$

ahol:

– a és b az összehasonlított két modell (az a az általánosabb), R_a és R_b a két többszörös korrelációs együttható;

– k_a és k_b a becsült együtthatók száma;

– $(k_a - k_b)$ és $(n - k_b)$ a két szabadságfok.

Ha az egyedek nem túl eltérőek, és értelmes összehasonlítani az eredményüket, egy-egy modell esetén az illeszkedés jóságára még egy mutatót számíthatunk; az egyedek többszörös korrelációinak négyzetes (kvadratikus) átlagát (root mean square):

$$RMS_a = \frac{1}{n}\left(\sum_{i=1}^{n} R_{ia}^2\right)^{\frac{1}{2}}.$$

14.6.7. Példa a PREFMAP-eljárásra (Rokeach-értékek a magyar és az amerikai társadalomban)

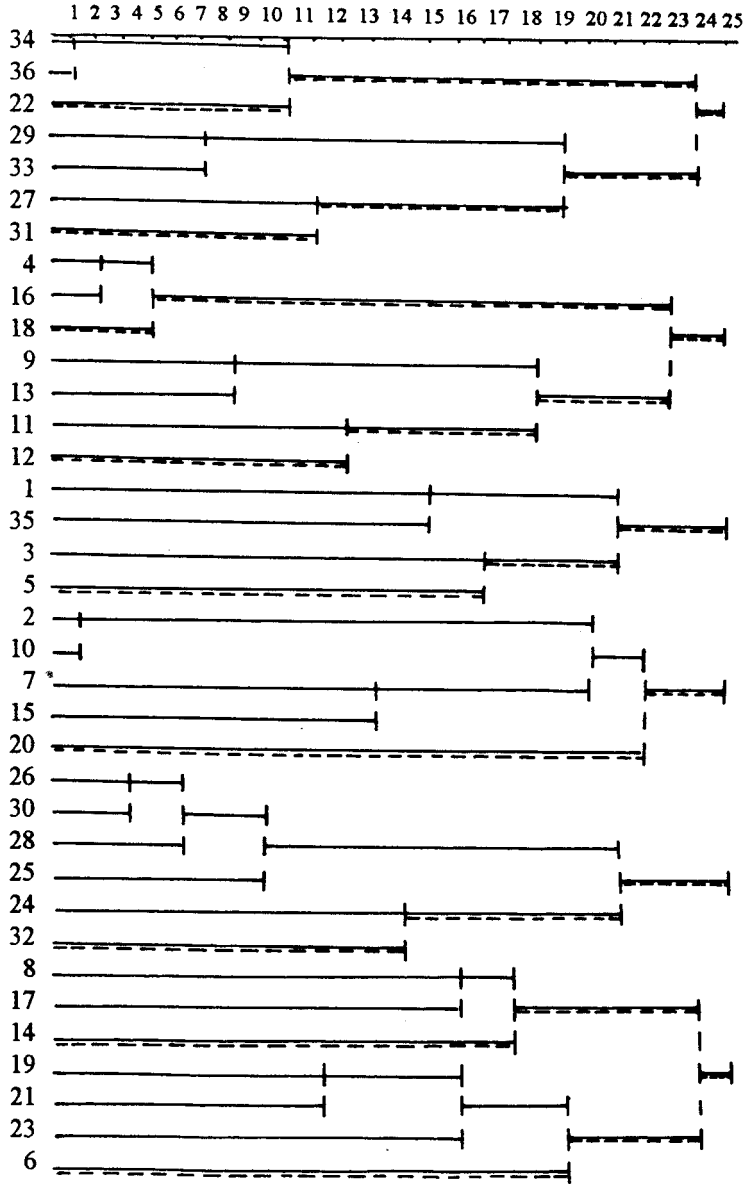
Milton Rokeach az amerikai nemzeti mintán (1409 felnőtt, húsz év feletti lakos) 36 eszköz- és célérték struktúráját vizsgálta 1967–68-ban. A 18 cél- és 18 eszközérték vizsgálatát az Életmód, Életminőség és Értékrendszer vizsgálat (Hankiss E., Manchin R., Füstös L. 1978) megismételte a magyar országos reprezentatív mintán (808 fő).

Először a magyar minta eredményeit közöljük.

Sorrendben: a faktorelemzés, a klaszterelemzés (fürtelemzés), a legkisebb térelemzés (MINISSA), és a preferencia illesztés (PREFMAP) eredményeit.

1. faktor	udvarias	0,68
variancia:	megbocsátó	0,67
17,75	tiszta	0,65
	engedelmes	0,64
	üdvözülés	0,62
	szeretettel teljes	0,53
	szépség világa	0,51
	bölcsesség	0,41
	jókedélyű	0,41
	kellemes élet	0,38
	igazi barátság	0,38
	érdekes élet	0,35
	igazi szerelem	0,34
2. faktor	alkotó szellemű	0,67
variancia:	értelmes	0,61
8,28	előítéletmentes	0,58
	logikus	0,49
	bölcsesség	0,48
	érdekes élet	0,45
	szépség világa	0,44
	belső harmónia	0,39
	igazi szerelem	0,38
	igazi barátság	0,38
3. faktor	béke	0,78
variancia:	haza biztonsága	0,74
5,88	szabadság	0,54
	egyenlőség	0,30
4. faktor	felelősségteljes	0,74
variancia:	hatékony	0,62
5,10	önálló	0,62
	fegyelmezett	0,61
	szavahihető	0,55
	logikus	0,55
	segítőkész	0,39
5. faktor	anyagi jólét	0,72
variancia:	törekvő	0,64
3,84	kellemes élet	0,54
	boldogság	0,37
	érdekes élet	0,31

14.16. táblázat. Rokeach-értékek faktorstruktúrája a magyar mintában



14.25. ábra. A Rokeach-értékek hierarchikus kapcsolódása (WARD-módszer) magyar mintában

A 14.25. ábrában szereplő értékek és sorszámaik egymáshoz rendelése.

tiszta	34	béke	2
udvarias	36	haza biztonsága	10
engedelmes	22	egyenlőség	7
megbocsátó	29	szabadság	15
szeretetteljes	33	bátor, gerinces	20
jókedélyű	27	hatékony	26
segítőkész	31	önálló	30
bölcsesség	4	logikus gondolkodású	28
szépség világa	16	felelősségteljes	25
üdvözülés	18	fegyelmezett	25
érdekes élet	9	szavahihető	32
kellemes élet	13	elvégzett munka	8
igazi barátság	11	társadalmi megbecsülés	17
igazi szerelem	12	emberi önérzet	14
anyagi jólét	1	alkotó szellemű	19
törekvő	35	előítéletektől mentes	21
boldogság	3	értelmes	23
családi biztonság	5	belső harmónia	6

A 36 emberi érték struktúrájának statisztikai elemzése, a faktorelemzés, klaszterelemzés és a legkisebb térelemzés eredményeinek összevetésével a következő összetartozó értékcsoportok különíthetők el:

A haza biztonsága, béke, szabadság, egyenlőség, bátorság
B anyagi jólét, boldoság, törekvő, családi boldogság
C/1 engedelmes, tiszta, udvarias
C/2 érdekes élet, barátság, szerelem, kellemes élet
C/3 bölcsesség, a szépség világa, üdvözülés
C/4 jókedély, megbocsátó, segítőkész
D fegyelmezett, felelősségteljes, hatékony, logikus, önálló, szavahihető
E/1 munka öröme, önérzet, megbecsülés
E/2 alkotó szellemű, előítéletektől mentes, értelmes

A fenti érték-klaszterek viszonya is látható az értékeknek a MINISSA-eljárás eredményeként kapott axiológiai terében. A PREFMAP-eljárással illesztjük ebbe a térbe az egyes társadalmi rétegeket az értékválasztásuk preferenciái alapján. A következő rétegképző ismérveket és rétegeket vizsgáltuk:

Település

1. Falu
2. Kisváros
3. Nagyváros
4. Budapest

Életkor

5. 20–29 éves
6. 30–39 éves
7. 40–49 éves
8. 50–59 éves
9. 60–69 éves
10. 70 év feletti

Iskolai végzettség (években)

11. 4 év vagy kevesebb
12. 5–8 év
13. 9–11 év
14. 12 év
15. 13–15 év
16. 16–17 év
17. 18 év vagy több

Jövedelem

18. 1000 Ft vagy kevesebb
19. 1001–2000 Ft
20. 2001–2500 Ft
21. 2501–3000 Ft
22. 3001–4000 Ft
23. 4001–6000 Ft
24. 6001 Ft vagy több.

Nem

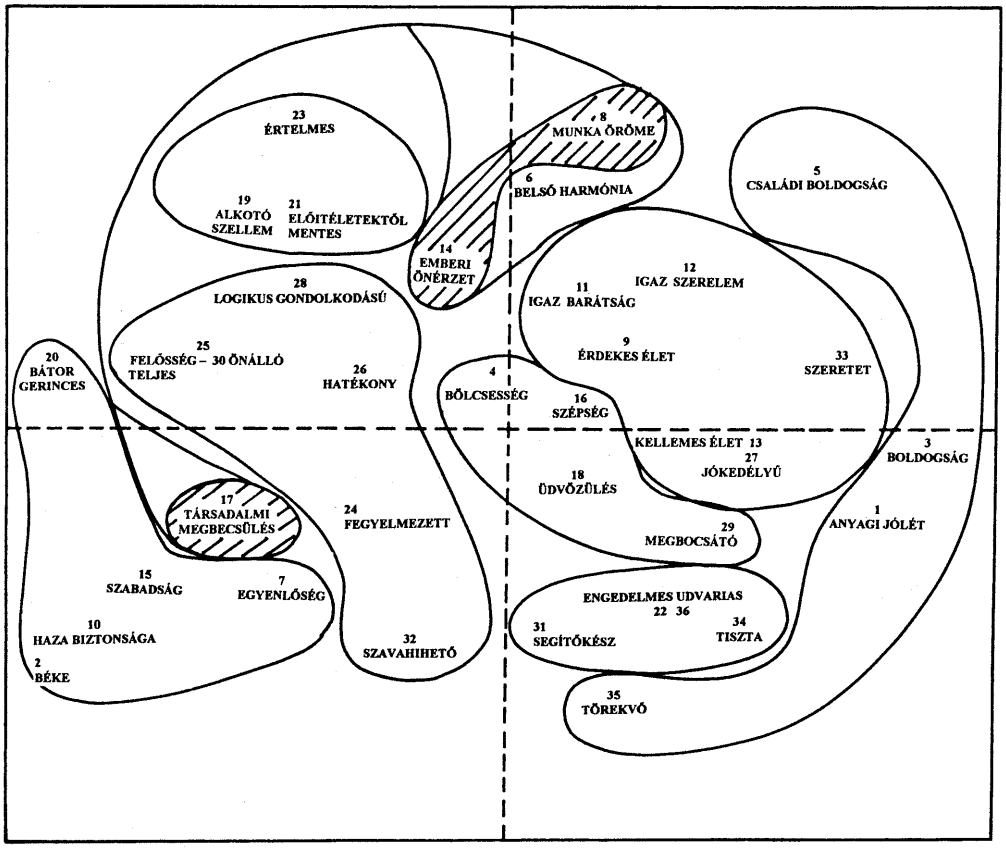
25. férfi
26. nő

A PREFMAP-eljárással kapott eredményeket a következő táblázatokban közöljük.

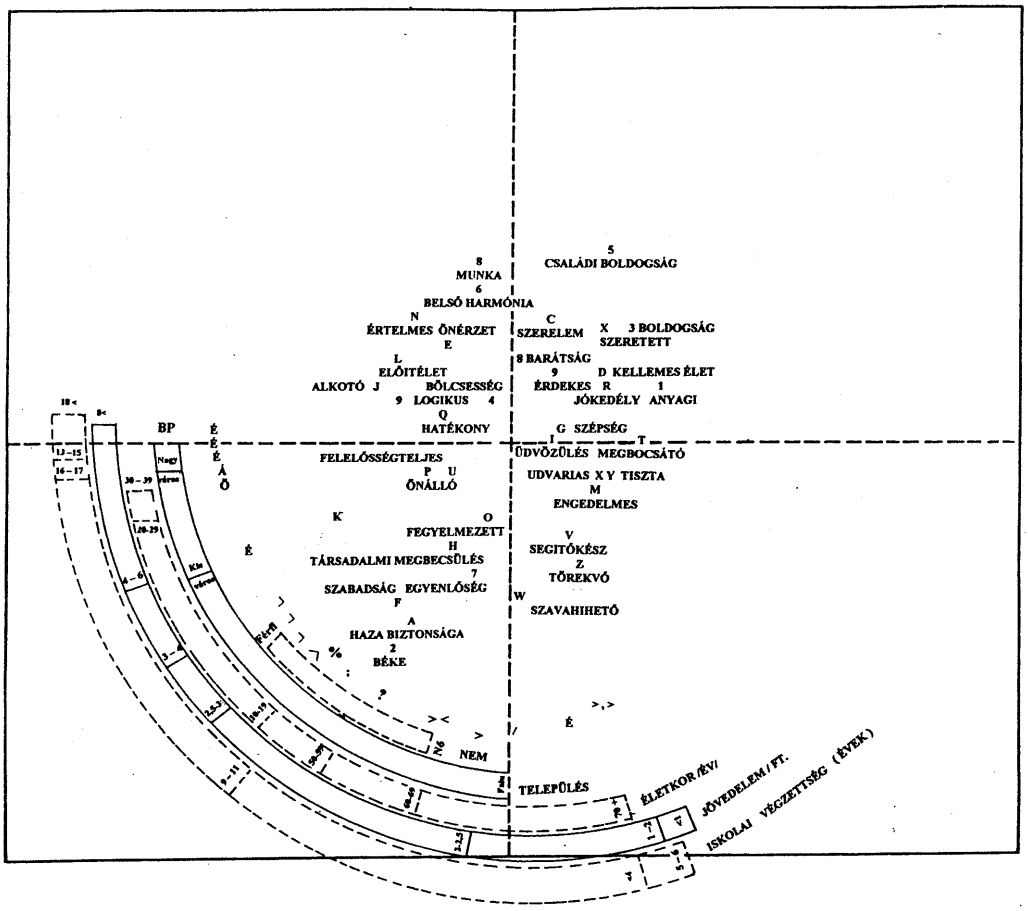
A társadalmi rétegek értékválasztásának különbségei a PREFMAP-eljárás eredményeként minden olvasó számára nyilvánvalóak. A különbségek igazolására a társadalmi rétegek között euklideszi távolságokat számítottunk, és a társadalmi rétegeket MINISSA-eljárással kétdimenziós ábrába vetítettük. Ezt mutatja a 14.28. ábra.

Falu	1	0,0062	−1,000
Kisváros	2	−0,9348	−0,3551
Nagyváros	3	−0,9944	−0,1056
Budapest	4	−0,9970	−0,0772
20–29 éves	5	−0,9806	−0,1962
30–39 éves	6	−0,9922	−0,1250
40–49 éves	7	−0,7205	−0,6935
50–59 éves	8	−0,5294	−0,8484
50–69 éves	9	−0,2419	−0,9703
70 év feletti	10	−0,3655	−0,9308
4 év iskola	11	0,3467	−0,9380
5–8 év	12	0,4617	−0,8871
9–11 év	13	−0,6698	−0,7425
12 év	14	−1,0000	−0,0066
13–15 év	15	−0,9993	−0,0379
16–17 év	16	−0,9954	−0,0955
18 év vagy több	17	−0,9997	0,0238
1000 Ft vagy kevesebb	18	0,4780	−0,8784
1001–2000 Ft	19	0,4398	−0,8981
2001–2500 Ft	20	−0,0747	−0,9972
2501–3000 Ft	21	−0,7365	−0,6765
3001–4000 Ft	22	−0,8643	−0,5030
4001–6000 Ft	23	−0,9430	−0,3328
6000 Ft felett	24	−0,9999	0,0168
Férfi	25	−0,7747	−0,6324
Nő	26	−0,2633	−0,9647
ÁTLAG	AVR	−0,8358	−0,5490

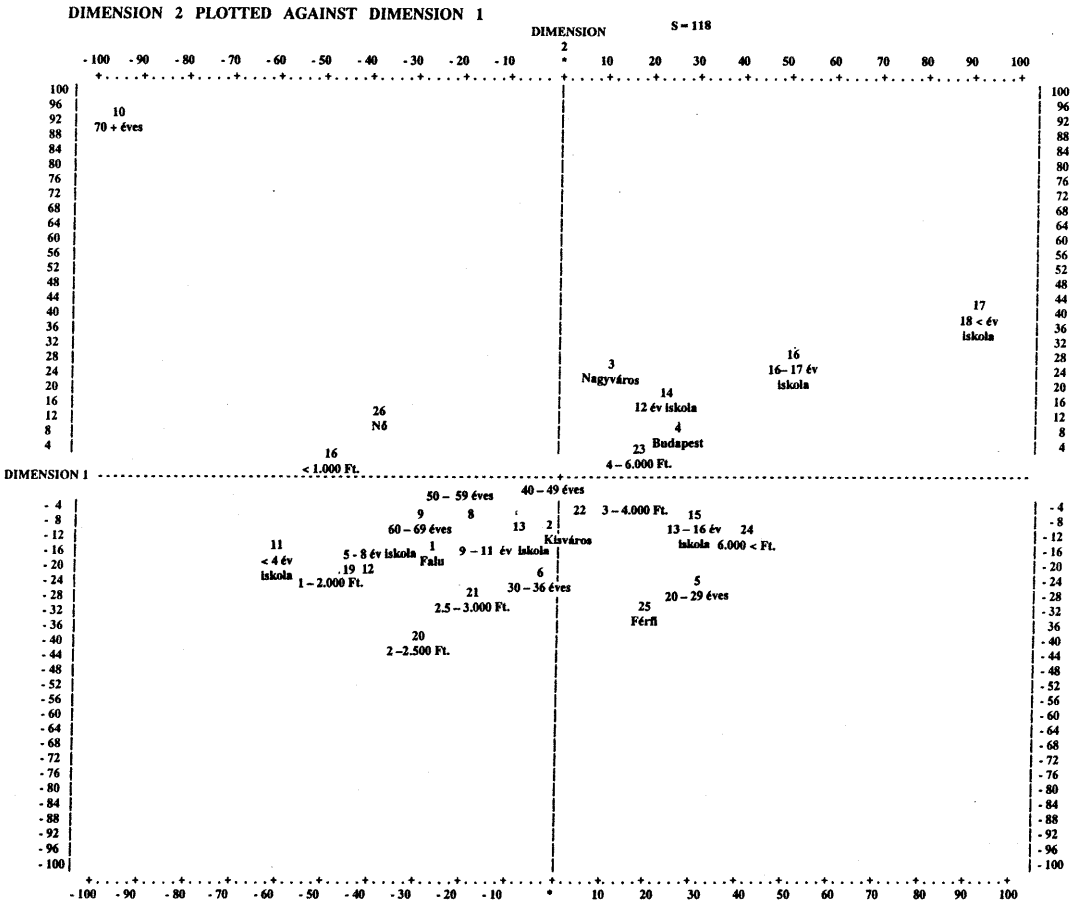
14.17. táblázat. A társadalmi rétegek iránykoszinuszai a Rokeach-értékek axiológiai terében (magyar adatok)



14.26. ábra. Rokeach-értékek axiológiai tere (MINISSA-eredmények, magyar adatok).
A csoportok kialakítása a faktor- és klaszterelemzés alapján.



14.27. ábra. A Rokeach-értékek axiológiai terébe illesztett társadalmi rétegek (magyar adatok)

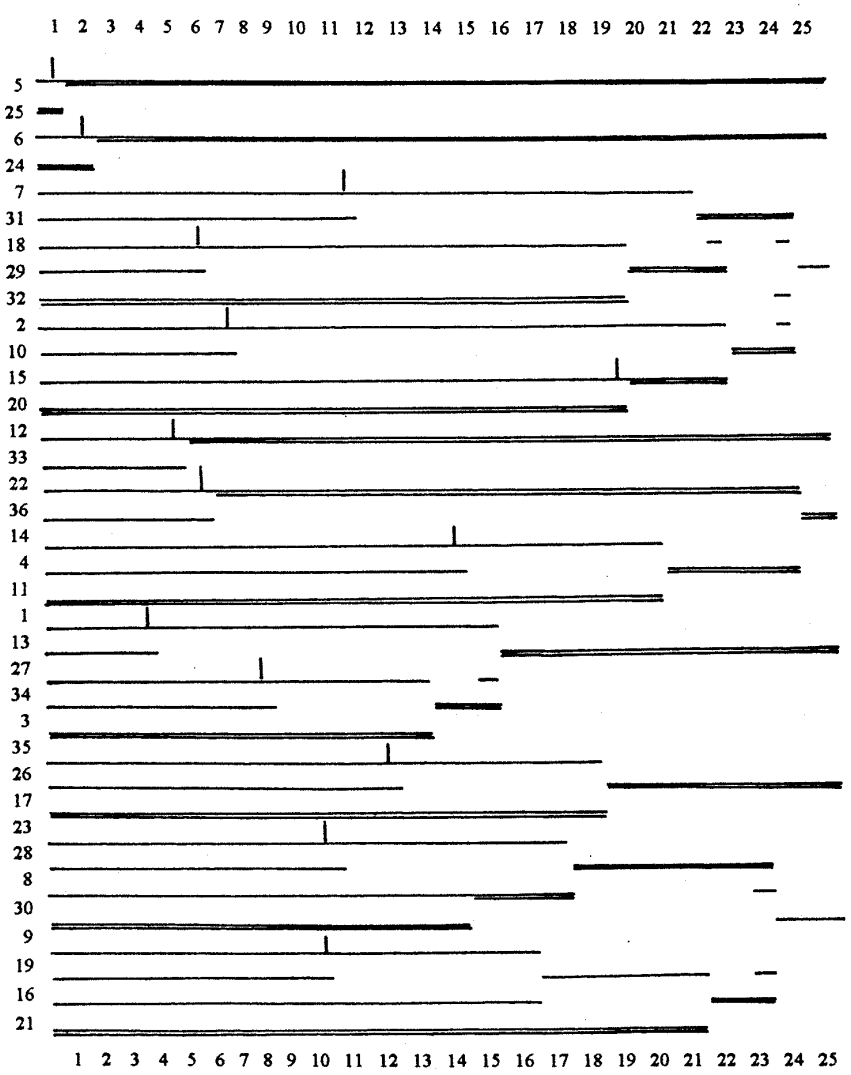


14.28. ábra. MINISSA-megoldás (magyar adatok). Társadalmi csoportok az értékek terében.

A következő táblázatokban az amerikai minta eredményei találhatók, amelyeket Milton Rokeach *The Nature of Human Values* c. könyvében közölt korrelációs mátrix és faktormátrix felhasználásával saját számításainkkal származtattunk.

1. faktor variancia: 8,2	Comfortable life (Anyagi jólét)	0,69	Wisdom (Bölcsesség)	−0,56
	Pleasure (Kellemes élet)	0,62	Inner harmony (Belső harmónia)	−0,41
	Clean (Tiszta)	0,47	Logical (Logikus)	−0,34
	An exciting life (Érdekes élet)	0,41	Self-controlled (Fegyelmezett)	−0,33
2. faktor variancia: 7,8	Logical (Logikus)	0,53	Forgiving (Megbocsátó)	−0,64
	Imaginative (Alkotó szellemű)	0,45	Salvation (Üdvözülés)	−0,56
	Intellectual (Intelligens)	0,44	Helpful (Segítőkész)	−0,39
	Independent (Önálló)	0,43	Clean (Tiszta)	−0,34
3. faktor variancia: 5,5	Obedient (Engedelmes)	0,52	Broadminded (Előítéletektől mentes)	−0,56
	Polite (Udvarias)	0,50	Capable (Hatékony)	−0,51
	Self-controlled (Fegyelmezett)	0,37		
	Honest (Becsületes)	0,37		
4. faktor variancia: 5,4	A world at peace (Béke)	0,61	True friendship (Barátság)	−0,49
	National security (Haza biztonsága)	0,58	Self-respect (Önérzet)	−0,48
	Freedom (Szabadság)	0,40		
5. faktor variancia: 5,0	A world of beauty (Szépség világa)	0,58	Family security (Családi biztonság)	−0,50
	Equality (Egyenlőség)	0,39	Ambitious (Törekvő)	−0,43
	Helpful (Segítőkész)	0,36	Responsible (Felelősségteljes)	−0,33
	Imaginative (Alkotó szellemű)	0,30	Capable (Hatékony)	−0,32
6. faktor variancia: 4,9	Social recognition (Társadalmi megbecsülés)	0,49	Mature love (Szerelem)	−0,68
	Self-respect (Emberi önérzet)	0,32	Loving (Szeretettel teljes)	−0,60
7. faktor variancia: 4,0	Polite (Udvarias)	0,34	Courageous (Bátor)	−0,70
			Independent (Önálló)	−0,33

14.18. táblázat. Rokeach-értékek faktorstruktúrája az amerikai mintában



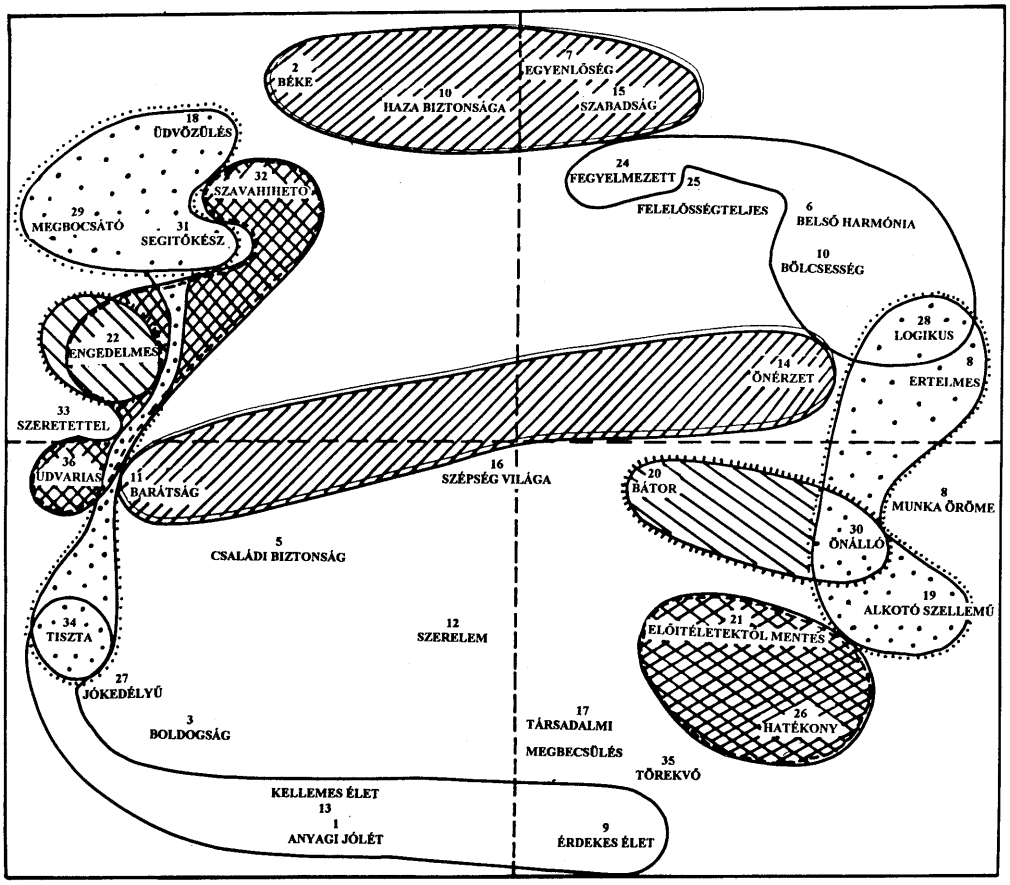
14.29. ábra. A Rokeach-értékek hierarchikus kapcsolódásai (WARD-módszer) az amerikai mintában

A 14.29. ábrában szereplő értékek és sorszámaik egymáshoz rendelése.

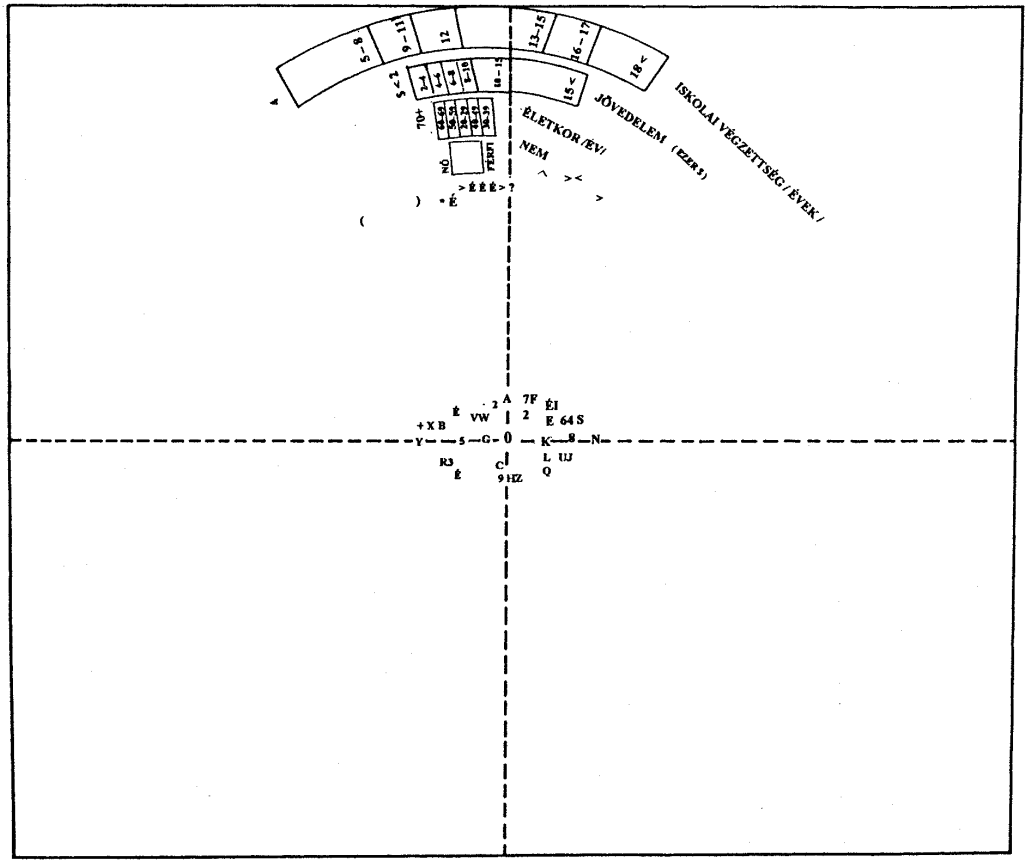
családi boldogság	5	bölcsesség	4
felelősségteljes	25	igazi barátság	11
belső harmónia	6	anyagi jólét	1
fegyelmezett	24	kellemes élet	13
egyenlőség	7	jókedélyű	27
segítőkész	31	tiszta	34
üdvözülés	18	boldogság	3
megbocsátó	29	törekvő	35
szavahihető	32	hatékony	26
béke	2	társadalmi megbecsülés	17
haza biztonsága	10	értelmes	23
szabadság	15	logikus gondolkodású	28
bátor, gerinces	20	elvégzett munka öröme	8
igazi szerelem	12	önálló	30
szeretetteljes	33	érdekes élet	9
engedelmes	22	alkotó szellem	19
udvarias	36	szépség világa	16
emberi önérzet	14	előítéletmentes élet	21

	Férfi	1	−0,1556	0,9878
	Nő	2	−0,2728	0,9621
J	2000	3	−0,4207	0,9072
ö	2–3999	4	−0,3510	0,9364
v	4–5999	5	−0,2888	0,9574
e	6–7999	6	−0,3285	0,9445
d	8–9999	7	−0,1620	0,9868
e	10–15000	8	−0,0074	1,0000
l	15001 +	9	−0,4228	0,9062
e				
m				
I	4 év	10	−0,7752	0,6317
s	5–8 év	11	−0,5298	0,8481
k	9–11 év	12	−0,3475	0,9377
o	12 év	13	−0,2125	0,9772
l	13–15 év	14	−0,1253	0,9921
a	16 év	15	−0,3498	0,9368
	18 évnél több	16	−0,5222	0,8528
É	20–29 éves	17	−0,2173	0,9761
l	30–39 éves	18	−0,1017	0,9948
e	40–49 éves	19	−0,2055	0,9786
t	50–59 éves	20	−0,2885	0,9575
k	60–69 éves	21	−0,2884	0,9575
o	70 +	22	−0,3174	0,9483
r				
		AVR	−0,1789	0,9839

14.19. táblázat. A társadalmi rétegek iránykoszinuszai a Rokeach-értékek axiológiai terében (amerikai adatok)



14.30. ábra. Rokeach-értékek axiológiai tere (MINISSA-eredmények, amerikai adatok).
(A csoportok kialakítása a faktorelemzés alapján).

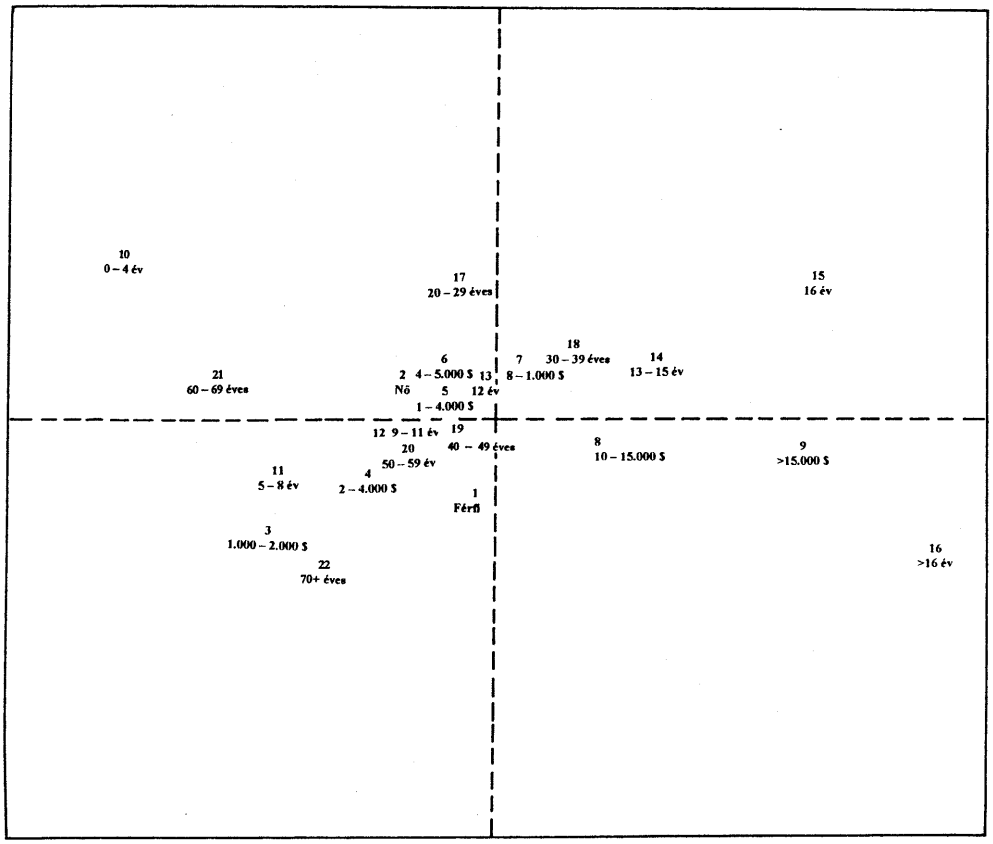


PT É	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
CHAR	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	G	H	I	J

PT É	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
CHAR	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	+

PT É	37	38	39	40	41	42	43	44	45	46	47	48	49	50	
CHAR	/	=	*			Ö	Á	%	?	<	(	)	,	;	¬

14.31. ábra. A Rokeach-értékek axiológiai terébe illesztett társadalmi rétegek (amerikai adatok)



14.32. ábra. Társadalmi rétegek az értékek terében (MINISSA-megoldás, amerikai adatok)

14.7. A PARAMAP-modell (PARAmetric MAPing)

A PARAMAP-eljárás n objektum sokdimenziós skálázását végzi el az objektumoknak m változóra vonatkozó megfigyelési értékei (vagy egy szimmetrikus különbözőségi mátrix) alapján.

A PARAMAP-eljárás az n objektumot az m dimenziós-megfigyelési térből átviszi az r-dimenziós ($r < m$) származtatott térbe úgy, hogy a folytonosságot maximalizálja.

A PARAMAP-eljárás feltételezése szerint a mért változók kifejezhetők r számú latens változó nemlineáris függvényével. Így az n objektum az m-dimenziós térből átvihető egy redukált r dimenziós térbe.

A PARAMAP-eljárás a fentiek alapján a nemlineáris faktorelemzés egy fajtája. A PARAMAP-programot Shepard és Carroll (1966), valamint Carroll és Chang (1973) fej-

lesztette ki. Különösen akkor előnyös az alkalmazása, ha a megfigyelt változók között erős nemlineáris, esetleg nemmonoton kapcsolat van.

A PARAMAP-program kétfajta adatmátrixot fogad el:

– $(n \times m)$ típusú, n objektum m változóra vonatkozó mérési eredményeit tartalmazó mátrixot; vagy

– egy $(n \times n)$-es szimmetrikus különbözőségi (vagy hasonlósági) mátrixot.

Az utóbbi esetben a szimmetrikus mátrix lehet távolságok, távolság-négyzetek, korrelációk vagy kovarianciák négyzetes mátrixa.

A korreláció- vagy kovarianciamátrix esetén a koszinusz-tétel alapján számítjuk a távolságokat:

$$d_{ij}^2 = c_{ii} + c_{jj} - 2c_{ij}.$$

Korreláció esetén természetesen $c_{ii} = c_{jj} = 1$ ezért:

$$d_{ij}^2 = 2(1 - r_{ij}).$$

Más különbözőségi mértékek (δ_{ij}) esetén e távolságot a következőképpen számíthatjuk ki:

$$d_{ij}^2 = \sum_{k=1}^{n} \left(\delta_{ik} - \delta_{jk}\right)^2.$$

14.7.1. A PARAMAP-modell

Tartalmazza az $\mathbf{Y}$ mátrix az n objektum m változóra vonatkozó megfigyelési értékeit:

$$\mathbf{Y} = \{y_{ik}\}, \qquad \text{ahol} \quad i = 1, 2, \ldots, n,$$
$$k = 1, 2, \ldots, m.$$

Tegyük fel, hogy létezik r nem megfigyelt (mögöttes, latens) változó (dimenzió):

$$x_1, x_2, \ldots, x_r,$$

amelyekkel a megfigyelt változók valamilyen f_k függvénnyel kifejezhetők:

$$y_k = f_k(x_1, x_2, \ldots, x_r)$$

vagy az i-edik egyedre (objektumra)

$$y_{ik} = f_k(x_{i1}, x_{i2}, \ldots, x_{ir}).$$

Vektorjelöléssel:

$$\mathbf{y} = F(\mathbf{x})$$
$$\mathbf{y}_i' = F(\mathbf{x}_i'),$$

ahol F írja elő az x térből az y térbe való illesztést.

Az x változókról feltételezzük, hogy számuk kevesebb, mint az y változók száma $r < m$.

A PARAMAP-eljárás keresi az r-dimenziós térben azt a megoldást, amely esetén a megfigyelési egységek (objektumok) között az eredeti térben mért „megfigyelt" távolságok (d_{ij}) és a származtatott térben mért távolságok (D_{ij}) jól illeszkednek a Carroll-féle folytonossági mérték (measure of continuity) szerint. A folytonossági mérték (KAPPA

κ) általános képlete:

$$\kappa = \left[\sum_{i \neq j}^{n} \left(d_{ij}^2\right)^a / \left(D_{ij}^2\right)^b\right] / \left[\sum_{i \neq j}^{n} \left(D_{ij}^2\right)^c\right]^{-b/c},$$

ahol $d_{ij}^2 = \sum_{k=1}^{m} \left(y_{ik} - y_{jk}\right)^2$ és $D_{ij}^2 = \sum_{k'=1}^{r} \left(x_{ik'} - x_{jk'}\right)^2 c\ \ell^2$, ahol: ℓ^2 egységnyi szórású paraméter, c normalizáló konstans.

Kruskal és Carroll (1968) a, b és c paraméterek következő megválasztását javasolta:

	$a = 1$	$b = 1$	$c = -1$
vagy	$a = 0{,}5$	$b = 1$	$c = -1$
vagy	$a = 0{,}5$	$b = 0{,}5$	$c = -1$
vagy	$a = 1$	$b = 2$	$c = -1$.

Leggyakrabban az utóbbi paraméterválasztást használják. Ebben az esetben a KAPPA a következő formát ölti:

$$\kappa = \left[\sum_{i \neq j}^{n} d_{ij}^2 / D_{ij}^4\right] \Bigg/ \left[\sum \frac{1}{D_{ij}^2}\right]^2.$$

A κ mutató a folytonosságot inverz módon méri.

Egyenletesebb függvénykapcsolatokhoz kisebb κ értékek tartoznak.

Az $a = 1, b = 2, c = -1$ paraméterek esetén Gower (1980) bizonyította, hogy a KAPPA-kritérium minimalizálása ekvivalens az

$$s_2(d, D) = \sum_{ij} w_{ij} \left(\frac{1}{d_{ij}^2} - \frac{1}{D_{ij}^2}\right)^2$$

minimalizálásával, és ekvivalens a

$$\rho_2^2(d, D) = \left[\sum_{ij} w_{ij} \frac{1}{d_{ij}^2} \frac{1}{D_{ij}^2}\right]^2 \left(\sum_{ij} w_{ij} / d_{ij}^4\right)$$

kifejezés maximalizálásával, ahol $w_{ij} = d_{ij}^2$.

Az $s_2(d, D)$ hasonló az

$$s_1(d, D) = \sum_{ij} w_{ij} (d_{ij} - D_{ij})^2$$

kritériumhoz, ha kifejtjük a négyzetes kifejezést, és a

$$w_{ij} / d_{ij}^4 D_{ij}^4 \approx 1 / d_{ij}^6$$

feltételezve a d_{ij} és a D_{ij} hasonló terjedelmét.

Így a PARAMAP-modellben a nagy távolságok nagyon kis súlyt kapnak, a kicsi távolságok pedig nagy súlyt. A PARAMAP-program a KAPPA-kritériumot minimalizálja, és eredményül a redukált térbe illesztett konfigurációt adja.

A számítási eljárás

A PARAMAP-eljárás a legmeredekebb lejtő (steepest descent) módszerével keresi az $(x_{ik'})$ értéket és az ℓ^2 értékét, amely optimalizálja a κ-át. A PARAMAP-program egy önkényes $(n \times r)$ típusú **x** mátrixból indul ki.

Ezután kiszámítja az ezzel összefüggő κ folytonossági mértéket, amelynek minimumát keressük. Ezután a „steepest descent" módszer iteratív eljárásával változtatjuk az **X** mátrixot, amíg el nem érünk egy stacionárius konfigurációhoz az adott dimenziószámú származtatott térben. Az iterációs eljáráson segíthetünk egy jól megválasztott kezdő konfiguráció megadásával, amelyet pl. kaphatunk más sokdimenziós skálázó módszer felhasználásával.

A PARAMAP-program az iteráció végén kapott **x** mátrixot normalizálja és a főtengelyekhez rotálja.

14.7.2. Példa a PARAMAP-eljárásra (Társadalmi csoportok az értékek terében)

Milton Rokeach az amerikai nemzeti mintán (1409 felnőtt, húsz év feletti lakos) 36 eszköz- és célérték struktúráját vizsgálta 1967–68-ban. A 18 cél- és 18 eszközérték vizsgálatát az Életmód, Életminőség és Értékrendszer vizsgálat (Hankiss E., Manchin R., Füstös L.) 1978-ban megismételte a magyar országos reprezantatív mintán (808 fő).

A PARAMAP-eljárást a társadalmi csoportok értékterének redukálására a magyar mintában két változatban alkalmaztuk.

Először: Nem adtunk meg kezdő konfigurációt a társadalmi csoportok kétdimenziós megjelenítéséhez. Az adathalmaz a 26 társadalmi csoport 36 értékválasztás mediánjaiból álló mátrix volt.

Másodszor: Kezdő kétdimenziós konfigurációnak megadtuk a MINISSA-eljárás kétdimenziós megoldását, amelyet a társadalmi csoportok között az érték-mediánok alapján számított euklideszi távolságokból számítottunk. Az adatmátrix ugyanaz volt, mint az első esetben.

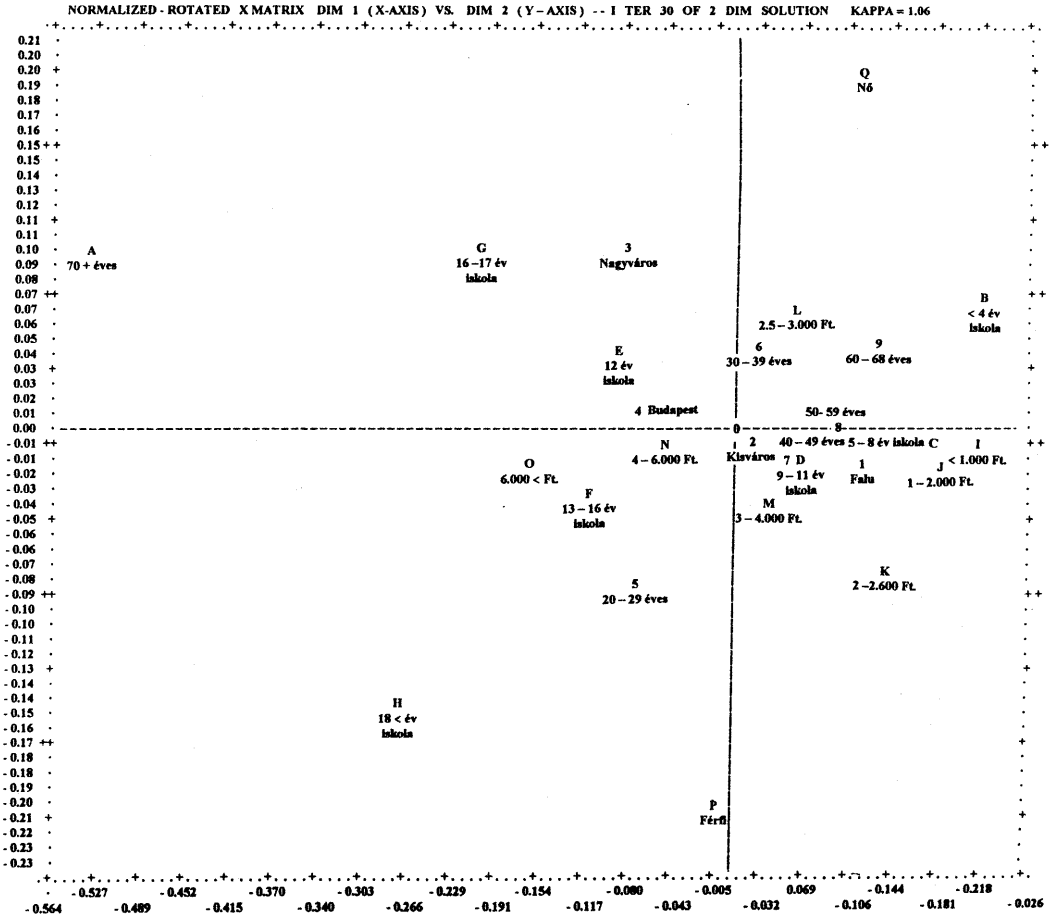
A két megoldás lényegében megegyezett.

Az első esetben azonban a 70 évesnél öregebb társadalmi csoport az első dimenzió ellenkező pólusára került. A 14.33. ábrán is jól látható, hogy az illesztés során valamelyik pont bizonyos pontokhoz közelebb került, míg másokhoz képest távolabb. Ez a pont a 70 évesnél öregebb társadalmi csoportot jelöli. Ez a csoport értékválasztásának nagyfokú inhomogenitását jelzi. A MINISSA-megoldással csaknem azonos eredményt kaptunk a PARAMAP-eljárás másodszori alkalmazásával. A pontábra ebben az esetben egy kicsit árnyaltabb, újabb különbségeket tudott felfedni azzal, hogy a távolabb lévő csoportok kisebb súlyt kaptak a PARAMAP-megoldásban.

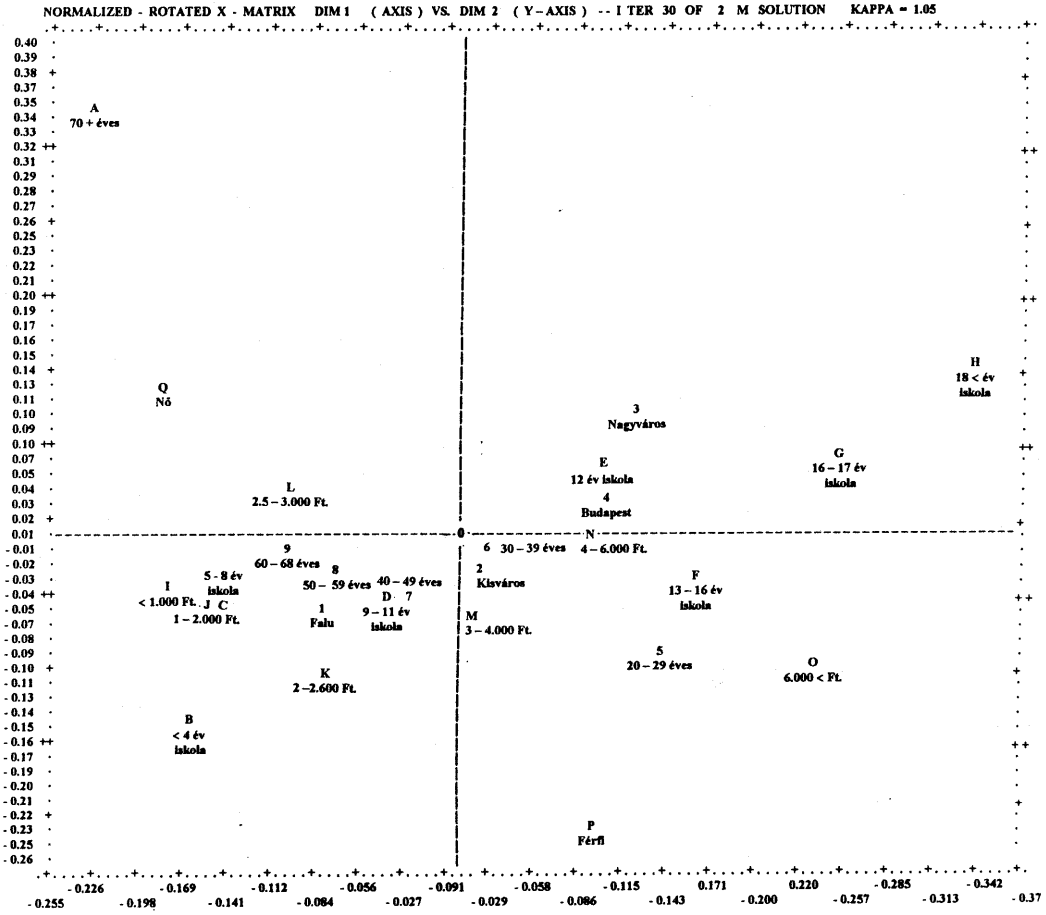
Az amerikai minta adatain végzett másodlagos feldolgozás során ugyancsak kiszámítottuk az értékek axiológiai terének koordinátáit a MINISSA-eljárással, és ennek alapján a PARAMAP-modell redukált terét.

E helyen a kapott eredményeket nem verbalizáljuk, azokat a kétdimenziós pontábráról bárki leolvashatja.

Megemlítjük, hogy a PREFMAP-modell fejezetében közölt eredmények kiegészítésül szolgálnak az itt szereplő ábrákhoz.



14.33. ábra. Társadalmi csoportok az értékek terében (a kezdő konfiguráció véletlen megadásával, magyar adatok)



14.34. ábra. Társadalmi csoportok az értékek terében (a kezdő konfiguráció a MINISSA-megoldás, magyar adatok)

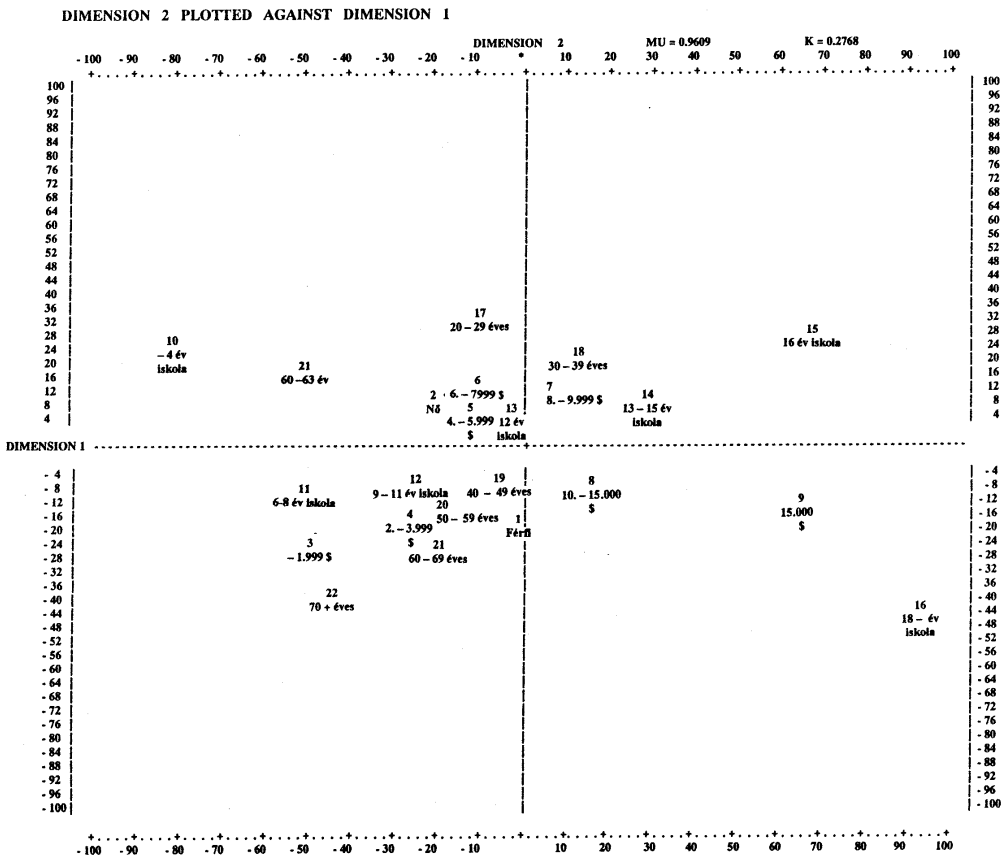
	N
1. Havi jövedelem	
1000 Ft alatt	260
1000–3000 Ft között	445
3000–5000 Ft között	517
5000 Ft felett	239
2. Lakóhely	
Község	703
Kisváros	327
Nagyváros (megyei jogú város)	118
Budapest	312
3. Életkor	
20 év alatt	15
20–30 év között	263
30–40 év között	299
40–50 év között	318
50–60 év között	303
60–70 év között	216
70 év felett	48
4. Nem	
Férfi	738
Nő	724
5. Iskolai végzettség	
0–4 év	135
5–8 év	495
9–11 év	338
12 év	131
13–15 év	187
16–17 év	77
18 évnél több	98
6. Foglalkozás	
Vezető	33
Értelmiségi	84
Egyéb szellemi	210
Szakmunkás	213
Betanított munkás	187
Segédmunkás	125
Kisiparos	24
Egyéni gazda	17
Inaktív (nyugdíjas, GYES stb.)	389
Mezőgazdasági fizikai dolgozó	93

14.20. táblázat. Társadalmi csoportok a magyar mintában

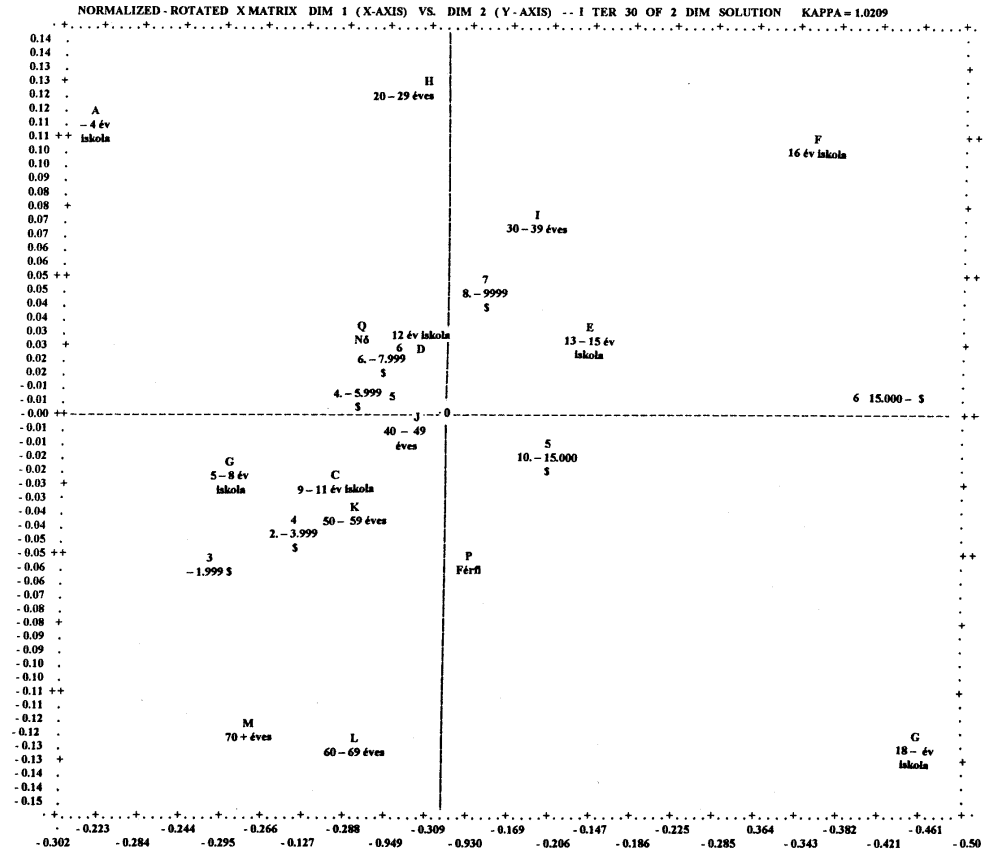
A következő táblázatban, ill. ábrákban az amerikai minta adatai találhatók, amelyeket Milton Rokeach „*The Nature of Human Values*” c. könyvében közölt korrelációs mátrix felhasználásával származtattunk.

Nem		N
1.	Férfi	665
2.	Nő	744
Jövedelem		
3.	1999 $ vagy kevesebb	139
4.	2000–3999 $ között	239
5.	4000–5999 $ között	217
6.	6000–7999 $ között	249
7.	8000–9999 $ között	178
8.	10000–14999 $ között	208
9.	15000 $ vagy több	95
Iskolai végzettség (években)		
10.	4 évnél kevesebb	64
11.	5–8 év	263
12.	9–11 év	320
13.	12 év	426
14.	13–15 év	180
15.	16 év	90
16.	18 évnél több	61
Életkor		
17.	20–29 éves	251
18.	30–39 éves	297
19.	40–49 éves	278
20.	50–59 éves	236
21.	60–69 éves	159
22.	70 éves vagy idősebb	167

14.21. táblázat. Társadalmi csoportok az amerikai mintában



14.35. ábra. Társadalmi csoportok az értékek terében (MINISSA-megoldás, amerikai adatok)



14.36. ábra. Társadalmi csoportok az értékek terében (a kezdő konfiguráció a MINISSA-megoldás, amerikai adatok)

14.8. Az MDPREF-modell (**MultiDimensional PREFerence scaling**)

Az MDPREF-eljárás preferencia-rendezések elemzését végzi vagy páronkénti összehasonlítások adataiból vagy a preferenciákat kifejező mérési eredményekből kiindulva. A Q halmazhoz tartozó objektumokra adott a P halmazhoz tartozó személyek preferencia-rendezése. Az MDPREF-modell a személyeket és az objektumokat egy r-dimenziós közös térbe illeszti úgy, hogy az objektumoknak a személyek vektoraira vetített értékeinek relatív nagyságai megfeleljenek a személyek objektumokra vonatkozó preferenciáinak. Az MDPREF-modell preferencia illesztése hasonlít a PROFIT-eljárásra, azonban míg a PROFIT-módszernél az objektumok *a priori* terébe illesztjük a személyek preferenciáinak legjobban megfelelő vektort, az MDPREF-eljárásnál az objektumok terét szimultán, a személyek terével együtt határozzuk meg.

14.8.1A z MDPREF-modell

Tegyük fel, hogy n megfigyelési egység (vizsgálati személy, a továbbiakban személy) valamilyen közös tulajdonság, ismérv alapján m számú objektumot (stimulust) rangsorol. Tehát minden személynek van egy preferencia-rendezése, amely azt tartalmazza, hogy az adott személy mely objektumokat részesít előnyben mely objektumokkal szemben.

Tartalmazza az i-edik személy preferenciáit a $\mathbf{P}_i$ mátrix. A $\mathbf{P}_i$ mátrix általános eleme: $p_{i,jk}$, ahol $i = 1, 2, \ldots n$; $j, k = 1, 2, \ldots, m$. A $p_{i,jk}$ három értéket vehet fel aszerint, hogy az i személy a j vagy k objektumot részesíti-e előnyben, vagy indifferens számára a kettő.

$$p_{i,jk} = \begin{cases} +1 & \text{ha az } i \text{ személy a } j \text{ objektumot preferálja } k\text{-val szemben,} \\ -1 & \text{ha az } i \text{ személy a } k \text{ objektumot preferálja } j\text{-vel szemben,} \\ 0 & \text{ha az } i \text{ személy nem részesíti egyik objektumot sem előnyben.} \end{cases}$$

Az MDPREF-modell feltételezi, hogy az objektumok és a személyek reprezentálhatók egy r-dimenziós térben úgy, hogy az objektumoknak megfelelő pontok vetületei a személyek vektoraira relatív nagyságban kifejezik a személyek preferenciáit.

Legyen $\mathbf{x}_j = [x_{j1}, x_{j2}, \ldots, x_{jr}]'$ a j objektum r-elemű vektora, az i személy koordinátáit az r-dimenziós térben pedig az $\mathbf{y}_i = [y_{i1}, y_{i2}, \ldots, y_{ir}]'$ vektor jelölje. Ekkor az i személynek és a j objektumnak a becsült preferenciaértékét a két vektor skaláris szorzatával definiálhatjuk:

$$s_{ij} = \mathbf{y}_i' \mathbf{x}_j .$$

Általánosabban az m objektum koordinátáit az r-dimenziós térben az $\mathbf{X} = \{x_{jt}\}$ jelöli, az n személy koordinátái pedig az $\mathbf{Y} = \{y_{it}\}$ mátrixban találhatók. Az $\mathbf{X}$ mátrix $(m \times r)$ típusú, az $\mathbf{Y}$ mátrix $(n \times r)$ típusú. A becsült preferenciaértékek:

$$\mathbf{S} = \{s_{ij}\} = \mathbf{Y}\,\mathbf{X}'.$$

A probléma most úgy fogalmazható meg, hogy hogyan lehet meghatározni olyan $\mathbf{Y}$ és $\mathbf{X}$ mátrixokat (olyan pontokat az adott r-dimenziós térben), hogy a belőlük számított preferenciaértékek ($\mathbf{S}$) minél jobban közelítsék a megfigyelt (mért) preferenciaadatokat. Carroll és Chang (1964) adott két eljárást a feladat megoldására. Az egyik egy iteratív eljárás, a másik pedig az Eckart–Young dekompoziciós eljárást használja. Az MDPREF-program az utóbbi eljárással dolgozik. Az Eckart-Young eljárás az $\mathbf{S}, \mathbf{S}'$ vagy az $\mathbf{S}'\mathbf{S}$ mátrix sajátértékét és sajátvektorát számítja. Carroll és Chang Monte Carlo elemzéssel arra a megállapításra jutott, hogy az Eckart–Young eljárás ugyanolyan jól dolgozik, mint az iteratív eljárás.

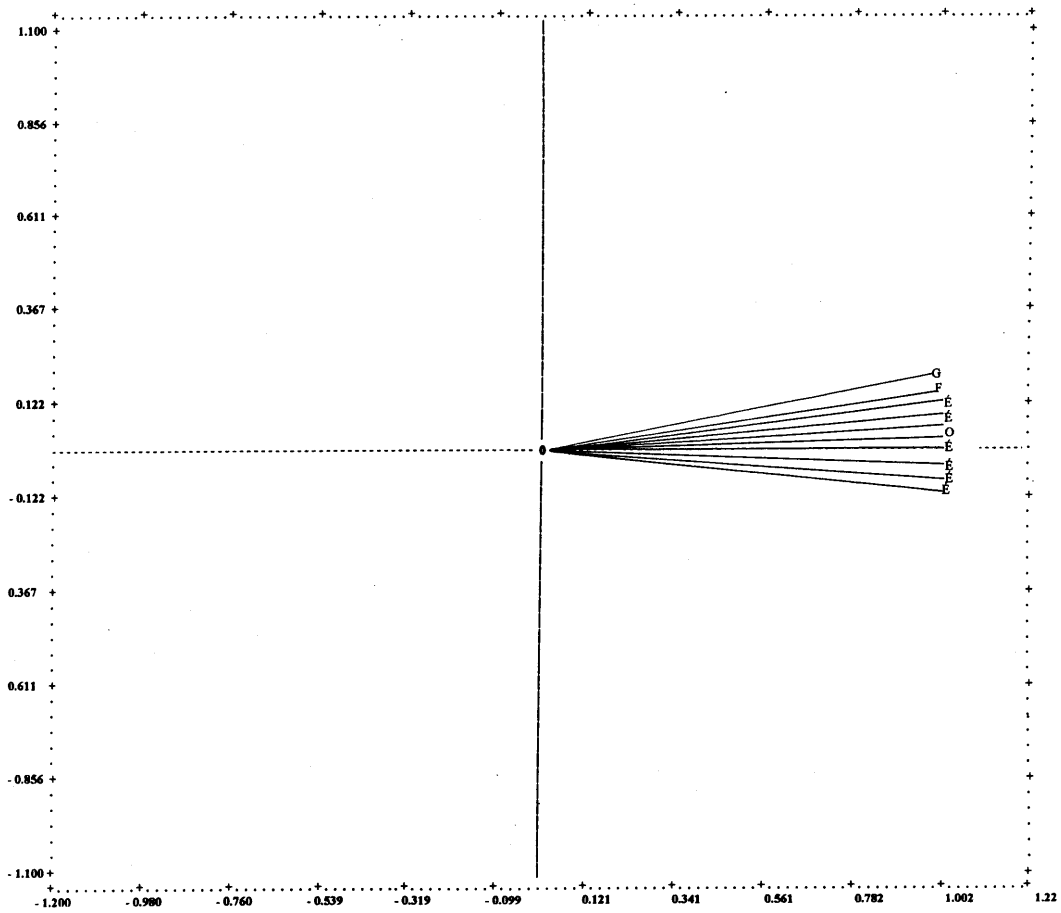
14.8.2. Példa az MDPREF-eljárásra (Társadalmi csoportok az értékek terében)

Az MDPREF-módszer bemutatására ugyanazokat az adatokat használjuk, amelyeket a MINIRSA-eljárásnál már ismertettünk. A huszonöt társadalmi csoport értékrendjét kifejező mediánokat az MDPREF-módszernél folytonos értékként adtuk meg. A megoldás koordinátáit és ábráit a következő táblázat és ábrák tartalmazzák. A társadalmi csoportok preferencia-tengelyeire ha levetítjük a Rokeach-értékek terében található pontokat, a társadalmi csoportok értékpreferenciáinak a lehető legjobban megfelelő rendezé-

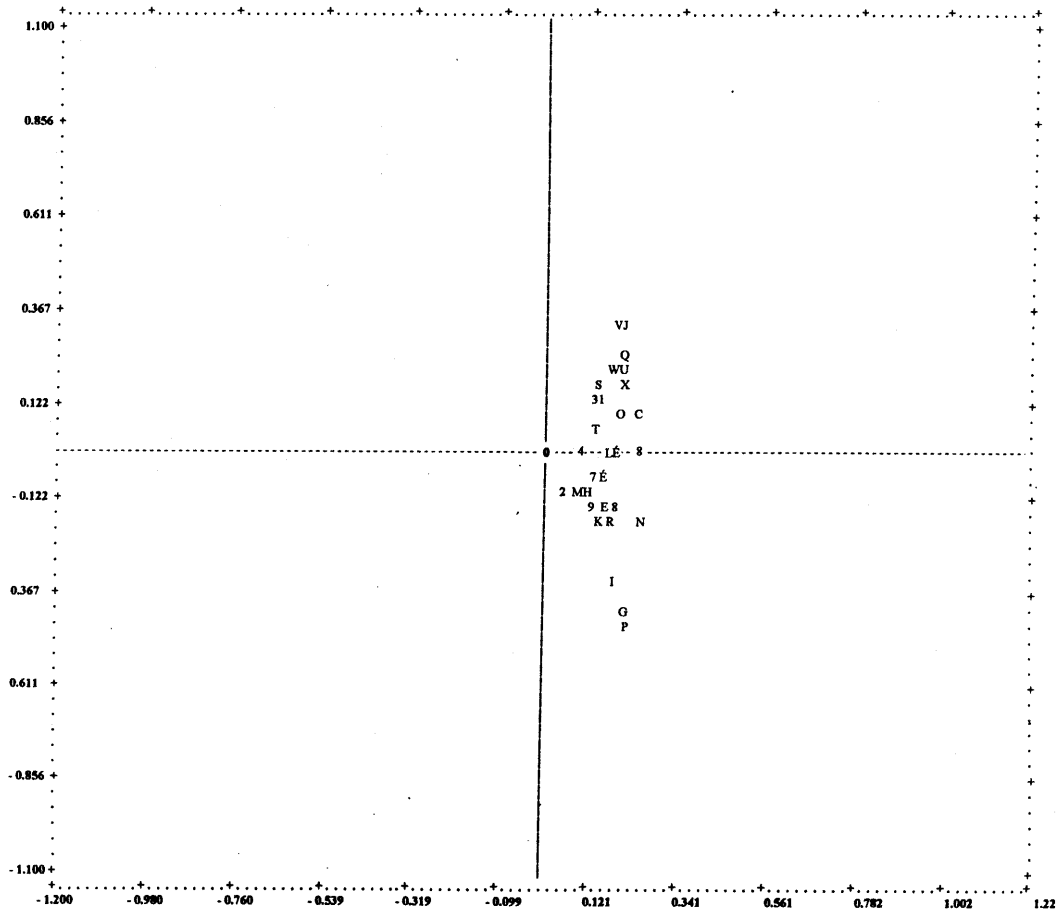
seket kapjuk. A Rokeach-értékeket és a társadalmi csoportokat a közös, együttes térben mutatja a harmadik ábra (14.39. ábra).

	1	2		1	2
1	0,9963	−0,0850	1	0,1426	0,1416
2	1,0000	0,0058	2	0,0481	−0,0629
3	0,0968	0,0796	3	0,1203	0,1183
4	0,9971	0,0763	4	0,0791	−0,0024
5	0,9976	0,0695	5	0,1601	−0,0562
6	1,0000	0,0034	6	0,1793	−0,0008
7	0,9998	−0,0193	7	0,1392	−0,0336
8	0,9986	−0,0528	8	0,2380	−0,0175
9	0,9969	−0,0789	9	0,1107	−0,1111
10	0,0930	−0,1184	10	0,1166	−0,0136
11	0,0902	−0,1395	11	0,1907	−0,1134
12	0,9991	−0,0349	12	0,2413	0,0844
13	0,9971	0,0767	13	0,1723	−0,0404
14	0,0941	0,1080	14	0,1557	−0,1135
15	0,9895	0,1438	15	0,1601	−0,0414
16	0,9749	0,2228	16	0,2057	−0,3758
17	0,9903	−0,1391	17	0,1199	−0,0896
18	0,9914	−0,1303	18	0,1964	−0,2781
19	0,9950	−0,0903	19	0,2204	0,3455
20	0,9999	−0,4500	20	0,1457	−0,1636
21	1,0000	0,0006	21	0,1709	0,0017
22	0,0977	0,0681	22	0,14040	−0,0869
23	0,9943	0,1062	23	0,2205	−0,1795
24	0,9991	0,0428	24	0,1891	0,0807
25	0,9965	−0,0826	25	0,2085	−0,3988
			26	0,2201	0,2633
			27	0,1698	−0,1484
			28	0,1463	0,1663
			29	0,1103	0,0464
			30	0,2104	0,2103
			31	0,2003	0,3109
			32	0,1773	0,1879
			33	0,2114	0,1824

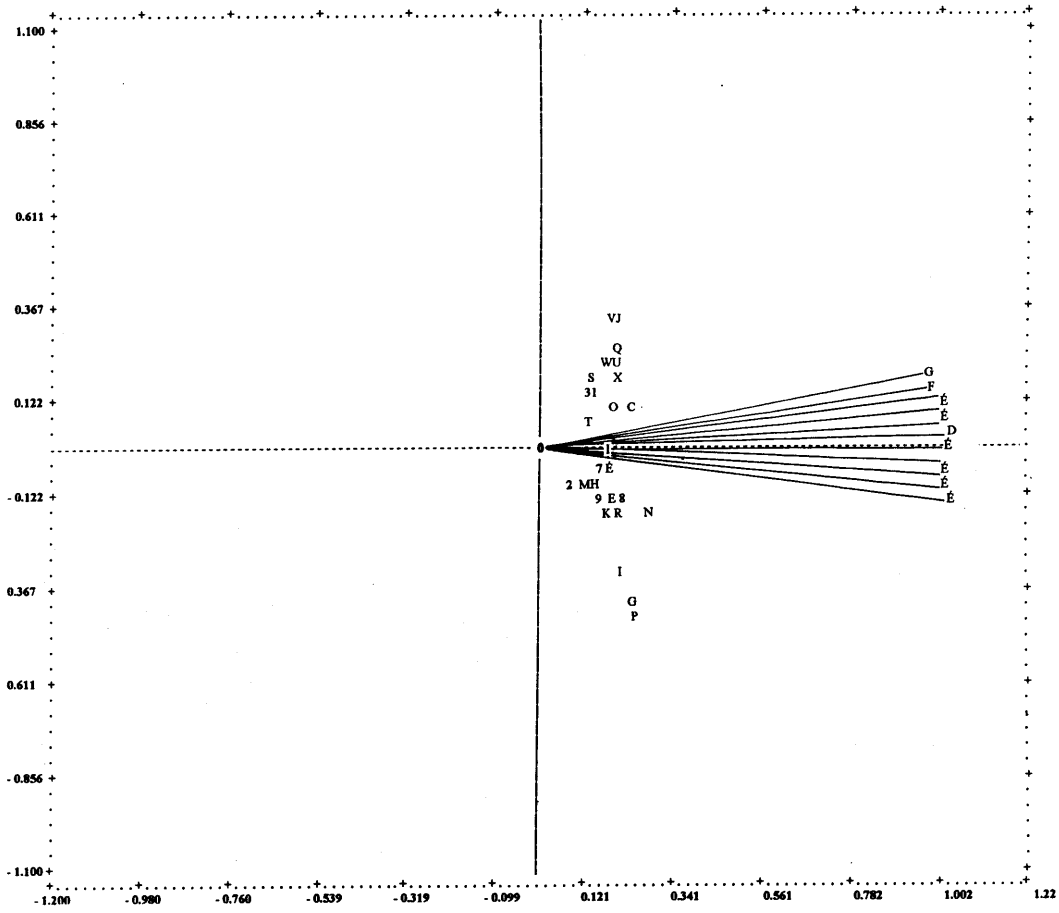
14.22. táblázat. A társadalmi csoportok és az értékek koordinátái a kétdimenziós preferenciatérben (MDPREF-megoldás)



14.37. ábra. A társadalmi csoportok preferenciatengelyei



14.38. ábra. Rokeach-értékek a preferenciatérben



14.39. ábra. Rokeach-értékek és társadalmi csoportok a preferenciatérben

14.9. A HICLUS-modell (HIerarchical CLUStering)

A HICLUS-modell objektumok optimálisan homogén csoportjait keresi. A csoportok hierarchikus rendszerét állítja elő az adatok monoton transzformációjára invariáns algoritmussal.

Az empirikus kutatások területén nagyon gyakran felmerülő probléma a nagyszámú összegyűjtött adat redukciója. Más szóval a probléma a nagyszámú adat belső struktúrájának a felismerése sokszor anélkül, hogy tiszta elméleti struktúrával rendelkeznénk. A HICLUS-eljárás az objektumok struktúráját az objektumok, ill. az objektumok csoportjainak (klasztereinek, fürtjeinek) hierarchikus elrendezésével próbálja felismertetni. A hierarchiában először minden egyes objektum különálló klaszterbe tartozik, a végén az összes objektum egy klaszterbe kerül. A hierarchiában a klaszterek optimálisan homogének, és a hierarchia független az adatok monoton transzformációjától.

A HICLUS-eljárás kifejlesztése és népszerűsítése S. C. Johnson (1967) nevéhez fűződik. Az MDS (X) programcsomagban szereplő HICLUS-programot is Johnson dolgozta ki.

14.9.1. A HICLUS-modell

A kiinduláskor rendelkezésünkre álló adatmátrix általában az n objektum m változóra vonatkozó megfigyelési értékeit tartalmazó $n \times m$ típusú mátrix ($\mathbf{X}$). Ebből a HICLUS inputját az objektumok között (vagy a változók között) hasonlósági/különbözőségi mértékek számításával nyerhetjük, így az input egy $n \times n$-es (vagy $m \times m$-es) hasonlósági/különbözőségi mátrix ($\mathbf{S}$). De hasonlósági mátrixhoz juthatunk az objektumok hasonlóságának közvetlen mérésével is. Vagyis a HICLUS-eljárás sok más eljárással ellentétben nem követeli meg feltétlenül, hogy az objektumok mint az euklideszi tér pontjai legyenek adottak. Az i-edik és j-edik objektum hasonlóságát jelöljük S_{ij}-vel. Az S_{ij}-ről feltételezzük, hogy kielégíti a következő két feltételt:

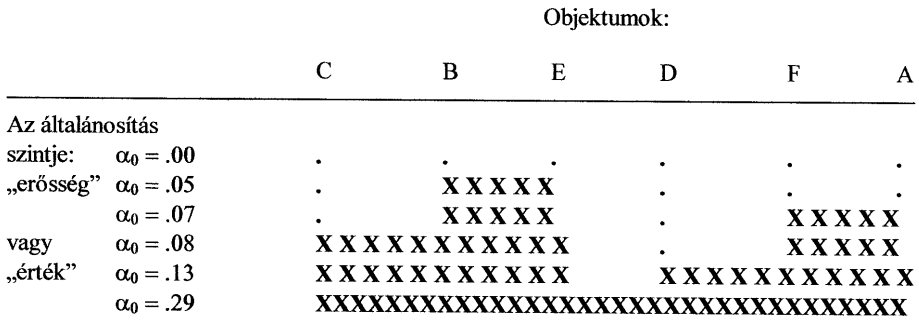
a) szimmetria: $S_{ij} = S_{ji}$ minden i, j-re $i, j = 1, \ldots, n$.

b) pozitivitás: $S_{ij} \geq 0$ minden i, j-re és $S_{ij} = 1$ akkor és csak akkor, ha $i = j$.

Kicsi S_{ij} érték az i-edik és j-edik objektum szoros kapcsolatára utal valamilyen nem definiált értelemben. (Hogy milyen értelemben, azt a hasonlósági mérték számítása vagy mérése határozhatja meg.)

Az n objektum között mért $n(n-1)/2$ hasonlósági mérték alapján a hierarchikus klaszterező módszer eredménye a klaszterek hierarchikus elrendezése.

Ha a hierarchikus szkémában nem engedjük meg az átfedő klasztereket, a hierarchikus szkémát dendrogramnak nevezzük. Ilyet mutat a következő ábra.



14.40. ábra. Dendrogram

Az ábrában legelőször (a legalacsonyabb szinten) minden objektum külön klaszter, így hat klaszterünk van. A legalsó szint „értéke” 0,00. A következő, első szinten öt klaszterünk van, mivel összevontuk a két leghasonlóbb klasztert, a B és E objektumot, amelyek ezentúl egy klaszterbe tartoznak. Az első szint „értéke” 0,05. A második szint 0,07 értéke mellett négy klaszterünk van: (C), (B, E), D, (F, A). A következő szinten a C kapcsolódik a (B, E) klaszterhez, a negyedik szinten pedig a D-t az (F, A) klaszterhez soroljuk. Az ötödik szinten 0,29 érték mellett mind a hat objektum egy klaszterbe kerül. Összefoglalva, a fenti ábra alapján azt mondhatjuk, hogy az α „érték” a 0-tól kezdődik, és monoton növekszik ahogy a klaszterek is „növekednek” hierarchikusan, minden klaszter (kivéve az első sort) az előző szint klasztereinek az összevonásából származik.

Most fogalmazzuk meg általánosabban a hierarchikus klaszterezés módszerét. Legyen adva az osztályozandó objektumok véges, nem üres halmaza:

$$S = \{s_1, s_2, \ldots, s_n\}.$$

Az objektumokat az indexek különböztetik meg, ami 1-től n-ig futhat, vagyis n objektumunk van.

Rendelkezünk a klaszterezésnek (osztályozásnak, csoportosításnak) egy sorozatával, $C_0, C_1, \ldots, C_p$, és minden C_i klaszterezéshez tartozó α_i számmal, ami az i-edik osztályozás értéke.

Megköveteljük, hogy C_0 esetben minden objektum külön klaszterbe tartozzon, ekkor $\alpha_0 = 0$, és C_p esetben minden objektum egy csoportba kerüljön. Megköveteljük az α_i értékek monotonitását, vagyis

$$\alpha_{i-1} \leq \alpha_i \qquad \text{minden } i = 1, 2, \ldots, p\text{-re}$$

és a klaszterezés monotonitását, vagyis

$$C_{i-1} \subset C_i,$$

más szóval minden klaszter a C_i osztályozásnál a C_{i-1} osztályozás klasztereinek összevonásával származhat.

Minden x, y objektumpárra definiálunk egy $d(x, y)$ mértéket: legyen j a [0, 1, ..., p] index-halmaznak azon legkisebb eleme, amelyhez tartozó C_j osztályozásban x és y azonos klaszterbe tartozik. Más szóval x és y legelső egyesüléséhez rendeljük a j-szintet.

Ekkor:

$$d(x, y) = \alpha_j.$$

Az előző ábrában például d(B, E) = .05, mivel B és E a .05 szinten került először egy klaszterbe, vagy d(D, A) = .13, mivel a negyedik lépésben vontuk egy klaszterbe őket. A teljes távolságmátrixot a következő táblázat tartalmazza:

d	A	B	C	D	E	F
A	0	0,29	0,29	0,13	0,29	0,07
B		0	0,08	0,29	0,05	0,29
C			0	0,29	0,08	0,29
D				0	0,29	0,13
E					0	0,29
F						0

14.23. táblázat. A 14.40. ábrával összefüggő távolságmátrix

A definícióból adódóan néhány dolog közvetlenül következik: például x és x ugyanabban a klaszterben van minden C_i-ben, és mivel 0 a j-index legkisebb értéke, következik hogy

$$d(x, x) = \alpha_0 = 0.$$

Viszont, ha $d(x, y) = 0$ valamely x-re és y-ra, ez azt jelenti, hogy x és y ugyanabba a klaszterbe tartozik a C_0-ban – de a C_0-ban csak egyelemű klaszterek vannak, ebből következően a $d(x, y) = 0$ akkor és csak akkor, ha $x = y$.

A $d(x, y) = d(y, x)$ minden x és y-ra, vagyis a $d(x, y)$ szimmetrikus. A d függvény metrika, ha teljesül rá a háromszög-egyenlőtlenség is. Ezt mutatjuk most be.

Legyen x, y, z három objektum, és

$$d(x, y) = \alpha_j$$

$$d(y, z) = \alpha_k.$$

Így x és y a C_j osztályozásnál ugyanabban a klaszterben van, míg y és z a C_k-ban kerül egy klaszterbe. Miután a klaszterezés hierarchikus, ezen klaszterek egyike tartalmazza a másikat, méghozzá az, amelyik j és k közül a nagyobbikkal függ össze. Jelöljük j és k közül a nagyobbikat ℓ-lel. Ekkor C_ℓ osztályozás produkál olyan klasztert, amelynek x, y és z is eleme. A d távolság definiciójából:

$$d(x, y) \leq \alpha_\ell.$$

De $\ell = \max[j, k]$, és α az indexek növekedésével együtt nő (nem csökken), ezért

$$\alpha_\ell = \max[\alpha_j, \alpha_k],$$

így végül

$$d(x, z) \leq \max[d(x, y), d(y, z)].$$

Ezt hívják *ultametrikus* egyenlőtlenségnek. Ez erősebb, mint a háromszög-egyenlőtlenség, amely szerint

$$d(x, z) \leq d(x, y) + d(y, z).$$

A távolságok és a hierarchikus szkéma között kölcsönösen egyértelmű megfeleltetés létesíthető.

A HICLUS-eljárás a hierarchikus szkémát vagy – mivel diszjunkt klasztereket kapunk – a dendrogramot a „minimum" vagy „maximum" módszerrel állítja elő. Ezek a módszerek invariánsak az adatok (különbözőségek) monoton transzformációjára.

14.9.2. A „minimum" módszer. A legközelebbi szomszéd (nearest neighbor) vagy egyszerű lánc (single linkage) módszer

A legközelebbi szomszéd módszer a következő: az első fázisban minden objektum külön klaszter, azután minden lépésben azt a két klasztert vonjuk össze, amelyek legközelebbi elemeinek a távolsága (vagyis a klaszterek távolsága) a legkisebb. A c klaszter ($c = \{j, k\}$) és az i-edik objektum (pont) közötti távolság a legközelebbi szomszéd módszer definíciója szerint:

$$d(c, i) = \min(d(i, j), d(i, k)).$$

Ez a módszer közbülső pontok miatt összekapcsolhat elkülönülő klasztereket (ez az ún. lánc-hatás). Ezért sokszor nehéz értelmezni a kapott eredményeket.

14.9.3. A „maximum" módszer. A legtávolabbi szomszéd (furthest neighbor) vagy teljes lánc (complete linkage) módszer

Az algoritmus megegyezik a legközelebbi szomszéd módszerével. A különbség a klaszterek távolságainak definiálásában van. A legtávolabbi módszerben a dendrogram valamely szintjén azokat a klasztereket vonjuk össze, amelyek legtávolabbi elemeinek a távolsága a legkisebb. Egy c klaszter és egy külső i pont közötti távolság a teljes lánc módszer szerint:

$$d(i, c) = \max(d(i, j), d(i, k)).$$

A definícióból adódóan ez a módszer láncosodásra kevésbé hajlamos, mint az előző módszer. Használata akkor célszerű, ha az objektumok halmaza kis átmérőjű csoportokat tartalmaz.

Tökéletes adatok esetén a két módszer megegyező eredményt ad. Ettől eltérő adatoknál azonban a két módszer különböző hierarchikus szkémát produkál. Ez a különbség éppen az adatok „tökéletlenségére" utal, vagyis arra, hogy milyen mértékben térnek el az ultrametrika feltételezésétől.

14.9.4. A HICLUS- és az MDS-módszerek kapcsolata

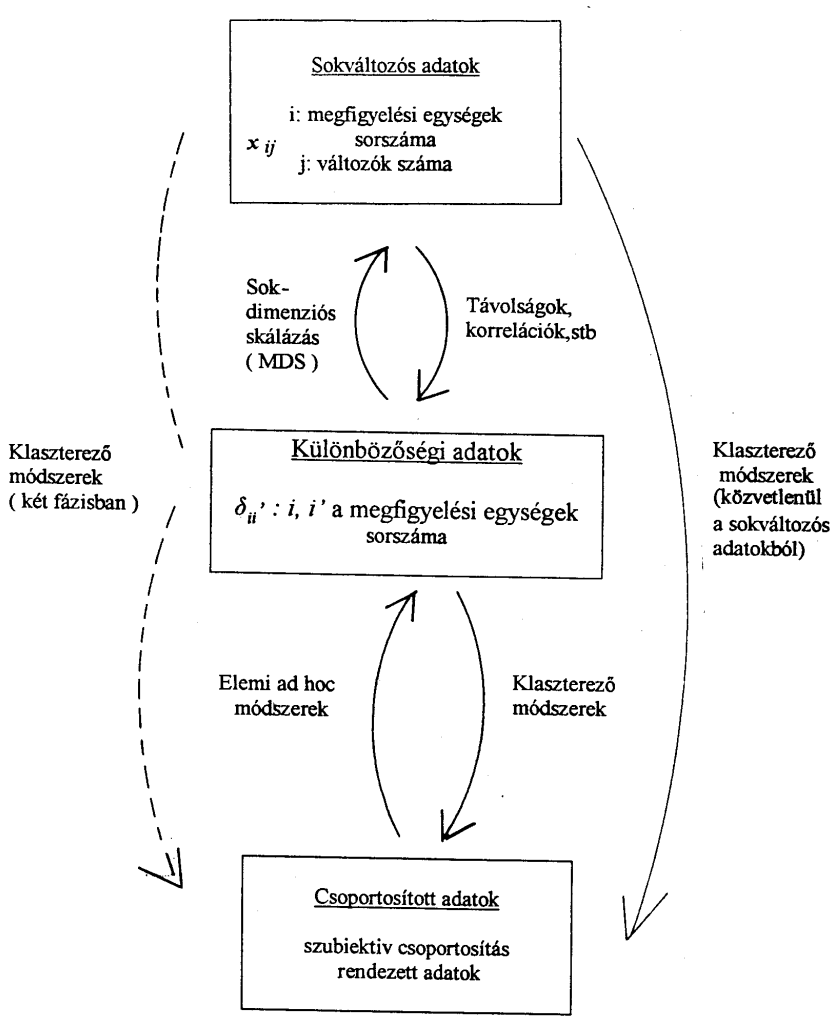
A klaszterező és a sokdimenziós skálázó (MDS) módszerek kapcsolatára kétfajta viszonyt szoktak említeni. Egyesek a két módszer konkurenciáját emelik ki. Még többen azt hangsúlyozzák, hogy a két módszer komplementáris kapcsolatban van egymással.

Nézzünk először egy Kruskaltól származó ábrát (14.41. ábra), amely a két módszer kapcsolatát jól szemlélteti.

Az ábrában három típusú adat van. Az első típusba az ún. sokváltozós adathalmaz tartozik. A sokváltozós adathalmaz sok változó megfigyelési értékeit tartalmazza a megfigyelési egységek halmazán. Ezt a x_{ij} jelöléssel szoktuk leírni, ahol az i-index a megfigyelési egységet jelöli, a j-index pedig a változókra vonatkozik. Az x tehát a j-edik változónak az i-edik megfigyelési egységre vonatkozó értéke. A különbözőségek (második típusú adat) halmaza jelenthet bármilyen hasonlósági vagy távolság- vagy korrelációs adatokat, amelyek két megfigyelési egység (objektum, az objektum lehet a változó, és lehet a megfigyelési egység is) közelségét vagy távolságát jelzik.

A harmadik típusa az adatoknak a csoportosított adatok. Általában ilyen típusú adatot az elemzés végén szoktunk kapni. Azonban sokszor előfordul, hogy a megfigyelési

egységektől változók, tulajdonságok, stimulusok csoportokba rendezését kérjük a hasonlóságuk alapján.



14.41. ábra. A klaszterelemzés és a sokdimenziós skálázás kapcsolódása

A klaszterelemző (fürtelemző) módszerek az első vagy második típusú adathalmazzal kezdenek, és ezeket csoportosított adatokká transzformálják. A klaszterelemzésnél fontos, de közbülső lépés a sokváltozós adatok különbözőségi adatokká való transzformálása.

A sokdimenziós skálázás viszont a különbözőségi adatokból indul ki, és eredményül sokváltozós adathalmazt ad. Ebből azután újra különbözőségi adatokat számíthatunk, amit viszont már másodrendű különbözőségeknek nevezünk.

Az elemzések során tetszőleges egymásutániságban járhatunk a három típusú adat között. Például a sokváltozós adathalmazból különbözőségeket számíthatunk (pl. euklideszi távolságokat), ezután a sokdimenziós skálázással egy redukált sokváltozós adat-

halmazhoz juthatunk, majd azt az adathalmazt klaszterezve eljuthatunk a csoportosított adatokhoz. A két módszer alapvető különbsége, hogy a sokdimenziós skálázás a különbözőségek alapján az objektumok térbeli reprezentálását végzi el, míg a klaszterelemzés az objektumokat kétdimenziós fa- (speciális gráf), hierarchikus szkéma, vagy dendrogram ábrában helyezi el.

Eric Holman (1972) a két módszer kompetitív kapcsolatát hangsúlyozza. Azt próbálja bizonyítani, hogy ha az adatok jól illeszkednek a klaszter-modellhez, akkor nagyon sok dimenzióra van szükség a sokdimenziós skálázás pontos megoldásához. Vagyis Holman szerint, ha az egyik modell nem illeszkedik jól (pl. a skálázó), akkor a másik modell illeszkedik jól.

Kruskal benyomása ezzel szemben az, hogy pozitív kapcsolat van aközött, hogy egyik és egy másik modell jól illeszkedik egy adott adathalmazra.

A klaszterelemzés és a sokdimenziós skálázás praktikus kapcsolatára néhány lehetőség:

– a kétdimenziós MDS-megoldásban a két módszer eredménye közvetlenül összevethető;

– a kétdimenziós MDS-pontábrát csak mint a klaszterek bemutatását használjuk akkor, ha az MDS-megoldás többdimenziós;

– az MDS térbeli értelmezéséhez – a dimenziómenti értelmezésen túl – segíthetnek a klaszterek.

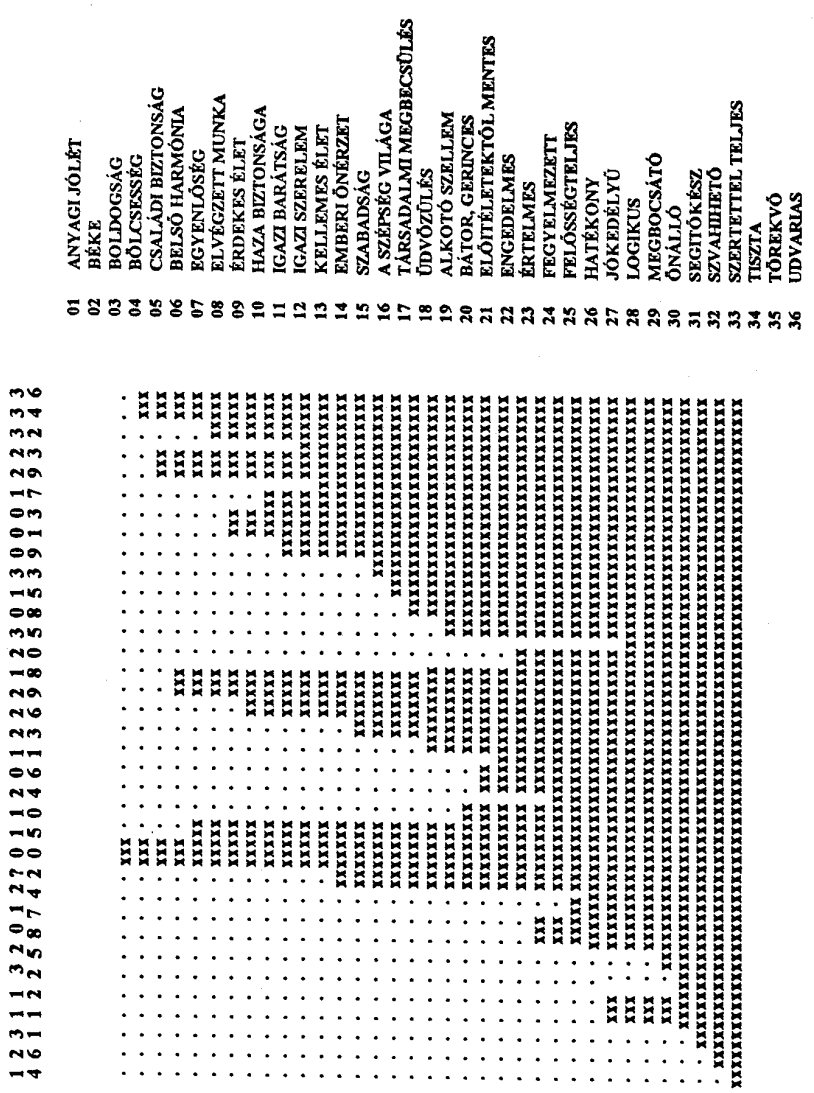
14.9.5. Példa a HICLUS-eljárásra (Rokeach-értékek klaszterei)

A példa szöveges leírása megegyezik az MRSCAL-modell példájával és input adatrendszerével.

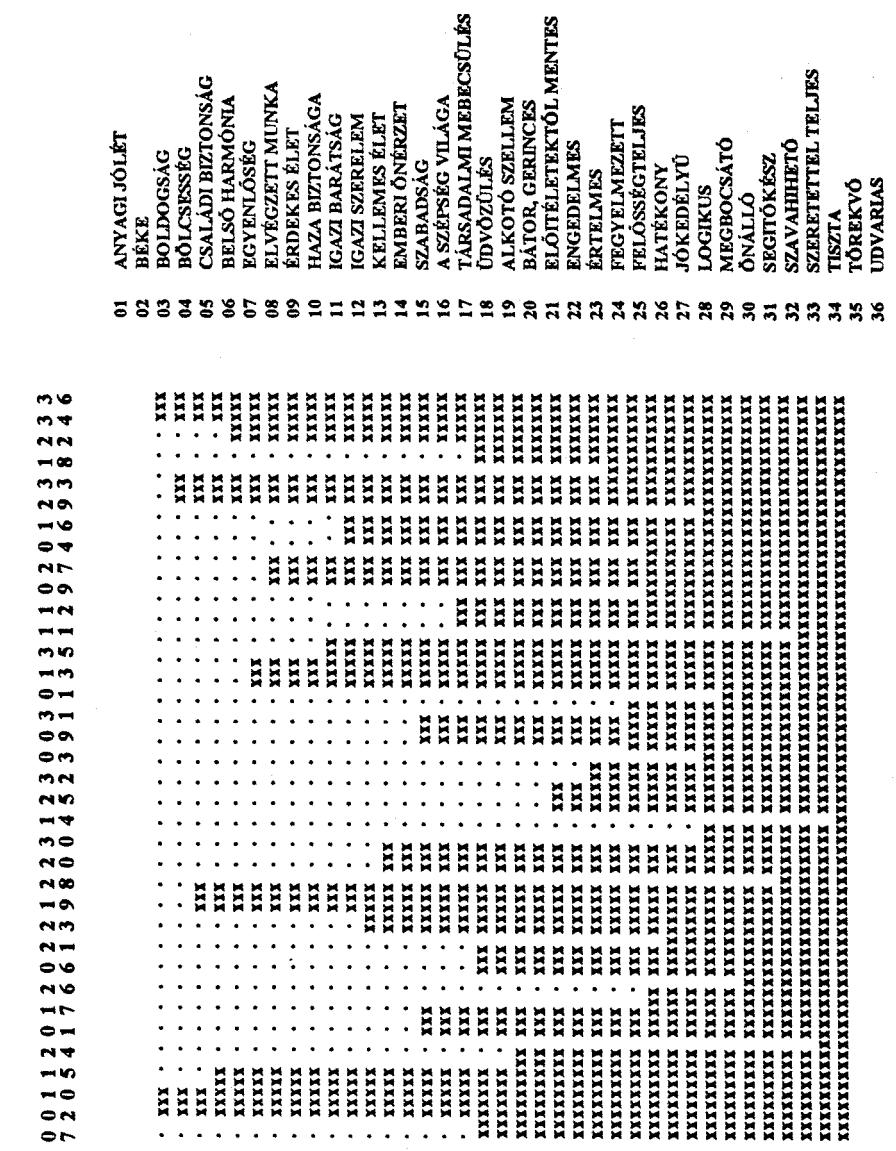
A HICLUS-módszer eredményeit négy ábrán mutatjuk be. Először a magyar minta, utána az amerikai minta megoldásait közöljük. A HICLUS inputját képező hasonlósági adatokat az értékek páronkénti korrelációs mátrixából képeztük a következő képlet szerint:

$$S_{ij} = \frac{r_{ij} + 1}{2}.$$

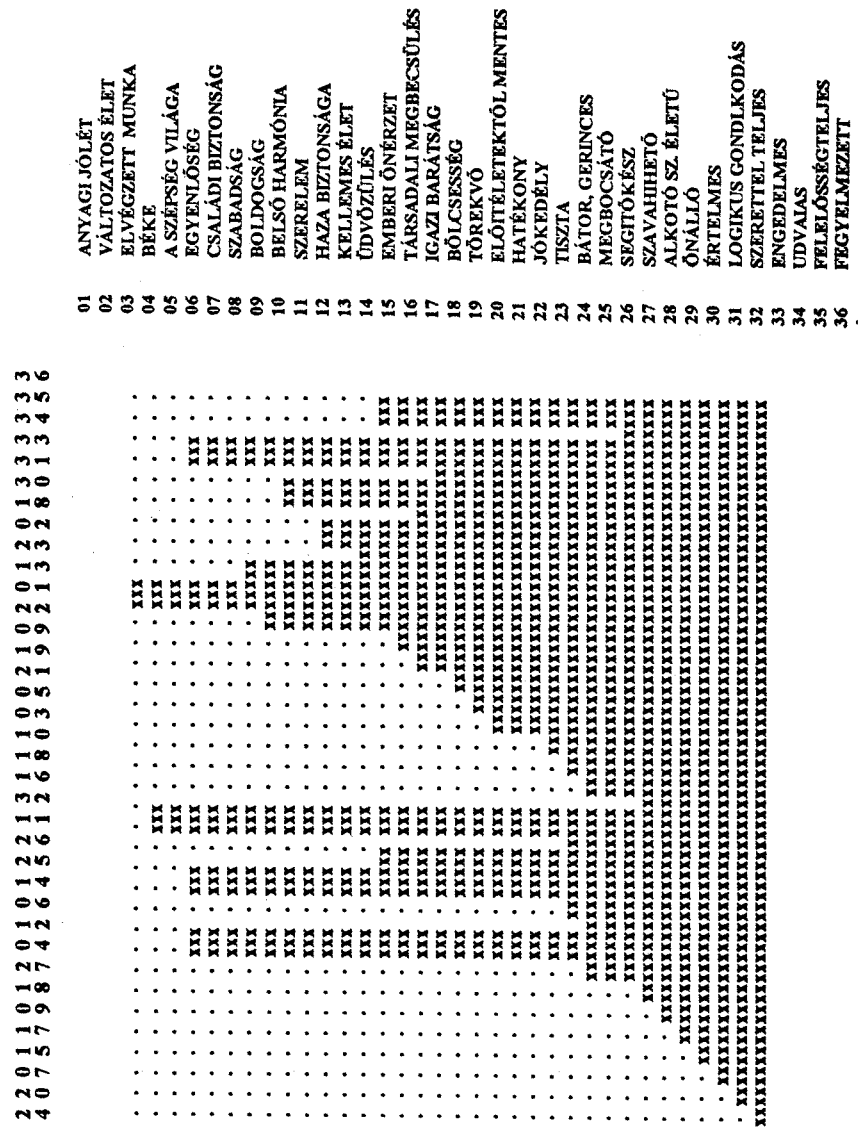
Az eredmények részletes értelmezésével ezen a helyen nem foglalkozunk.



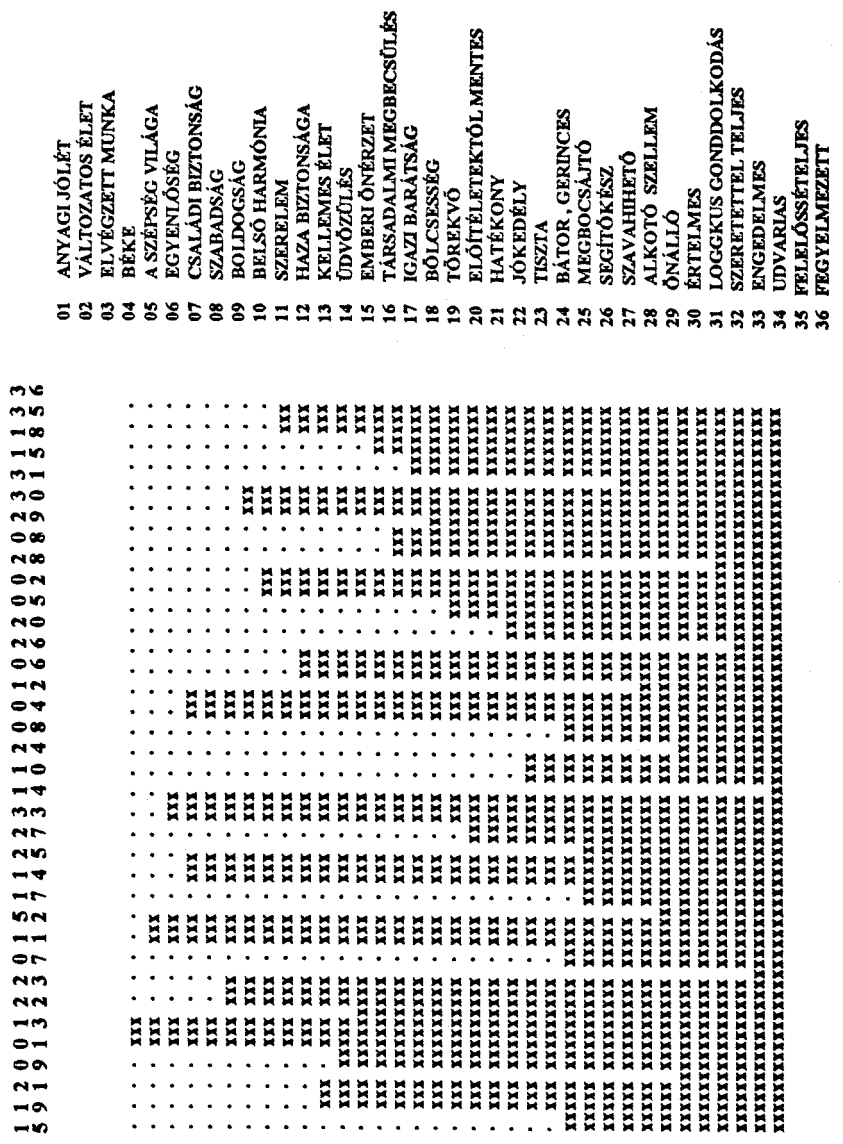
14.42. ábra. Rokeach-értékek klaszterei Magyarországon. (Legközelebbi szomszéd módszer)



14.43. ábra. Rokeach-értékek klaszterei Magyarországon. (Legtávolabbi szomszéd módszer)



14.44. ábra. Rokeach-értékek az USA-ban. (Legközelebbi szomszéd módszer)



14.45. ábra. Rokeach-értékek az USA-ban. (Legtávolabbi szomszéd módszer)

14.10. Az UNICON-modell
(UNIdimensional CONjoint measurement for multifaceted design)

Az UNICON-eljárás maximum öt független változónak egy függő változóra vonatkozó összetett, közös hatását (conjoint effect) határozza meg additív, szubtraktív vagy multiplikatív, illetve ezen elvek kombinációjával felállított modell segítségével.

A társadalomtudományokban gyakran vizsgáljuk egy független változóhalmaz együttes hatását egy függő változóra. Például, amikor azt gondoljuk, hogy az életkor és a nem multiplikatívan hat a keresetre, vagy hogy a társadalmi státus és az értékrendszer együttesen befolyásolja az emberek önértékelését, akkor egyesített mérést (conjoint measurement) hajtunk végre: a független változókat (amelyek gyakran minőségi – nominális, vagy ordinális – tulajdonságok) egyesítjük valamilyen formában, hogy így előállíthassuk a függő változó értékeit.

A legszélesebb körben ismert és gyakran alkalmazott formája az egyesített mérésnek az N-utas szóráselemzés (N-way analysis of variance – ANOVA). Az ANOVA-modellben a független változók hatását additív módon összesítjük.

Az ANOVA-modellben a függő változóról feltételezzük, hogy legalább intervallummérési szintű változó, a független változók pedig nominális mérési szintű változók. Kruskal (1964) mutatta meg, hogy ha a függő változó ordinális szintű, egy nemmetrikus MDS (MultiDimensional Scaling) eljárást használhatunk a függő változó értékeinek újraskálázására úgy, hogy a független változók hatása jó közelítéssel additív legyen. Ezt az eljárást Kruskal MONANOVA-nak (Monotone ANOVA) nevezte el.

14.10.1A z UNICON-modell

Az UNICON továbbfejlesztése a MONANOVA és a Guttman-féle CM (conjoint measurement) módszereknek. Az UNICON-eljárásban lehetőség van az additív modell mellett szubtraktív és multiplikatív, illetve ezek kombinációjából álló modellek becslésére is.

Legyen $Q = \{p_{jk\ell}\}$ a függő változó, és
$A = \{\alpha_j\}$ $j = 1, 2, \ldots, m_\alpha$
$B = \{\beta_k\}$ $k = 1, 2, \ldots, m_\beta$
$C = \{\gamma_\ell\}$ $\ell = 1, 2, \ldots, m_\gamma$
jelölje a független változókat.

Az adatok ekkor egy háromutas táblázatba rendezhetők. A táblázat általános eleme $\{z_{jk\ell}\}$ az A változó j-edik, a B változó k-adik és a C változó ℓ-edik kategóriájába (a táblázat $\{jk\ell\}$-háromdimenziós cellájába) eső megfigyelési egységeknek a függő változóra vonatkozó átlagos értéke.

Az általános modellben a függő változó értéke a független változók (ismeretlen) f függvényével fejezhető ki:

$$Q = z_{jk\ell} = (A_j, B_k, C_\ell).$$

Az additív modell:

$$f(A_j, B_k, C_\ell) = \alpha_j + \beta_k + \gamma_\ell.$$

A szubtraktív modell:

$$f(A_j, B_k, C_\ell) = \alpha_j - \beta_k - \gamma_\ell.$$

A multiplikatív modell:

$$f(A_j, B_k, C_\ell) = \alpha_j \beta_k \gamma_\ell.$$

Egy komplex modell pl.:

$$f(A_j, B_k, C_\ell) = (\alpha_j - \beta_k)\gamma_\ell.$$

Az UNICON-eljárás a független változók lehetséges kategóriáihoz numerikus skálaértékeket rendel:

$$a_j = g_A(\alpha_j)$$

$$b_k = g_B(\beta_k)$$

$$c_\ell = g_C(\gamma_\ell)$$

úgy, hogy az ezekkel produkált értékek ($\widehat{z}_{jk\ell}$) és (például: $\widehat{z}_{jk\ell} = a_j + b_k + c_\ell$) maximálisan illeszkedjenek az eredeti érékekhez, a $z_{jk\ell}$-hez.

Az illeszkedés jóságát a Stress Form2-függvénnyel mérjük, amely három független változó esetén a következő:

$$S_2 = \sqrt{\frac{\sum_j \sum_k \sum_\ell \left(z_{jk\ell} - \widehat{z}_{jk\ell(h)}\right)^2 \theta_{jk\ell(h)}}{\sum_j \sum_k \sum_\ell n_{jk\ell} \left(z_{jk\ell} - \overline{z}\right)^2}},$$

ahol

$z_{jk\ell} = f(a_j, b_k, c_\ell)$ a specifikált modell szerint

$\widehat{z}_{jk\ell}$ = monoton regressziós becslése $z_{jk\ell}$-nek a h. ismétlésben

$e_{jk\ell(h)} = 1$, ha j, k, ℓ a h. ismétlésben sorrendezett

$= 0$, ha j, k, ℓ hiányzó adat a h. ismétlésben

$$n_{jk\ell} = \sum_h e_{jk\ell(h)}$$

$$\overline{z} = \left(\sum_j \sum_k \sum_\ell n_{jk\ell} z_{jk\ell}\right) \Big/ \left(\sum_j \sum_k \sum_\ell n_{jk\ell}\right).$$

A $z_{jk\ell(h)}$ értéke nincs meghatározva, amikor $e_{jk\ell(h)} = 0$, de ez nincs hatással S_2-re.

Az UNICON-eljárásban S minimalizálása a legmeredekebb lejtő (steepest descent) módszerével történik.

A függő változó becsült értékeit pedig Kruskal-féle monoton regressziós eljárással számítjuk.

14.10.2. Példa az UNICON-eljárásra (A modernizáció hatása az önértékelésre)

Az MTA Szociológiai Intézet Értékszociológiai és Társadalmi Elemzések Műhelye az Életmód, Életminőség és Értékrendszer vizsgálatában (Hankiss E., Manchin R., Füstös L. 1978) megismételte a magyar országos reprezentatív mintán (808 fő) Milton Rokeach értéktesztjét. A 18 cél- és 18 eszközérték megfigyelési terében (mintatér) az embereket értéksorrendjeik alapján homogén csoportokba soroltuk a nemhierarchikus klaszterelemzés (fürtelemzés) módszerével (MacQueen-féle k-középpontú eljárással). A kialakult csoportokat, fürtöket a csoport átlagos értékválasztása, értéksorrendje alapján a

következőképpen neveztük el (erről részletesebben lásd Hankiss E., Manchin R., Füstös L., Szakolczai Árpád: *Folytonosság és Szakadás* c. könyvében):

Értékcsoportok

1. *Ideologikus elit*
2. Hagyományos, vallásos, örömtelen, passzív (*Kispolgári-ideologikus*)
3. Gazdag, örömteli, szociábilis személyiség (*Örömelvű*)
4. Hagyományos, vallásos, örömteli, törekvő (*Kispolgári-anyagias*)
5. Kemény, törekvő, alacsony szociábilitás (*Törekvő*)
6. Személyiségkorlátozó, fegyelmezett, etikus, örömtelen, (*Szolgálatelvű)*
7. Pragmatikus, technokrata, ideológiamentes (*Pragmatikus*)

Az embereket iskolai végzettségük szerint három kategóriába soroltuk:

Iskolai végzettség szerinti csoportok:

1. Általános iskola (8 év iskola)
2. Szakmunkásképzés (9–11 év iskola)
3. Érettségi v. felsőfokú iskola (12 év iskola vagy több).

Az emberek iskolai végzettsége és értékrendje az UNICON-modell két független változója. A modellben arra a kérdésre keressük a választ, hogy az iskolai végzettség és az értékrend együttesen hogyan befolyásolja, hogyan hat az emberek önértékelésére. A modell függő változója tehát az önértékelés. Ezt a következő kérdéssel mértük:

Kérdés:

Az ember gyakran viszonyítja saját helyzetét más emberekéhez. Például, ha megkérdezném Önt, hogy hol az Ön helye a társadalomban, ha mondjuk megkérném, hogy helyezze el magát a legjobb-módú és a legszegényebb emberek között, hova helyezné magát? Itt van egy kilenc fokú létra. Ha a legfelső fokon vannak a leggazdagabb emberek, a legalsó fokon a legszegényebbek, akkor Ön hova, melyik fokra helyezné Önmagát?

A példánkban most három létrát emelünk ki:

1. hatalom
A legnagyobb befolyással, hatalommal rendelkező emberek

9
8
7
6
5
4
3
2
1

Akiknek nincs beleszólásuk semmibe, nincs hatalmuk

2. hasznosság
A leghasznosabb, legnélkülözhetetlenebb emberek

9
8
7
6
5
4
3
2
1

A legnélkülözhetőbb, a legkevésbé hasznos emberek

3. jólét
A legmagasabb életszínvonalon élő emberek

9
8
7
6
5
4
3
2
1

A legalacsonyabb életszínvonalon élő emberek

Az embereket az iskolai végzettség és a választott értékrend alapján egy kétdimenziós (kétutas) táblázat celláiba soroltuk, és az egy cellába kerülteknek kiszámítottuk az átlagos pozíciójukat a Hatalom, Hasznosság és Jólét önértékelési létrán. Így minden cellában három átlagértékhez jutottunk. Ezután a Hatalom önértékelési skála átlagértékeit külön elhelyeztük az iskolai végzettség és az értékrend kétdimenziós celláiba, és ugyanezt tettük a Hasznosság és Jólét átlagértékeivel is. Vagyis három kétutas táblázatunk volt, amelyekre külön-külön illesztettünk UNICON-modellt.

Mindegyik táblázathoz három modellt illesztettünk, egy additív, egy szubtraktív és egy multiplikatív modellt.

A különböző modellek illeszkedésének jósága

	Hatalom	Hasznosság	Jólét
	Stress érték		
Additív modell $z_{jk} = a_j + b_k$	0,448	0,581	0,469
Szubtraktív modell $z_{jk} = a_j - b_k$	0,430	0,560	0,399
Multiplikatív modell $z_{jk} = a_j \times b_k$	0,032	0,011	0,385

A táblázatból látható, hogy csak a multiplikatív modell illeszkedik jól a Hatalom és a Hasznosság önértékelési skálához.

Az UNICON multiplikatív modell megoldása

		A függő változó	
Független változók		Hatalom	Hasznosság
Iskolai végzettség			
1. Általános isk.	a_1	0,021	−0,005
2. Szakmunkásképzés	a_2	1,064	2,080
3. Érettségi v. több	a_3	0,004	0,016
Értékrend			
1. Ideologikus	b_1	0,045	0,017
2. Kispolgári-ideol.	b_2	0,008	−0,012
3. Öröm-társas	b_3	0,073	0,018
4. Kispolgári-anyagias	b_4	0,896	1,816
5. Törekvő-felhalmozó	b_5	2,838	1,541
6. Szolgálatelvű	b_6	0,043	0,006
7. Pragmatikus	b_7	0,043	0,017

Az UNICON-modell a két független változó lehetséges kategóriájához skálaértéket rendel. A multiplikatív modell szerint ezeknek a skálaértékeknek a szorzata adja a függő változó becslését.

14.11. A PROFIT-modell (PROperty FITting)

Tételezzük fel, hogy rendelkezésünkre áll n pontról (ez lehet n stimulus, n egyed, megfigyelési egység stb.)

– egy *a priori* konfiguráció, az n egyed r dimenzióra vonatkozó koordinátája, és

– egy vagy több, az n egyedre (stimulusra) vonatkozó és egymástól függetlenül meghatározott mérési eredmény, amelyek az n stimulust valamilyen szempontból jellemzik, amelyeket tulajdonságoknak nevezünk.

A PROFIT-eljárás feladata az, hogy minden tulajdonsághoz találjon az r-dimenziós térben egy olyan vektort, amelyik maximálisan korrelál az adott tulajdonságokkal. Ezt az optimális „megfelelés"-t definiálhatjuk a lineáris korreláció vagy a nemlineáris korreláció értelmében. Másképpen fogalmazva, keressük az n stimulus pont adott *a priori* terében az n stimulus tulajdonságainak legjobban illeszkedő vektorát. Az n stimulus r dimenzióra vonatkozó koordinátái származhatnak más módszerek eredményeiből, így például a MINISSA-eljárás kevés dimenziószámú megoldásának koordinátái is lehetnek.

A feladat lineáris esetét Miller, Shepard és Chang (1964) dolgozta ki, míg a nemlineáris megoldás Carroll (1964) nevéhez fűződik. A PROFIT számítógépprogramját Carroll és Chang (1968) dolgozta ki. Ezt a programot illesztették az MDS (X) programcsomaghoz.

14.11.1. Tulajdonságillesztés lineáris regresszióval

A PROFIT matematikai modellje a következő jelöléseket tartalmazza:

$\mathbf{X} = \{x_{ij}\}$ mátrix n egyed (objektum, stimulus) koordinátáit tartalmazza az r-dimenziós térben

$i = 1, 2, \ldots, n; \quad j = 1, 2, \ldots, r.$

Feltételezzük, hogy az r-dimenziós tér origója az n egyed középpontjában van, azaz az $\mathbf{X}$ mátrix oszlopösszegei egyenlők nullával.

$\mathbf{p}'$ egy tulajdonság n egyedre vonatkozó megfigyelési értékeit tartalmazó sorvektor, $i = 1, 2, \ldots, n$

$\mathbf{t} = [t_j]$ az illesztett vektor iránykoszinusz[1] oszlopvektora, $j = 1, 2, \ldots, r$

$\mathbf{h}' = [h_i]$ az n pontnak az illesztett vektorra vonatkozó vetületei sorvektora.

A PROFIT-elemzés feladata megkeresni az iránykoszinuszok $\mathbf{t}$ és a vetületek $\mathbf{h}$ vektorát, amelyekre az

$$|\mathbf{p} - \mathbf{h}|^2 \quad \text{kifejezés minimális, ahol a} \quad \mathbf{h} = \mathbf{X}\,\mathbf{t}.$$

A feladat a lineáris regresszió feladatával azonos. A megoldást a legkisebb négyzetek módszere értelmében a következőképpen kapjuk:

$$\mathbf{t} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{p},$$

ami alapján

$$\mathbf{h} = \mathbf{X}\mathbf{t} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{p}.$$

Az eljárás

1. lépés

Az eljárás első lépésében normalizáljuk $\mathbf{X}$ mátrixot úgy, hogy elemeiből kivonjuk a megfelelő oszlop átlagát. Minden oszlop egy dimenziót reprezentál. Ezután számítjuk az XMAT-mátrixot, ahol

$$\mathbf{XMAT} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

2. lépés

Minden tulajdonságra a következő műveleteket végezzük el:

1. kiszámítjuk az iránykoszinuszokat (az illesztett vektor iránykoszinuszait);
2. kiszámítjuk az n pont vetületeit az illesztett vektorra;
3. kiszámítjuk a korrelációs együtthatót (RHO) az n pont illesztett vektorra vonatkozó vetületei és a megfelelő tulajdonság értékei között.

[1] Iránykoszinuszoknak az adott vektor és a koordinátatengelyek által bezárt szögek koszinuszait nevezzük.

14.11.2. Tulajdonságillesztés nemlineáris regresszióval

Carroll (1964.) definiált egy általános indexet egy p változó és egy függő x változó közötti nemlineáris korreláció mérésére:

$$K = \frac{1}{s^2} \sum_{\substack{i \neq j \\ i,j}}^{n} w_{ij}(x_i - x_j)^2,$$

ahol $w_{ij} = f(|p_i - p_j|)$, f egy monoton csökkenő függvény, $s^2 = \frac{1}{n} \sum_{i=1}^{n}(x_i - \overline{x})^2$.

A PROFIT-elemzésben a független, magyarázó változó (p) a tulajdonság, a függő, becsült változó (x) az n pontnak a vetületei r-dimenziós térben, a w_{ij} súlyokat pedig a következőképpen definiáljuk:

$$w_{ij} = 1/(p_i - p_j)^2 + \text{ konstans.}$$

A K a nemlineáris korrelációnak egy inverz mértéke, így a PROFIT-elemzésben az eljárás célja K minimalizálása. A feladat megtalálni az iránykoszinuszokat úgy, hogy a stimulusok vetületei a lehető legközelebb legyenek a megfelelő tulajdonság értékéhez. Az eljárás első lépéseként a kezdeti konfiguráció koordinátáit ortonormál rendszerré transzformáljuk ($\mathbf{X}'\mathbf{X} = \mathbf{I}$) a főtengely-transzformáció végrehajtásával, és minden tengelyt egységnyi hosszúságúra standardizálunk.

Az illesztett vektor iránykoszinuszát és a stimulus vetületeit úgy számítjuk ki, hogy először a szimmetrikus $\mathbf{X}'\mathbf{A}\mathbf{X}$ mátrixot számítjuk, ahol $a_{ij} = -w_{ij}$ minden $i \neq j$ és $a_{ii} = n \sum_{i \neq j} w_{ij}$.

Az $\mathbf{X}'\mathbf{A}\mathbf{X}$ legkisebb nem nulla karakterisztikus gyöke lesz a K minimális értéke, és az ezzel összefüggő karakterisztikus vektor adja az iránykoszinuszok vektorát, amiből: $\mathbf{h} = \mathbf{X}\mathbf{t}$.

Az algoritmus

1. lépés

Az $\mathbf{X}$ mátrix oszlopainak átlagát kivonjuk a megfelelő oszlop elemeiből, és ortonormált vektorokká transzformáljuk őket.

2. lépés

Minden tulajdonságra a következő műveleteket végezzük el:

1. kiszámítjuk K első négy momentumát, majd a ferdeség és csúcsosság mérőszámát,
2. kiszámítjuk Z és Z négyzet (ZSQ) értékét, ahol Z a nemlineáris korreláció standardizált indexe. A ZSQ megközelítőleg khi négyzet eloszlású r (a dimenziók száma) szabadságfokkal.
   A Z értékét a következőképpen számítjuk:
   $$Z = \frac{K - M_k}{O_k},$$
   ahol M_k a K átlaga és O_k a szórása.
   A harmadik és a negyedik momentum számítása K eloszlásának a normális eloszlástól való különbözősége mérésére szolgál. Ha a K eloszlása normális, a ferdeség és csúcsosság mérőszámának értéke 0, ill. 3.
3. Kiszámítjuk az $\mathbf{X}'\mathbf{A}\mathbf{X}$ szimmetrikus mátrixot, és megkeressük a legkisebb karakterisztikus gyökét (K). A megfelelő karakterisztikus vektor ($\mathbf{t}$) tartalmazza az irányko-

szinuszokat, amelyek alapján számított vetületek (n pont vetülete) maximális nemlineáris korrelációban vannak (K értelme szerint) az adott tulajdonság értékeivel.

4. Kiszámítjuk az n pont vetületeit (**h**) a **t** iránykoszinuszok alapján.

3. lépés

Amikor a program minden vektort megtalált, minden vektorpárra kiszámítja a hajlásszög koszinuszát.

4. lépés

Minden vektorpárra kirajzolja a vektorokat és az n pont vetületeit.

5. lépés

Az X konfigurációt visszatranszformálja az eredeti koordinátákhoz.

6. lépés

Kiszámítja minden vektornak az iránykoszinuszát.

7. lépés

Kirajzolja a vektorok vetületeit a stimulus pontokkal együtt az eredeti térben.

14.11.3. Példa a PROFIT-eljárásra (Társadalmi csoportok értékpreferenciái az alapértékek terében)

Az 1978-as életminőség-vizsgálatban[2] a megkérdezettek 32 olyan emberi érték közül választották ki a nekik legfontosabb tizet, amelyeket az emberek (a korábbi vizsgálatok tanúsága szerint) általában fontosnak szoktak tartani. (Lásd részletesen az Életmód, Életminőség, Értékrendszer Országos vizsgálat, 1978. Alapadatok I. Hankiss Elemér, Manchin Róbert, Füstös László.) A 32 érték választásának kapcsolódását egy asszociációs mutatóval, a kontingenciaegyütthatóval mértük. Ennek alapján a MINISSA-eljárással kerestük meg az értékrendszer kétdimenziós szemantikai terét. Ebből az értékek belső kapcsolódási rendje által kifeszített szemantikai térből könnyen kiolvashatjuk az értékrendszer elemeit és egymáshoz való viszonyukat, az emberi értékek rendszerét. A különböző társadalmi csoportok értékválasztásának különbözőségét és eltérését a teljes minta szemantika terétől az INDSCAL-eljárás súlyainak kiszámításával kapjuk meg.

Az értéktérben egy-egy társadalmi csoport helyzetét, a csoportok az adott értéktérbe való illesztésével határozhatjuk meg. Mielőtt a PROFIT-eljárással kapott illesztést bemutatnánk, először a MINISSA-módszer eredményeit közöljük.

Az egy főre jutó négy jövedelmi csoport (1000 Ft alattiak, 1001–3000 Ft közöttiek, 3001–5000 Ft közöttiek, 5001 Ft felettiek) illesztését a MINISSA-eljárással meghatározott kétdimenziós szemantikai térbe a jövedelmi csoportok átlagos értékválasztása alapján a PROFIT-módszerrel végeztük.

[2] Részletes elemzés található a PROFIT-modell hasonló alkalmazásánál: Hankiss E., Manchin R., Füstös L., Szakolczai Á.: *Kényszerpályán?* MTA Szociológiai Kutatóintézet. Értékszociológiai Műhely Kiadványai, Budapest, 1982.

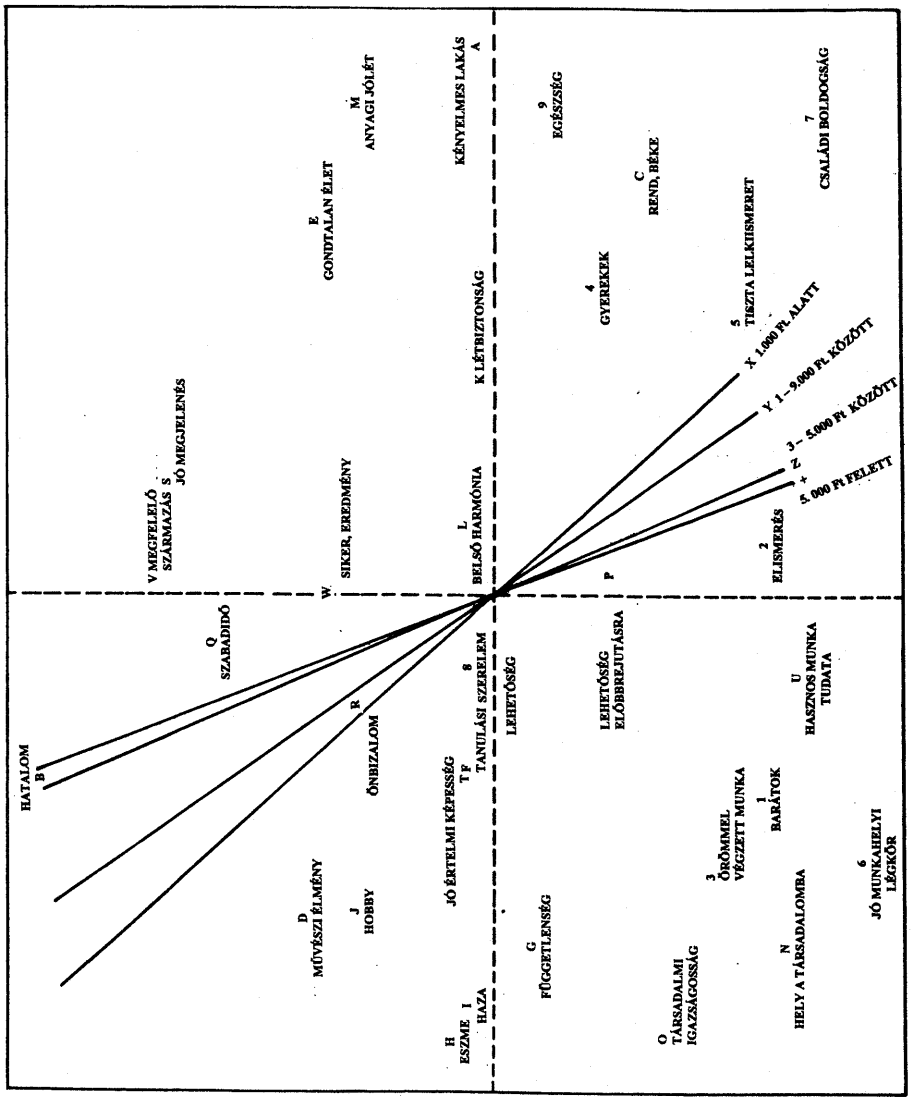
	1	2	3
1	−0,3655	−0,9236	−0,1288
2	0,0428	−0,4548	0,7767
3	−0,6279	−0,6426	0,0086
4	0,7955	−0,3347	0,2219
5	0,6401	−0,7402	−0,2496
6	−0,7213	−0,8782	0,7731
7	1,1602	−0,8266	0,2077
8	−0,1724	0,0522	−0,1586
9	1,1888	−0,0452	−0,4447
10	1,0171	−0,0297	0,7743
11	−0,3550	1,3031	0,3354
12	0,8115	−0,4432	−0,7260
13	−0,6082	0,5016	−0,6034
14	0,7005	0,5123	−0,7553
15	−0,2977	0,0509	0,4102
16	−0,9913	0,1820	0,1700
17	−0,7094	0,0022	−0,7752
18	−0,3022	−0,2441	−0,9713
19	−0,7913	0,3575	−0,2824
20	0,5690	0,0706	−0,3716
21	0,1607	0,0042	−0,5713
22	1,1070	0,5648	0,1185
23	−0,8417	−0,6458	0,3377
24	−0,8135	−0,4817	−0,5506
25	0,0769	0,0114	0,8061
26	−0,2646	0,9126	−0,2322
27	−0,1847	0,3668	0,0600
28	0,3114	1,0540	0,2576
29	−0,3718	−0,0234	0,2187
30	−0,1554	−0,8467	0,4206
31	−0,0042	0,9045	0,7453
32	−0,0034	0,5099	0,1793

14.24. táblázat. Az alapértékek szemantikai tere háromdimenziós MINISSA-megoldásának outputja

	1	2
1	−0,4315	−0,8348
2	0,1689	−0,7867
3	−0,6344	−0,6261
4	0,7687	−0,2844
5	0,7315	−0,7054
6	−0,6352	−1,1928
7	1,2978	−0,9079
8	−0,1897	0,1160
9	1,3405	−0,1138
10	1,3422	0,1811
11	−0,4905	1,4042
12	1,0827	−0,4698
13	−0,8053	0,5891
14	0,8563	0,7083
15	−0,2355	0,0097
16	−1,0220	0,2273
17	−1,0573	−0,0259
18	−0,9767	−0,0558
19	−0,7590	0,4457
20	0,5499	0,0816
21	0,2149	0,1207
22	1,1797	0,6069
23	−0,7998	−0,8523
24	−1,0382	−0,4364
25	0,1260	−0,3802
26	−0,2608	0,9017
27	−0,1934	0,3891
28	0,2337	1,1469
29	−0,2639	−0,0497
30	−0,1261	−0,9532
31	0,0600	1,1996
32	−0,0333	0,5475
ÁTLAG:	0,0000	0,0000
SZÓRÁS:	0,7444	0,6677

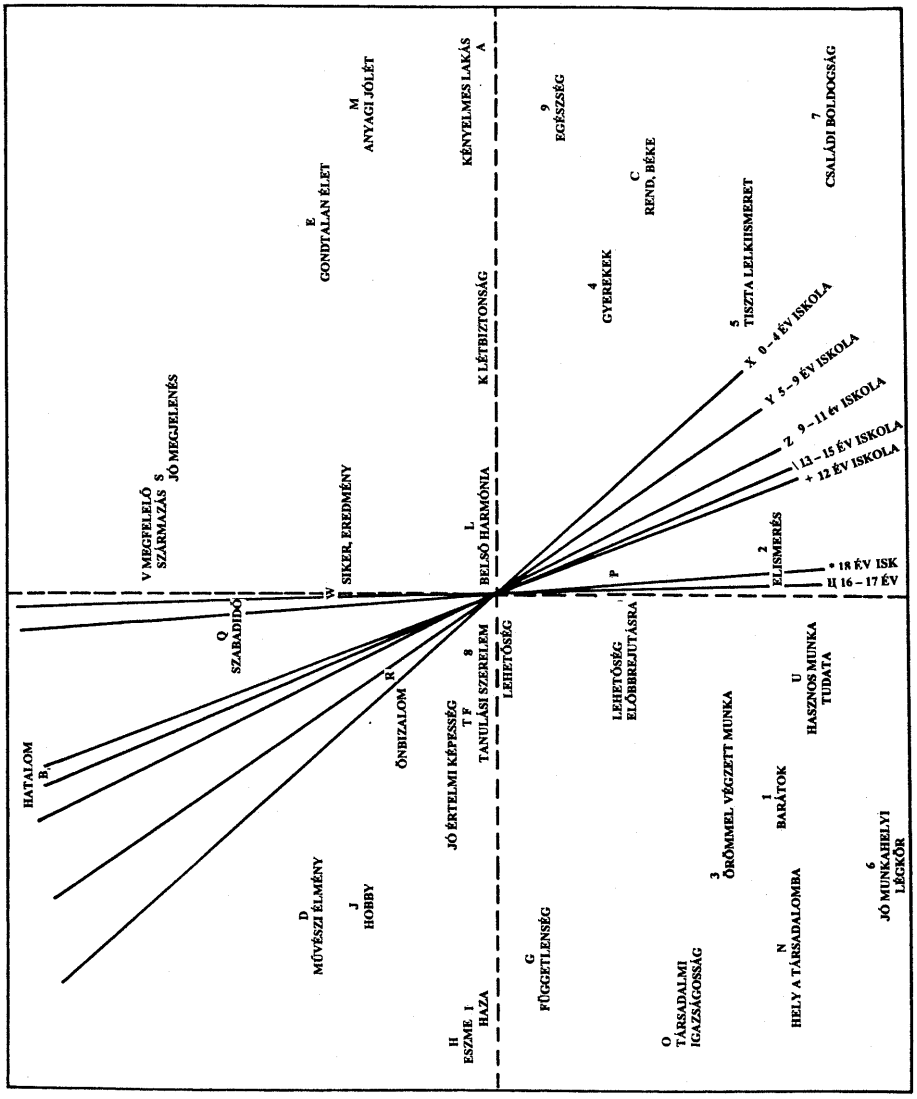
14.25. táblázat. Az alapértékek szemantikai tere kétdimenziós MINISSA-megoldásának outputja

A négy jövedelmi csoport értékválasztás szerinti illesztése az értékek szemantikai terében azt mutatja, hogy jelentős különbség van a 3000 Ft alattiak és felettiek kategóriái között. Ha az értékeket rávetítjük a kategóriákhoz tartozó tengelyekre, az értékválasztás jelentős eltéréseit figyelhetjük meg a kapott rangsorok alapján. Ezeket az összehasonlításokat bárki könnyen leolvashatja a következő ábráról.



14.46. ábra. Négy jövedelmi kategória illesztése az értékek szemantikai terében. PROFIT-megoldás (lineáris modell)

Az iskolai végzettség szerinti csoportok között két helyen találhatunk nagy ugrást: az általános iskolát végzettek és ennél többet végzettek csoportjai között, valamint a középiskolát (főiskolát) és egyetemet végzettek között. A következő ábrában az iskolai végzettség szerinti tengelyekre vetítjük az egyes értékeket, és ezen rangsorok eltérései alapján az olvasó az iskolai végzettség szerinti társadalmi csoportok értékválasztásának jelentős különbségeit veheti észre.



14.47. ábra. Iskolai végzettség szerinti csoportok az értékek szemantikai terében.
PROFIT-megoldás (lineáris modell)

15. fejezet

Korreszpondencia-modell

A korreszpondenciaelemzés egy adatmátrix sorait és oszlopait mint pontokat illeszti egy alacsony dimenziószámú térbe. A leggyakrabban a mérési eredményeket egy kontingenciatáblázatba rendezzük, és a gyakorisági táblázatból indulunk ki. A korreszpondenciaelemzés két vektorteret keres, egyet a gyakorisági táblázat sorainak, egyet pedig az oszlopainak reprezentálására, grafikus ábrázolására. Az elméletet Roberts (1981) adatai alapján mutatjuk be. A következő táblázat a bőrrák egyik típusának (malignant melanoma) vizsgálatából származik, négyszáz beteg megoszlását mutatja a tumor elhelyezkedése (fej, nyak, törzs, végtagok) és történeti típusai (Hutchison's melatonic freckle, Superficial spreading melanoma, Nodular, Interminate) szerint:

	A tumor elhelyezkedése		
	Fej, nyak	Törzs	Végtagok
Történeti típusok	(*h*)	(*t*)	(*e*)
Hutchison's melatonic freckle (*H*)	22	2	10
Superficial spreading melanoma (*S*)	16	54	115
Nodular (*N*)	19	33	73
Interminate (*I*)	11	17	28

15.1. táblázat.

A fenti adatokat a (4×3) típusú $\mathbf{X}$ mátrixba helyezzük. A korreszpondenciaelemzés elvégzése előtt (az összehasonlítás lehetősége miatt) végezzük el az $\mathbf{X}$ mátrix szinguláris érték felbontását (Singular Value Decomposition; SVD):

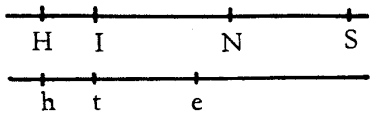
$$\mathbf{X} = \mathbf{A}\mathbf{D}_\lambda\mathbf{B}'.$$

$$\mathbf{X} = \begin{bmatrix} 0{,}087 & 0{,}906 & 0{,}221 \\ 0{,}818 & -0{,}292 & 0{,}109 \\ 0{,}526 & 0{,}215 & 0{,}187 \\ 0{,}217 & 0{,}219 & -0{,}951 \end{bmatrix} \begin{bmatrix} 156{,}369 & 0 & 0 \\ 0 & 22{,}140 & 0 \\ 0 & 0 & 4{,}083 \end{bmatrix} \times$$

$$\times \begin{bmatrix} 0{,}175 & 0{,}418 & 0{,}891 \\ 0{,}982 & -0{,}144 & -0{,}125 \\ -0{,}075 & -0{,}897 & 0{,}436 \end{bmatrix}$$

$$\mathbf{X} = 156{,}369 \begin{bmatrix} 0{,}015 & 0{,}036 & 0{,}078 \\ 0{,}143 & 0{,}342 & 0{,}729 \\ 0{,}092 & 0{,}220 & 0{,}469 \\ 0{,}038 & 0{,}091 & 0{,}194 \end{bmatrix} + 22{,}140 \begin{bmatrix} 0{,}889 & -0{,}130 & -0{,}114 \\ -0{,}287 & 0{,}042 & 0{,}037 \\ 0{,}211 & -0{,}031 & -0{,}027 \\ 0{,}215 & -0{,}031 & -0{,}027 \end{bmatrix}$$
$$+ 4{,}083 \begin{bmatrix} -0{,}017 & -0{,}198 & 0{,}096 \\ -0{,}008 & -0{,}098 & 0{,}048 \\ -0{,}014 & -0{,}167 & 0{,}081 \\ 0{,}072 & 0{,}853 & -0{,}414 \end{bmatrix}$$

Mivel az első szinguláris érték kétszer nagyobb, mint a második, egy dimenzió is elégséges a tumor típusának, és egy a helyének a reprezentálására. Az első bal oldali szinguláris vektor elemei a tumor típusát reprezentáló pontok H, I, N, S sorrendben, az első jobb oldali szinguláris vektor a tumor elhelyezkedéséhez tartozó pontok koordinátái (h, t, e sorrendben).



15.1. ábra. A tumoradatok SVD felbontása egy dimenzióban.
Az első tengely a történeti típus, a második a tumor elhelyezkedése

15.1. A korreszpondencia-modell

Tekintsük az $\mathbf{X}$ $(n \times p)$ típusú mátrixot (a gyakorlatban legtöbbször kontingenciatáblázatot). Két vektorteret keresünk, az egyik az $\mathbf{X}$ mátrix sorait, a másik az oszlopait reprezentálja.

Az $\mathbf{X}$ mátrix i-edik sorprofilja a mátrix i-edik sora a sorösszeggel standardizálva (r_i az i-edik sor összege, súlya). Hasonlóan az $\mathbf{X}$ mátrix j-edik oszlopprofilja a mátrix j-edik oszlopa standardizálva az oszlopösszeggel (c_j a j-edik oszlop összege, súlya).

Az $\mathbf{X}$ mátrix sorprofiljai:

$$\mathbf{D}_r^{-1}\mathbf{X}, \quad \text{ahol} \quad \mathbf{D}_r = \text{diag}(r_1, \ldots, r_n).$$

Az $\mathbf{X}$ mátrix oszlopprofiljai:

$$\mathbf{D}_c^{-1}\mathbf{X}', \quad \text{ahol} \quad \mathbf{D}_c = \text{diag}(c_1, \ldots, c_p).$$

A sor- és oszlopprofilokat két vektortérben reprezentáljuk az $\mathbf{X}$ mátrix szinguláris érték felbontásának általánosításával.

Tekintsük az $\mathbf{X}$ mátrix általánosított szinguláris érték felbontását:

$$\mathbf{X} = \mathbf{A}\mathbf{D}_\lambda\mathbf{B}', \quad \text{ahol} \quad \mathbf{A}'\mathbf{D}_r^{-1}\mathbf{A} = \mathbf{B}'\mathbf{D}_c^{-1}\mathbf{B} = \mathbf{I}.$$

Az $\mathbf{A}$ mátrix ortonormált bázisa az $\mathbf{X}$ oszlopvektorainak, normalizálva a $\mathbf{D}_r^{-1}$-vel. A $\mathbf{B}$ mátrix ortonormált bázisa az $\mathbf{X}$ sorainak.

A sorprofilok a szinguláris érték dekompozíció felhasználásával a következőképpen írhatók fel:

$$\mathbf{D}_r^{-1}\mathbf{X} = \mathbf{D}_r^{-1}\mathbf{A}\mathbf{D}_\lambda\mathbf{B}',$$

feltéve, hogy

$$(\mathbf{D}_r^{-1}\mathbf{A})'\mathbf{D}_r^{-1}(\mathbf{D}_r^{-1}\mathbf{A}) = \mathbf{B}'\mathbf{D}_c^{-1}\mathbf{B} = \mathbf{I}.$$

Legyen $\mathbf{U} = \mathbf{D}_r^{-1}\mathbf{A}$, így

$$\mathbf{D}_r^{-1}\mathbf{X} = \mathbf{U}\mathbf{D}_\lambda\mathbf{B}', \quad \text{és} \quad \mathbf{U}'\mathbf{D}_r^{-1}\mathbf{U} = \mathbf{B}'\mathbf{D}_c^{-1}\mathbf{B} = \mathbf{I}. \tag{15.1}$$

Hasonlóan az oszlopprofilokat is kifejezhetjük a következőképpen:

$$\mathbf{D}_c^{-1}\mathbf{X}' = \mathbf{D}_c^{-1}\mathbf{B}\mathbf{D}_\lambda\mathbf{A}',$$

feltételezve, hogy

$$\mathbf{A}'\mathbf{D}_r^{-1}\mathbf{A} = (\mathbf{D}_c^{-1}\mathbf{B})'\mathbf{D}_c(\mathbf{D}_c^{-1}\mathbf{B}) = \mathbf{I}.$$

Legyen $\mathbf{V} = \mathbf{D}_c^{-1}\mathbf{B}$, ekkor

$$\mathbf{D}_c^{-1}\mathbf{X}' = \mathbf{V}\mathbf{D}_\lambda\mathbf{A}', \quad \text{és} \quad \mathbf{A}'\mathbf{D}_r^{-1}\mathbf{A} = \mathbf{V}'\mathbf{D}_c\mathbf{V} = \mathbf{I}. \tag{15.2}$$

A (15.1) egyenlet azt mutatja, hogy a sorprofilokat az $\mathbf{U}\mathbf{D}_\lambda$ térben reprezentáljuk, ahol a sorprofilokat a $\mathbf{B}$ rotációs mátrix transzformálja a tér pontjaiba. A gyakorlatban az $\mathbf{U}\mathbf{D}_\lambda$ első k oszlopvektorát, mint a k dimenziós tér legkisebb négyzetek módszere szerinti legjobb becslését tekintjük a sorprofilok reprezentálására lehetőleg alacsony dimenzió-számú térben.

Hasonlóan mutatja a (15.2) egyenlet az oszlopprofilok reprezentálását a $\mathbf{V}\mathbf{D}_\lambda$ térben az $\mathbf{A}$ rotáció után.

Példa

A tumor-példában az $\mathbf{X}$ mátrixot normalizáljuk a minta elemszámával, 400-zal. Az $\mathbf{X}$ mátrix SVD felbontása:

$$\mathbf{X} = \begin{bmatrix} 0{,}055 & 0{,}005 & 0{,}025 \\ 0{,}040 & 0{,}135 & 0{,}288 \\ 0{,}048 & 0{,}083 & 0{,}183 \\ 0{,}028 & 0{,}043 & 0{,}070 \end{bmatrix}$$

$$\mathbf{D}_r = \text{diag}[0{,}085,\quad 0{,}463,\quad 0{,}313,\quad 0{,}140]$$

$$\mathbf{D}_c = \text{diag}[0{,}170,\quad 0{,}265,\quad 0{,}565]$$

$$\mathbf{D}_r^{-1}\mathbf{X} = \begin{bmatrix} 0{,}647 & 0{,}059 & 0{,}294 \\ 0{,}086 & 0{,}292 & 0{,}622 \\ 0{,}152 & 0{,}264 & 0{,}584 \\ 0{,}196 & 0{,}304 & 0{,}500 \end{bmatrix}$$

$$\mathbf{D}_c^{-1}\mathbf{X}' = \begin{bmatrix} 0{,}324 & 0{,}235 & 0{,}279 & 0{,}162 \\ 0{,}019 & 0{,}509 & 0{,}311 & 0{,}160 \\ 0{,}044 & 0{,}509 & 0{,}323 & 0{,}124 \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} 0{,}085 & 0{,}269 & -0{,}050 \\ 0{,}463 & -0{,}255 & -0{,}166 \\ 0{,}313 & -0{,}036 & -0{,}131 \\ 0{,}140 & 0{,}021 & 0{,}346 \end{bmatrix} \begin{bmatrix} 1{,}0 & 0 & 0 \\ 0 & 0{,}403 & 0 \\ 0 & 0 & 0{,}047 \end{bmatrix} \times$$

$$\times \begin{bmatrix} 0{,}170 & 0{,}265 & 0{,}565 \\ 0{,}374 & -0{,}153 & -0{,}222 \\ 0{,}029 & 0{,}414 & -0{,}443 \end{bmatrix}$$

$$\mathbf{U} = \begin{bmatrix} 1 & 3{,}167 & -0{,}591 \\ 1 & -0{,}550 & -0{,}358 \\ 1 & -0{,}116 & -0{,}418 \\ 1 & 0{,}153 & 2{,}474 \end{bmatrix}$$

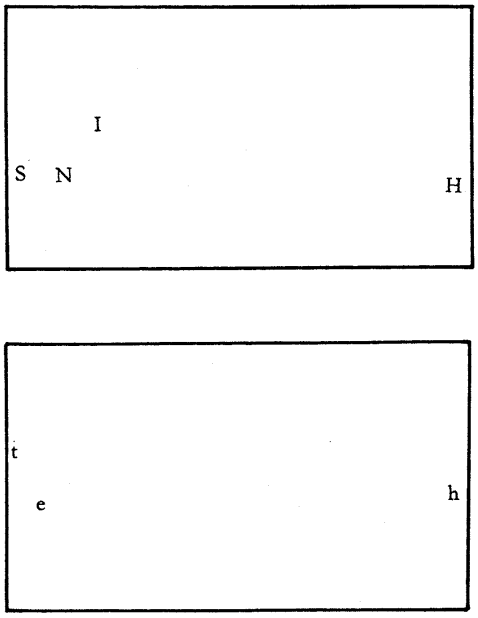
$$\mathbf{V} = \begin{bmatrix} 1 & 2{,}203 & 0{,}172 \\ 1 & -0{,}576 & 1{,}563 \\ 1 & -0{,}393 & -0{,}785 \end{bmatrix}$$

$$\mathbf{UD}_\lambda = \begin{bmatrix} 1 & 1{,}276 & -0{,}023 \\ 1 & -0{,}222 & -0{,}017 \\ 1 & -0{,}047 & -0{,}020 \\ 1 & 0{,}062 & 0{,}116 \end{bmatrix}$$

$$\mathbf{VD}_\lambda = \begin{bmatrix} 1 & 0{,}888 & 0{,}008 \\ 1 & -0{,}232 & -0{,}073 \\ 1 & -0{,}158 & -0{,}037 \end{bmatrix}$$

A példában láthatjuk, hogy az első szinguláris érték 1-gyel egyenlő, és a neki megfelelő vektor $\mathbf{1}$.

Ez azért van, mivel $\mathbf{D}_r^{-1}\mathbf{X1} = \mathbf{1}$ (a $\mathbf{D}_r$ a sorösszegeket tartalmazza). Hasonlóan $\mathbf{D}_c^{-1}\mathbf{X}'\mathbf{1} = \mathbf{1}$.



15.2–15.3. ábra. A tumoradatok korreszpondencia-elemzése
Az első tér a történeti típus, a második a tumor elhelyezkedése

A triviális első szinguláris érték és vektor elhagyása miatt a

$$\mathbf{D}_r^{-1}\mathbf{X} - \mathbf{1c}' \quad \text{és} \quad \mathbf{D}_c^{-1}\mathbf{X}' - \mathbf{1r}'$$

mátrixokat elemezzük.

A 15.2–15.3. ábrák a triviális dimenzió elhagyása után az $\mathbf{UD}_\lambda$ szinguláris vektorait – a tumor típusát reprezentálva – és a $\mathbf{VD}_\lambda$ szinguláris vektorokat – a tumor elhelyezkedését – mutatják.

15.2. Az inercia

A sorprofilok szóródásának mértéke a teljes inercia (teljes nyomaték). A teljes inerciát kifejezhetjük a kontingenciatáblázatoknál a függetlenség feltételezésével számított χ^2 statisztika segítségével:

$$I = N^{-1}\chi^2 = N^{-1}\sum_i\sum_j \frac{\left(x_{ij} - \frac{x_{i+}x_{+j}}{N}\right)^2}{\frac{x_{i+}x_{+j}}{N}},$$

ahol x_{i+} az x_{+j} a sor- és oszlopösszeget jelöli.

Másképpen írva:

$$N^{-1}\chi^2 = \sum_i \left\{ x_{i+} \sum_j \frac{\left(\frac{x_{ij}}{x_{i+}} - \frac{x_{+j}}{N}\right)^2}{x_{+j}} \right\} =$$

$$= \sum_i r_i(\mathbf{r}_i - \mathbf{c})'\mathbf{D}_c^{-1}(\mathbf{r}_i - \mathbf{c}) = I,$$

ahol r_i az i-edik sor összege,

$\mathbf{r}_i$ az i-edik sorprofil (vektor)

$\mathbf{c}$ „átlagos sorprofil", az oszlopösszegek profilja.

Látható, hogy I, a sorprofilok teljes inerciája a sorprofilok és az „átlagos sorprofil" közötti súlyozott távolság négyzetek súlyozott összege.

Hasonlóan az oszlopprofilok teljes inerciája:

$$I = c_j(\mathbf{c}_j - \mathbf{r})'\mathbf{D}_r^{-1}(\mathbf{c}_j - \mathbf{r}),$$

ami egyenlő a sorprofilok teljes inerciájával a χ^2 statisztika szimmetrikussága miatt.

A teljes inerciát a következőképpen is írhatjuk:

$$I = tr\left(\mathbf{D}_r\left(\mathbf{D}_r^{-1}\mathbf{X} - 1\mathbf{c}'\right)\mathbf{D}_c^{-1}\left(\mathbf{D}_r^{-1}\mathbf{X} - \mathbf{1c}'\right)'\right),$$

ahol $\left(\mathbf{D}_r^{-1}\mathbf{X} - \mathbf{1c}'\right)$ a sorprofil a triviális megoldás kivonásával.

Helyettesítsük $\left(\mathbf{D}_r^{-1}\mathbf{X} - \mathbf{1c}'\right)$-t egyszerűen $\mathbf{D}_r^{-1}\mathbf{X}$-szel feltételezve, hogy a triviális megoldást már kivontuk.

A teljes inercia ekkor:

$$\begin{aligned} I &= tr\left(\mathbf{D}_r\left(\mathbf{D}_r^{-1}\mathbf{X}\right)\mathbf{D}_c^{-1}\left(\mathbf{D}_r^{-1}\mathbf{X}\right)'\right) \\ &= tr\left(\mathbf{A}\mathbf{D}_\lambda\mathbf{B}'\right)\mathbf{D}_c^{-1}(\mathbf{B}\mathbf{D}_\lambda\mathbf{A}')\mathbf{D}_r^{-1} \\ &= tr\left(\mathbf{A}\mathbf{D}_\lambda^2\mathbf{A}'\mathbf{D}_r^{-1}\right) = tr\left(\mathbf{D}_\lambda^2\mathbf{A}'\mathbf{D}_r^{-1}\mathbf{A}\right) \\ &= tr(\mathbf{D}_\lambda^2). \end{aligned}$$

A teljes inercia egyenlő a szinguláris értékek négyzetösszegével.

A sor- és oszlopprofilok reprezentálására szükséges dimenziószámot megítélhetjük úgy, hogy k dimenzió milyen arányban járul hozzá a teljes inerciához:

$$\sum_{i}^{k} \lambda_i^2 \Big/ \sum_{i}^{n} \lambda_i^2,$$

ahol n a nem egységnyi szinguláris értékek száma.

A példánkban a teljes inercia értéke = 0,1645, és az első dimenzió 98,6%-ban járul ehhez hozzá.

15.3. A többszörös korreszpondencia-modell

A korreszpondenciaelemzést a gyakorlatban legtöbbször kétdimenziós kontingenciatáblázatok elemzésére alkalmazzuk, feltételezve azt, hogy $\mathbf{X}$ adatmátrix minden eleme nemnegatív.

A korreszpondenciaelemzést kiterjeszthetjük három- és többdimenziós kontingenciatáblázatok elemzésére, bevezetve egy ún. indikátor-változót. Az indikátor-változó segítségével konvertáljuk a többdimenziós táblázatot kétdimenziós táblázattá. Tételezzük fel, hogy a k-dimenziós táblázatban az i-edik szempont (változó) kategóriáinak a száma c_i. Egy indikátor-változót jelölünk ki a táblázat minden változó (dimenzió, szempont) minden kategóriájához, ezzel $J = \sum_{i}^{k} c_i$ számú dummy, 0 és 1 értékű indikátor-változót definiálunk.

Minden megfigyelési egységhez (N) J számú indikátor-változó tartozik, így egy ($N \times J$) típusú, ún. indikátormátrixot határozunk meg. Az indikátormátrix minden sora (megfigyelési egység) k egyest és $J - k$ nulla értéket tartalmaz.

A mintapéldában 400 személy két változó összesen 7 kategóriájához tartozó adatait egy (400×7) típusú indikátormátrixba rendezhetjük. Az indikátor-változók $I_1 = H$, $I_2 = S$, $I_3 = N$, $I_4 = I$, $I_5 = h$, $I_6 = t$, $I_7 = e$. Az indikátormátrix első 22 sora a következő lenne: [1, 0, 0, 0, 1, 0, 0].

A 15.4. ábra a mintapélda korreszpondenciaelemzését mutatja az indikátormátrix alkalmazásával. Az ábrán látható, hogy a tumor négy típusa és három testrészen való elhelyezkedése hasonló eredményt ad, mint amit a korábbi elemzésnél kaptunk, azzal a különbséggel, hogy a két tengely nincs súlyozva a szinguláris értékeivel. Általában a két módszer sajátértékei között a következő kapcsolat van:

$$\rho = (2\rho_I - 1)^2,$$

ahol ρ az eredti adatmátrix (kontingenciatáblázat) sajátértéke, ρ_I az indikátormátrix sajátértéke.

Egy k-dimenziós kontingenciatáblázat indikátormátrixa:

$$\mathbf{Z} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_k],$$

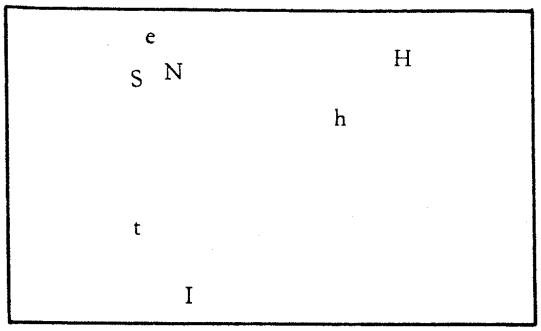
ahol $\mathbf{Z}_i$ egy ($N \times c_i$) típusú adatmátrix, N megfigyelési egység c_i indikátor-változóra vonatkozó megfigyelt értéke.

A $\mathbf{B} = \mathbf{Z}'\mathbf{Z}$ mátrix a Burt-mátrix:

$$\mathbf{B} = \begin{bmatrix} \mathbf{Z}_1'\mathbf{Z}_1 & \mathbf{Z}_1'\mathbf{Z}_2 & \dots & \mathbf{Z}_1'\mathbf{Z}_k \\ \mathbf{Z}_2'\mathbf{Z}_1 & \mathbf{Z}_2'\mathbf{Z}_2 & \dots & \mathbf{Z}_2'\mathbf{Z}_k \\ \vdots & \vdots & & \\ \mathbf{Z}_k'\mathbf{Z}_1 & \mathbf{Z}_k'\mathbf{Z}_2 & \dots & \mathbf{Z}_k'\mathbf{Z}_k \end{bmatrix},$$

ahol a $\mathbf{Z}'_i\mathbf{Z}_j$ partíció az i-edik és j-edik változó kétdimenziós kontingenciatáblázata, $\mathbf{Z}_i'\mathbf{Z}_i$ a diagonálisban elhelyezkedő partíció, diagonális mátrix, diagonális elemei az oszlop-összegek (az i-edik változó perem gyakoriságai).

A Burt-mátrix szimmetrikus, pozitív definit.



15.4. ábra. A tumoradatok korreszpondencia-elemzése az indikátor-mátrix alapján

15.4. Példa a többszörös korreszpondencia-modellre

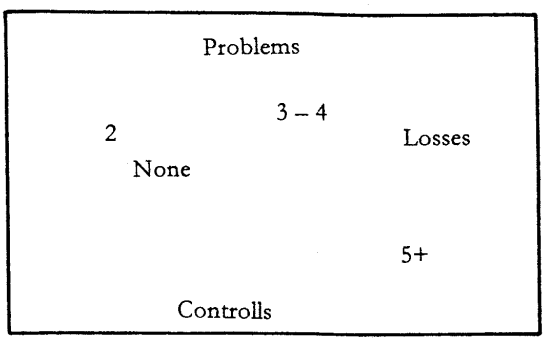
(Gyermekhalandóság vizsgálata)

Az adatok Plackett (1981) könyvében a gyermekhaladóságra (pl. halvaszületés), a születés sorszámára és a családban előfordult korábbi problémára vonatkoznak.

Az anyák száma	Születés sorszáma					
	2		3–4		5+	
	Probléma	Kontroll	Probléma	Kontroll	Probléma	Kontroll
Elvesztett	20	10	26	16	27	14
Nem	82	54	41	30	22	23

15.2. táblázat. Gyermekhalandóság kapcsolata a születés sorrendjével és a családban korábban előfordult gyermekkel kapcsolatos problémával

A Burt-mátrixot a táblázatból meghatározva a többszörös korreszpondenciaelemzés eredményét a 15.5. ábra mutatja. Plackett azt állapította meg, hogy a gyermekhalandóságra csak a gyermek születésének a sorszáma hat. A 15.5. ábra ezzel megegyező eredményt mutat, mivel a „Probléma/Controll” tengely közel merőleges az „Elvesztett/Nem” tengellyel. A születés sorrendje még szorosabb kapcsolatban lenne a gyermekhalandósággal, ha az „5+” kategória közelebb esne az „Elvesztett” kategóriához.



15.5. ábra. Korreszpondenciaelemzés a gyermekhaladósági adatokra

15.5. Példa a korreszpondencia-modellre *(Rokeach-értéktípusok és az iskolai végzettség kapcsolata)*

Az MTA Szociológiai Kutatóintézet Értékszociológiai Műhelyében a Rokeach-értéktesztet vizsgáltuk országos reprezentatív mintán a következő években:

Évek	Az érvényes esetek száma
1978	676
1982	2080
1990	1063
1993	1146
1996	1296

A nemhierarchikus klaszterelemzés McQueen-féle eljárásával határoztuk meg az értéktípusokat.

Az értéktípusok összegző leírása

CL1 Felvilágosult racionalista (hivatalos szocialista)
CL2 Poszt-vallásos
CL3 Klasszikus szociáldemokrata
CL4 Poszt-felvilágosult
CL5 Vallásos
CL6 Hedonista
CL7 Materialista (poszt-kommunista)
CL8 Tradicionális-fegyelmező
CL9 Ideológiai szociáldemokrata
CL10 Ideológiai hedonista

Az értéktípusok iskolai végzettség szerinti megoszlása adja az induló táblázatunkat:
Korreszpondencia-táblázat

spaceRokeach-klaszterek 5 év együttesére	Legmagasabb iskolai végzettsége?				
	kevesebb, mint 8 osztály	általános iskola (8 osztály)	közép-iskola	főiskola, egyetem	Összesen
Felvilágosult racionalista	79	322	267	154	822
Poszt-vallásos	156	342	51	12	561
Klasszikus szociáldemokrata	61	154	92	29	336
Poszt-felvilágosult	26	121	210	182	539
Vallásos	245	393	140	50	828
Hedonista	64	339	228	49	680
Materialista	69	406	268	105	848
Tradicionális-fegyelmező	96	259	105	15	475
Ideológiai szociáldemokrata	80	286	129	59	554
Ideológiai hedionista	92	280	154	38	564
Összesen	968	2902	1644	693	6207

A korreszpondencia-modell illesztésének eredményét tartalmazza a következő táblázat és ábra.

Summary

Dimension	Singular Value	Inertia	Chi Square	Sig.	Proportion of Inertia		Confidence Singular Value	
					Accounted for	Cumulative	Standard Deviation	Correlation 2
1	,357	,128			,836	,836	,012	,282
2	,148	,022			,143	,979	,013	
3	,057	,003			,021	1,000		
Total		,153	947,923	,000[a]	1,000	1,000		

a. 27 degrees of freedom

Az eredmények azt mutatják, hogy a kétdimenziós térben jól tudjuk reprezentálni az érték-típusokat és az iskolai végzettséget. Az együttes térben a különböző iskolai végzettség és értékrendszer kapcsolódások egyértelműen mutatkoznak.

A főiskolai vagy egyetemi végzettség a „Poszt-felvilágosult" értéktípushoz tartozik, ahol olyan értékeket preferálnak, mint (Kurziváltuk azokat az értékeket, amelyek különösen jellemzőek az adott értéktípusra.):

Típus	**Jóval az átlag fölött preferált értékek**	**Jóval az átlag alatt preferált értékek**
CL4	*Belső harmónia*, bölcsesség, szerelem, emberi önérzet, a szépség világa, *logikus, alkotó szellemű, értelmes,* előítéletektől mentes	*A haza biztonsága*, béke, egyenlőség, anyagi jólét, *engedelmes*, megbocsátó, tiszta, udvarias, törekvő

A középiskolások értékrendje a „Materialista" értékrendhez áll közel:

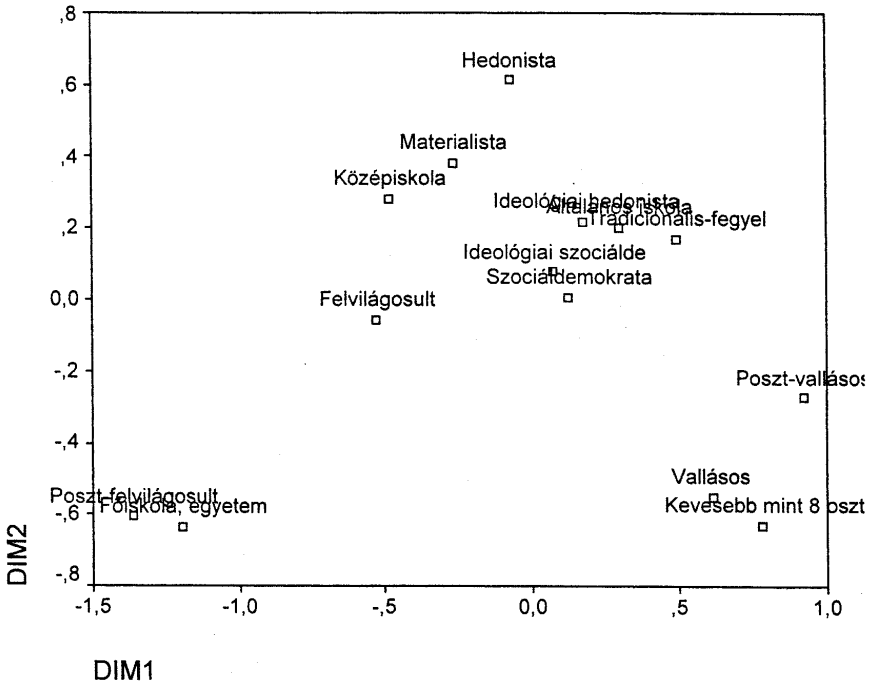
Típus	**Jóval az átlag fölött preferált értékek**	**Jóval az átlag alatt preferált értékek**
CL7	*Anyagi jólét*, boldogság, a haza biztonsága, *önálló*, hatékony, alkotó szellemű, logikus, értelmes, béke, szerelem	*Egyenlőség, megbocsátó, segítőkész*, engedelmes, udvarias, szeretettel teljes, szabadság, emberi önérzet, társadalmi megbecsülés

Az általános iskolát végzettek értékrendje két értéktípushoz áll közel:

Típus	Jóval az átlag fölött preferált értékek	Jóval az átlag alatt preferált értékek
CL10	*Béke*, a haza biztonsága, szabadság, anyagi jólét, *jókedélyű, tiszta*, udvarias, szeretettel teljes	Társadalmi megbecsülés, belső harmónia, munka öröme, egyenlőség, *felelősségteljes, fegyelmezett*, engedelmes
CL8	*Boldogság*, családi biztonság, béke, szerelem, a haza biztonsága, engedelmes, tiszta, törekvő, segítőkész, szeretettel teljes	*Társadalmi megbecsülés, hatékony, alkotó szellemű*, értelmes, a szépség világa, barátság, emberi önérzet

A 8 általánosnál kevesebb iskolát végzettek a „Vallásos" értékrendhez tartoznak leginkább:

Típus	Jóval az átlag fölött preferált értékek	Jóval az átlag alatt preferált értékek
CL5	*Üdvözülés, megbocsátó, szeretettel teljes*, engedelmes, segítőkész, udvarias, tiszta, jókedélyű, szavahihető, béke	Változatos élet, szabadság, *alkotó szellemű, önálló, logikus*, bátor, hatékony, előítéletektől mentes



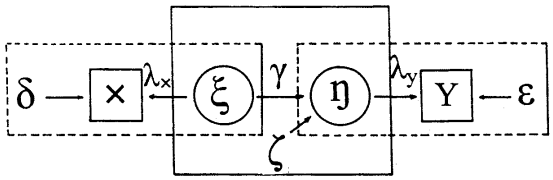
15.6. ábra. Az értéktípusok és iskolai végzettség kategóriáinak együttes ábrája

16. fejezet

LISREL-modell

(Analysis of Linear StructuraR elationships by the Method of Maximum Likelihood)

Tekintsük először az általános modell szkémáját:



ahol X: a megfigyelt, manifeszt független (exogén) változók halmaza, Y: a megfigyelt, manifeszt függő (endogén) változók halmaza, ξ: nem megfigyelt, latens független (exogén) változó(k), η: nem megfigyelt, latens függő (endogén) változó(k), λ_x: a független változók faktorsúlya(i), λ_y: a függő változók faktorsúlya(i), δ: a megfigyelt független változók mérési hibája, ε: a megfigyelt függő változók mérési hibája, ζ: a latens függő változó(k) sztochasztikus reziduális tagja(i).

Az általános modellen belül megkülönböztetjük:

1. *A mérési modelleket*, amelyek azt fejezik ki, hogy a manifeszt változók két komponensre bonthatók; az egyik a szisztematikus komponens, amit a modellben a latens változó fejez ki, a másik a mérési hiba.

A független és függő változókra külön-külön írjuk fel a mérési modellt. (Az ábrán szaggatott vonallal jelöljük.)

2. *A strukturális modellt*, amely a latens változók kauzális összefüggéseit írja le (az általános modellben a szaggatott vonallal körülhatárolt mérési modellek között helyezkedik el).

Az általános modell feltételei kétfélék:

a) rendszerfeltételek, a modell becsléséhez szükséges általános feltételek,

b) modellfeltételek, amelyekkel speciális modelleket definiálhatunk.

Az általános modell két latens változó halmaz struktúráját írja le azzal a feltételezéssel, hogy a latens változók nem mérhetők, hanem mögöttes kifejezői a közvetlenül mérhető manifeszt változóknak, vagyis az egymással kauzális kapcsolatban lévő latens változók mögöttes okai a (mérési modellekben hozzájuk rendelt) megfigyelt változóknak.

A strukturális egyenletekben játszott kauzális kapcsolódásaik alapján a változókat függőnek nevezzük, ha az okozat szerepét, és függetlennek, ha az ok szerepét játsszák. Szokásos ehelyett az eredmény- és a magyarázó változó elnevezések használata is.

A változók egy másik osztályozása az, amikor nem egy-egy strukturális egyenletben, hanem az egész modellben játszott kauzális szerepüket ítéljük meg. Eszerint a változók kétfélék, endogén és exogén változók. A strukturális egyenletekben az exogén változók

csak az ok szerepét játszhatják, és az okozatok az endogén változók. Az endogén változók között azonban kölcsönös kauzális kapcsolatok is lehetnek, de az exogén változók között ezt nem engedjük meg. A modellt szimultánnak nevezzük, ha az endogén változók között kölcsönös kauzális kapcsolatok vannak (mind az ok, mind az okozat szerepét betöltik), a modell rekurzív, ha az endogén változók ugyan lehetnek okok és okozatok is, de közöttük kölcsönös kauzális összefüggés nincsen.

Legyen a latens függő változók vektora $\eta' = [\eta_1, \eta_2, \ldots, \eta_m]$ és a latens független változók vektora $\xi' = [\xi_1, \xi_2, \ldots, \xi_n]$.

A strukturális egyenletek modellje:

$$\mathbf{B}\eta = \Gamma\xi + \zeta, \tag{16.1}$$

ahol

$\mathbf{B}$: a függő latens változók közötti regressziós együtthatók mátrixa, $(m \times m)$ típusú,

Γ: a független latens változók és a függő latens változók közötti regressziós együtthatók mátrixa, $(m \times n)$ típusú,

$\zeta' = [\zeta_1, \zeta_2, \ldots, \zeta_m]$ a függő latens változók reziduális komponense (sztochasztikus reziduális tag).

A modellben szereplő változókról feltételezzük, hogy az átlaguktól való eltéréseket tartalmazzák, és a véletlen sztochasztikus tagok korrelálatlanok egymással és a független változókkal, valamint hogy a $\mathbf{B}$ mátrix nem szinguláris:

$$E(\eta) = E(\zeta) = \mathbf{0} \quad \text{és} \quad E(\xi) = \mathbf{0},$$
$$E(\xi\zeta') = \mathbf{0} \quad \text{és} \quad \exists\mathbf{B}^{-1}.$$

A mérési modellek:

a) a függő megfigyelt, manifeszt (endogén) változók $\mathbf{y}' = [y_1, y_2, \ldots, y_p]$ mérési modellje:

$$\mathbf{y} = \Lambda_{\mathbf{y}}\eta + \epsilon, \tag{16.2}$$

ahol

Λ_y: $(p \times m)$ típusú mátrix a függő latens változók hatását fejezi ki a függő megfigyelt változóra (faktorsúlyok mátrixa),

ϵ: a függő változók mérési hibája.

Feltételezzük a fentiekhez hasonlóan, hogy

$$E(\mathbf{y}) = E(\epsilon) = \mathbf{0} \quad \text{és} \quad E(\eta\epsilon') = \mathbf{0}.$$

b) a független megfigyelt, manifeszt (exogén) változók $\mathbf{x}' = [x_1, x_2, \ldots, x_q]$ mérési modellje:

$$\mathbf{x} = \Lambda_x\xi + \delta, \tag{16.3}$$

ahol

Λ_x: $(q \times n)$ típusú mátrix a független latens változók hatását fejezi ki a független megfigyelt változókra (faktorsúlyok mátrixa),

δ: a független változók mérési hibája.

Tekintsük át ezután a modell változóit és a változók kapcsolódásait:

Változók	függő (endogén)	független (exogén)	eltérés (reziduális)
Megfigyelt (manifeszt)	$\mathbf{y}$	$\mathbf{x}$	–
Nem megfigyelt (latens)	η	ξ	ζ, ϵ, δ

16.1. táblázat. A változók különböző fajtái az általános modellben

Változók	függő (endogén)	független (exogén)	eltérés (reziduális)
Változók	függő (endogén)	független (exogén)	eltérés (reziduális)
	η	ξ	$\zeta\epsilon\delta$
Nem megfigyelt η	β	Γ	**I00**
Megfigyelt			
$\mathbf{y}$	Λ_y	**0**	**0I0**
$\mathbf{x}$	**0**	Λ_x	**00I**

16.2. táblázat. A közvetlen hatások lehetséges irányai az általános modellben

Az 16.2. táblázat az általános modell közvetlen hatásait mutatja. Az első sora a következő egyenletet adja:

$$\eta = \beta\eta + \Gamma\xi + \mathbf{I}\zeta. \tag{16.4}$$

Ezt az egyenletet átrendezhetjük:

$$(\mathbf{I} - \beta)\eta = \Gamma\xi + \zeta,$$

vagy a (16.1) egyenletben szereplő formában:

$$\mathbf{B}\eta = \Gamma\xi + \zeta,$$

ahol $\mathbf{B} = (\mathbf{I} - \beta)$, $\beta = (\mathbf{I} - \mathbf{B})$.

A fentiek alapján a (16.1) egyenletben szereplő **B** együtthatómátrixból a közvetlen hatások mátrixát úgy kapjuk meg, hogy vesszük az együtthatókat ellenkező előjellel, a diagonális elemek pedig egyenlők lesznek 0-val, (mivel a **B** mátrix diagonális elemei 1-gyel egyenlők).

Az első három egyenlethez tartozó feltételek mellett a következőkben olyan feltételeket említünk, amelyeket minden modell esetén érvényesnek tekintünk, és azért *rendszerfeltételeknek* nevezzük őket:

1) a mérési hibák és a sztochasztikus reziduális tag között a kovariancia (korreláció) nulla:

$$E(\zeta\epsilon') = \mathbf{0} \qquad E(\zeta\delta') = \mathbf{0},$$

2) a két mérési modell hibája között a kovariancia nulla:

$$E(\epsilon\delta') = \mathbf{0},$$

3) a latens változók és a manifeszt változók mérési hibája között a kovariancia (korreláció) nulla:

$$E(\xi\delta') = \mathbf{0} \qquad E(\eta\epsilon') = \mathbf{0}.$$

Ezenkívül minden modellre külön feltételeket specifikálhatunk, amelyeket modellfeltételeknek nevezünk. Ezeket majd a későbbiekben tárgyaljuk.

A modelleket akkor tekintjük adottnak, ha a következő nyolc paramétermátrixot specifikáltuk:

B: a latens (endogén) függő változók közötti regressziós együtthatók mátrixa ($m \times m$) típusú, általános eleme β_{ij} a j-edik endogén változónak az i-edik endogén változóra kifejtett közvetlen hatását mutatja az előjel megfordításával.

Γ: a független (endogén) latens változónak a függő (endogén) latens változóra vonatkozó regressziós együtthatók mátrixa, ($m \times n$) típusú, általános eleme γ_{ij} a j-edik exogén változónak az i-edik endogén változóra kifejtett közvetlen hatását mutatja.

Φ: a független (exogén) latens változók közötti kovariancia(korreláció)mátrix, ($n \times n$) típusú, általános eleme Φ_{ij} az i-edik és j-edik exogén változók közötti kovarianciát (korrelációt) fejezi ki ($E(\xi_i \xi_j)$), a Φ_{ii} diagonális elem az i-edik exogén latens változó varianciája $E(\xi_i \xi_i)$.

Ψ: az endogén változók sztochasztikus reziduális tagjai között számított kovariancia(korreláció)mátrix, ($m \times m$) típusú, általános eleme Ψ_{ij} az i-edik és j-edik reziduális komponens között számolt kovariancia (korreláció) $E(\zeta_i \zeta_j)$, és a Ψ_{ii} jelöli az i-edik reziduum varianciáját $E(\zeta_i \zeta_i)$.

Λ_x: a megfigyelt független (exogén) változók faktorsúlymátrixa, ($q \times n$) típusú, az általános eleme λ_{xij} a j-edik latens exogén változó közvetlen hatását fejezi ki az i-edik megfigyelt változóra.

Θ_δ: a megfigyelt független (exogén) változók hibatagjainak kovarianciamátrixa, ($q \times q$) típusú, általános eleme $\Theta_{\delta ij}$ az i-edik és j-edik mérési exogén változók hibatagjainak kovarianciája $E(\delta_i \delta_j)$, a $\Theta_{\delta ii}$ jelöli az i-edik változó hibájának a varianciáját $E(\delta_i \delta_i)$.

Λ_y: a megfigyelt függő (endogén) változók faktorsúlymátrixa, ($p \times m$) típusú, általános eleme λ_{yij} a j-edik latens változó közvetlen hatását fejezi ki az i-edik megfigyelt endogén változóra.

Θ_ε: a megfigyelt függő (endogén) változók hibatagjainak kovarianciamátrixa, ($p \times p$) típusú, általános eleme $\Theta_{\varepsilon ij}$ az i-edik és j-edik mérési endogén változók hibatagjainak kovarianciája $E(\varepsilon_i \varepsilon_i)$, a $\Theta_{\varepsilon ii}$ jelöli az i-edik változó hibájának a varianciáját $E(\varepsilon_i \varepsilon_i)$.

16.1. A strukturális egyenlet redukált formája

A következőkben először a strukturális egyenlet redukált formáját vizsgáljuk meg, majd a megfigyelt változók kovarianciamátrixát a paraméterek függvényeként határozzuk meg.

A (16.1) strukturális egyenletet ($\mathbf{B}\eta = \Gamma \xi + \zeta$) balról megszorozzuk $\mathbf{B}^{-1}$ inverzmátrixszal; így a strukturális egyenlet redukált formájához jutunk:

$$\eta = \mathbf{D}\xi + \zeta_r. \tag{16.5}$$

A **D** mátrix a redukált forma együtthatóit tartalmazza:

$$\mathbf{D} = \mathbf{B}^{-1}\Gamma \quad \text{és} \quad \zeta_r = \mathbf{B}^{-1}\zeta. \tag{16.6}$$

A függő változók (η) variancia-kovarianciamátrixa:

$$\mathbf{C} = E(\eta\eta') = E[(\mathbf{B}^{-1}\Gamma\xi + \mathbf{B}^{-1}\zeta)(\mathbf{B}^{-1}\Gamma\xi + \mathbf{B}^{-1}\zeta)],$$

mivel $E(\xi\xi') = \Phi$, $E(\zeta\zeta') = \Phi$, $E(\xi\zeta') = \mathbf{0}$, $E(\zeta\xi') = \mathbf{0}$ és $E(\zeta_r \xi_r') = \mathbf{B}^{-1}\Phi\mathbf{B}'^{-1}$.

A fentiek alapján

$$\mathbf{C} = \mathbf{D}\Phi\mathbf{D}' + \mathbf{B}^{-1}\Phi\mathbf{B}'^{-1}. \tag{16.7}$$

16.2. A megfigyelt változók variancia-kovarianciamátrixa

A függő és független megfigyelt változókat helyezzük a $\mathbf{z}$ vektorba, és tegyük fel, hogy a megfigyelt változókat az átlaguktól való eltérésekkel mértük: $\mathbf{z}' = [\mathbf{y}', \mathbf{x}']$. Ekkor

$$\boldsymbol{\Sigma} = E(\mathbf{z}\mathbf{z}') = \begin{pmatrix} E(\mathbf{y}\mathbf{y}') & E(\mathbf{y}\mathbf{x}') \\ E(\mathbf{x}\mathbf{y}') & E(\mathbf{x}\mathbf{x}') \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{pmatrix}. \tag{16.8}$$

A $\boldsymbol{\Sigma}$ különböző elemeit a modell segítségével a következőképpen kapjuk meg:

$$\boldsymbol{\Sigma}_{yy} = E(\mathbf{y}\mathbf{y}') = E[(\boldsymbol{\Lambda}_y\boldsymbol{\eta} + \boldsymbol{\epsilon})(\boldsymbol{\Lambda}_y\boldsymbol{\eta} + \boldsymbol{\epsilon})'],$$

mivel $E(\eta\epsilon') = \mathbf{0}$, $E(\epsilon\eta') = \mathbf{0}$, $E(\epsilon\epsilon') = \Theta_\varepsilon$ és $E(\eta\eta') = \mathbf{C}$, így

$$\boldsymbol{\Sigma}_{yy} = \boldsymbol{\Lambda}_y\mathbf{C}\boldsymbol{\Lambda}_y' + \boldsymbol{\Theta}_\varepsilon.$$

Hasonló módon kapjuk a $\boldsymbol{\Sigma}_{yx}$ elemeit is:

$$\boldsymbol{\Sigma}_{yx} = E(\mathbf{y}\mathbf{x}') = E[(\boldsymbol{\Lambda}_y\boldsymbol{\eta} + \boldsymbol{\epsilon})(\boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta})'],$$

$$E(\eta\delta') = \mathbf{0}, \quad E(\epsilon\xi') = \mathbf{0}, \quad E(\epsilon\delta') = \mathbf{0},$$

mivel $E(\zeta\xi') = \mathbf{0}$, $E(\eta\xi') = \mathbf{D}\Phi$, így

$$\boldsymbol{\Sigma}_{yx} = \boldsymbol{\Lambda}_y\mathbf{D}\Phi\boldsymbol{\Lambda}_x'.$$

Mivel $\boldsymbol{\Sigma}_{xy} = \boldsymbol{\Sigma}_{yx}$, így ismerjük már a $\boldsymbol{\Sigma}_{xy}$-t is.

Az alsó diagonális blokk $\boldsymbol{\Sigma}_{xx}$:

$$\boldsymbol{\Sigma}_{xx} = E(\mathbf{x}\mathbf{x}') = E[(\boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta})(\boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta})'],$$

mivel

$$E(\xi\xi') = \Phi, \quad E(\xi\delta') = \mathbf{0}, \quad E(\delta\xi') = \mathbf{0} \text{ és}$$
$$E(\delta\delta') = \Theta_\delta, \quad \text{így}$$
$$\boldsymbol{\Sigma}_{xx} = \boldsymbol{\Lambda}_x\,\Phi\,\boldsymbol{\Lambda}_x' + \boldsymbol{\Theta}_\delta.$$

A fentiek alapján a megfigyelt változók variancia-kovarianciamátrixát a paraméterek mátrixaival kifejezve:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Lambda}_y\,\mathbf{C}\boldsymbol{\Lambda}_y' + \boldsymbol{\Theta}_\varepsilon & \boldsymbol{\Lambda}_y\mathbf{D}\Phi\boldsymbol{\Lambda}_x' \\ \boldsymbol{\Lambda}_x\Phi\mathbf{D}'\boldsymbol{\Lambda}_y' & \boldsymbol{\Lambda}_x\Phi\,\boldsymbol{\Lambda}_x' + \boldsymbol{\Theta}_\delta \end{pmatrix},$$

vagy a $\mathbf{C}$ és $\mathbf{D}$ mátrixokat a (16.6) és (16.7) azonosságokkal helyettesítve:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Lambda}_y(\mathbf{B}^{-1}\Gamma\Phi\Gamma'\mathbf{B}'^{-1} + \mathbf{B}^{-1}\Phi\mathbf{B}'^{-1})\boldsymbol{\Lambda}_y + \boldsymbol{\Theta}_\varepsilon & \boldsymbol{\Lambda}_y\mathbf{B}^{-1}\Gamma\Phi\boldsymbol{\Lambda}_x' \\ \boldsymbol{\Lambda}_x\Phi\Gamma'\mathbf{B}'^{-1}\boldsymbol{\Lambda}_y' & \boldsymbol{\Lambda}_x\Phi\boldsymbol{\Lambda}_x' + \boldsymbol{\Theta}_\delta \end{pmatrix}. \tag{16.9}$$

A (16.8) egyenlet a megfigyelt változók variancia-kovarianciamátrixa ($\boldsymbol{\Sigma}$) és a modell $\mathbf{B}$, Γ, Φ, Φ, $\boldsymbol{\Lambda}_y$, $\boldsymbol{\Lambda}_x$, $\boldsymbol{\Theta}_\varepsilon$, $\boldsymbol{\Theta}_\delta$ paraméterei közötti kapcsolatot írja le. A (16.8) egyenlet alapján az általános modell vizsgálatának fontossága abban látszik, hogy a $\boldsymbol{\Sigma}$ mátrix és a modell paraméterei közötti kapcsolatot nem kell minden speciális modell esetén megvizsgálni, hanem az általános modell eredményei alkalmazhatók a speciális modellekre is.

16.3. Standardizálás

Az eddigiekben feltételeztük, hogy a megfigyelt változókat a várható értéküktől való eltéréseikkel mértük. Ismeretes, hogy ha a várható értéküktől való eltérésekkel mért változókat elosztjuk a szórásukkal, standardizált változókhoz jutunk, amelyekre az a jellemző, hogy a várható értékük nulla, szórásuk pedig egy.

Nézzük először azt az esetet, amikor a latens változók standardizáltak. Az η és ξ várható értéke feltételezésünk szerint nulla.

Az η és ξ változókat úgy standardizáljuk, hogy az $\mathbf{A}_\eta^{-1}$ és az $\mathbf{A}_\xi^{-1}$ mátrixokkal beszorozva újraskálázzuk őket, ahol $\mathbf{A}_\eta = (\text{diag}.\mathbf{C})^{1/2}$ és $\mathbf{A}_\xi = (\text{diag}.\Phi)^{1/2}$.

A Λ_y, Λ_x, $\mathbf{B}$, Γ, Φ, Φ paramétermátrixokat a standardizált latens változók esetén a következőképpen transzformáljuk:

$$\begin{aligned}
\mathbf{\Lambda}_y^* &= \mathbf{\Lambda}_y \mathbf{A}_\eta \\
\mathbf{\Lambda}_x^* &= \mathbf{\Lambda}_x \mathbf{A}_\xi \\
\mathbf{B}^* &= \mathbf{A}_\eta^{-1} \mathbf{A}_\eta \\
\mathbf{\Gamma}^* &= \mathbf{A}_\eta^{-1} \mathbf{\Gamma} \mathbf{A}_\eta \\
\mathbf{\Phi}^* &= \mathbf{A}_\xi^{-1} \mathbf{\Phi} \mathbf{A}_\xi^{-1} \\
\mathbf{\Phi}^* &= \mathbf{A}_\eta^{-1} \mathbf{\Phi} \mathbf{A}_\eta^{-1}.
\end{aligned} \tag{16.10}$$

A latens változók a standardizálás után is előállítják a megfigyelt változók eredeti értékeit és így a kovarianciamátrixot is:

$$\begin{aligned}
\mathbf{y} &= \mathbf{\Lambda}_y \mathbf{A}_\eta \mathbf{A}_\eta^{-1} \boldsymbol{\eta} + \boldsymbol{\epsilon} = \mathbf{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\epsilon} \\
\mathbf{x} &= \mathbf{\Lambda}_x \mathbf{A}_\xi \mathbf{A}_\xi^{-1} \boldsymbol{\xi} + \boldsymbol{\delta} = \mathbf{\Lambda}_x \boldsymbol{\xi} + \boldsymbol{\delta}.
\end{aligned}$$

Ha $\mathbf{y}$ és $\mathbf{x}$ megfigyelt változókat standardizáljuk, megszorozva megfigyelt értékeiket az $\mathbf{A}_y^{-1}$ és $\mathbf{A}_x^{-1}$ mátrixokkal, ahol

$$\begin{aligned}
\mathbf{A}_y &= (\text{diag } \mathbf{\Sigma}_{yy})^{1/2} \\
\mathbf{A}_x &= (\text{diag } \mathbf{\Sigma}_{xx})^{1/2}.
\end{aligned}$$

Λ_y, Λ_x, valamint Θ_ε és Θ_δ együtthatók a következőképpen transzformálódnak:

$$\begin{aligned}
&\mathbf{A}_y^{-1} \mathbf{\Lambda}_y^* \qquad \mathbf{A}_y^{-1} \mathbf{\Theta}_\varepsilon \mathbf{A}_y^{-1} \\
&\mathbf{A}_x^{-1} \mathbf{\Lambda}_y^* \qquad \mathbf{A}_x^{-1} \mathbf{\Theta}_\delta \mathbf{A}_x^{-1}.
\end{aligned} \tag{16.11}$$

Megjegyezzük, hogy standardizált változók esetén a variancia-kovarianciamátrix egyenlő a korrelációmátrixszal, és hogy a standardizált együtthatókat útegyütthatóknak nevezik.

16.4. Identifikáció

Mielőtt a paraméterek becslését megadnánk, az identifikáció problémáját kell megvizsgálnunk. Feltételezzük, hogy a megfigyelt változók valószínűségeloszlása jól jellemezhető az első és második momentumaikkal és így a magasabbrendű momentumok elhanyagolhatók. Gyakorlatilag ezzel feltételezzük, hogy a megfigyelt változók együttes eloszlása normális. Mivel a változók várható értékeiről feltételeztük, hogy egyenlők nullával, a $\mathbf{z}' = [\mathbf{y}', \mathbf{x}']$ eloszlását a $\boldsymbol{\Lambda}_y, \boldsymbol{\Lambda}_x, \mathbf{B}, \boldsymbol{\Gamma}, \boldsymbol{\Phi}, \boldsymbol{\Phi}, \boldsymbol{\Theta}_\varepsilon, \boldsymbol{\Theta}_\delta$, paraméterektől függő variancia-kovarianciamátrixszal jellemezzük. Jelölje π vektor az összes paramétert, és legyen t a vektor elemeinek a száma. A variancia-kovarianciamátrix $\boldsymbol{\Sigma}$ általános elemére felírhatjuk a következő összefüggést:

$$\sigma_{ij} = f_{ij}(\boldsymbol{\pi}), \tag{16.12}$$

amely azt fejezi ki, hogy a megfigyelt változók varianciái és kovarianciái a $\pi_1, \pi_2, \ldots, \pi_t$ paraméterek függvényei. A $\sigma_{ij} = f_{ij}(\pi)$ függvényekről feltételezzük, hogy folytonosak, és folytonosak az első deriváltjai is. A $\boldsymbol{\Sigma}$ variancia-kovarianciamátrixról feltételezzük, hogy a lehetséges paramétertér minden π pontjában pozitív definit.

Egy specifikált modell, egy adott struktúra ($\boldsymbol{\Lambda}_y, \boldsymbol{\Lambda}_x, \mathbf{B}, \boldsymbol{\Gamma}, \boldsymbol{\Phi}, \boldsymbol{\Phi}, \boldsymbol{\Theta}_\varepsilon, \boldsymbol{\Theta}_\delta$, vagy egyszerűen π) egy és csak egy $\boldsymbol{\Sigma}$-t generál. Ugyanakkor különböző struktúrák is generálhatják ugyanazt a $\boldsymbol{\Sigma}$-t. Ha két vagy több struktúra ugyanazt a $\boldsymbol{\Sigma}$-t generálja, akkor ezeket empirikusan ekvivalens struktúráknak nevezzük. Ha egy paraméter az összes ekvivalens struktúrában azonos értéket vesz fel, akkor a kérdéses paraméter identifikálható. Ha a modell valamennyi paramétere identifikálható, akkor a modellt identifikálhatónak nevezzük. Más szóval a modellt identifikálhatónak nevezzük, ha a paramétertér bármely két $\pi_1 \neq \pi_2$ vektorára igaz, hogy $\boldsymbol{\Sigma}(\pi_1) \neq \boldsymbol{\Sigma}(\pi_2)$, vagyis a variancia-kovarianciamátrixot ($\boldsymbol{\Sigma}$) egy és csak egy π generálja. Ha a modell nem identifikálható, abból még nem következik, hogy a modellnek nincs identifikálható paramétere.

Az identifikálhatóság a modell megválasztásától függ. Az identifikálhatóságot vizsgálva induljunk ki a (16.12) egyenletből:

$$\sigma_{ij} = f_{ij}(\boldsymbol{\pi}) \quad i \leq j.$$

Összesen $(1/2)(p+q)(p+q+1)$ egyenletünk van t ismeretlennel (a π paramétervektor elemeinek száma). Eszerint a modell identifikálhatóságának szükséges feltétele, hogy a

$$t \leq (1/2)(p+q)(p+q+1)$$

reláció teljesüljön. Ha egy paramétert a $\boldsymbol{\Sigma}$ mátrixból meg tudunk határozni, a kérdéses paraméter identifikálható, különben nem. Azonban előfordulhat, hogy néhány paramétert többféleképpen is meghatározhatunk (a (16.12)-ből különböző egyenleteket felhasználva). Ez a túlidentifikáltság feltétele. Mivel a (16.12) egyenlet gyakran nemlineáris, az egyenletek megoldása bonyolult, és a π explicit megoldása ritkán létezik, az identifikáció vizsgálatára egy ún. információs mátrix vizsgálata alapján döntjük el, hogy a modell, illetve annak paraméterei identifikálhatók-e. Ha az információs mátrix pozitív definit, akkor a modell identifikálható. Ha az információs mátrix szinguláris, akkor a modell nem identifikálható, és az információs mátrix rangja alapján ítélhető meg, hogy mely paraméterek nem identifikálhatók.

16.5. A paraméterek becslése

Feltételezzük, hogy a megfigyelt változók eloszlása a várható értékekkel és a kovarianciamátrixszal jellemezhető.

A kovarianciamátrix $\boldsymbol{\Sigma}$ a π paraméter függvénye. A paraméterek a gyakorlatban ismeretlenek, és az n elemű független minta alapján becsüljük őket. Az $(\mathbf{y}', \mathbf{x}')$ megfigyelt változók mintabeli kovarianciamátrixát $\mathbf{S}$ mátrix jelölje. A gyakorlatban a változók mérésénél a mérési skála egysége gyakran önkényes, vagy indifferens, ezért a kovarianciamátrix helyett szoktuk a korrelációmátrixot használni ($\mathbf{R}$).

Mivel a változókat a várható értéküktől való eltérésükkel jellemezük, a becslés problémája tulajdonképpen az, hogy a π paraméterektől függő $\boldsymbol{\Sigma}$ mátrixot illesszük a mintabeli $\mathbf{S}$ kovarianciamátrixhoz. Az illesztésre három eljárást tekintsünk át.

Az első a legegyszerűbb. Minimalizáljuk a megfigyelt és a modell által reprodukált varianciák s kovarianciák különbségeinek négyzetösszegét:

$$U = \sum_{i}^{p+q} \sum_{j}^{j<i} (s_{ij} - \sigma_{ij})^2 = \sum_{i}^{p+q} \sum_{j}^{j<i} (s_{ij} - f_{ij}(\boldsymbol{\pi}))^2. \tag{16.13}$$

Vagy mátrixjelöléssel:

$$U = \frac{1}{2} tr(\mathbf{S} - \boldsymbol{\Sigma})^2. \tag{16.14}$$

Ez a közismert legkisebb négyzetek módszere (unweighted least squares, ULS), amellyel a reziduális variancia-kovarianciákat minimalizáljuk.

A második illesztő eljárás az általánosított legkisebb négyzetek módszere (generalized least squares, GLS). Ennek lényege, hogy az $(\mathbf{S} - \boldsymbol{\Sigma})$ reziduális kovarianciamátrixot egy $\mathbf{V}$ súlymátrixszal megszorozzuk, és a súlyozott reziduális négyzetösszeget minimalizáljuk:

$$G = \frac{1}{2} tr[\mathbf{V}(\mathbf{S} - \boldsymbol{\Sigma}])^2. \tag{16.15}$$

Ha a súlymátrix egységmátrix, akkor visszakapjuk a súlyozatlan legkisebb négyzetek módszerét. A gyakorlatban súlynak a $\mathbf{V} = \mathbf{S}^{-1}$ mátrixot szokták megadni, ekkor

$$G = \frac{1}{2} tr(\mathbf{I} - \mathbf{S}^{-1}\boldsymbol{\Sigma})^2. \tag{16.16}$$

Mindkét eljárás nagy előnye, hogy a változók eloszlására nem tesz feltételt.

A harmadik eljárás a maximum likelihood-módszer. Ennél az eljárásnál a $\boldsymbol{\Sigma}$ adott értékeire az $\mathbf{S}$ elemeinek sűrűségfüggvényét ismernünk kell. Ha ismerjük az $\mathbf{y}$ és $\mathbf{x}$ változók eloszlását, akkor $\mathbf{S}$ sűrűségfüggvényét is meghatározhatjuk. Jelöljük ezt a sűrűségfüggvényt $f(\mathbf{S} \mid \boldsymbol{\Sigma})$-val. Ez a függvény $\boldsymbol{\Sigma}$-n keresztül függ a π paramétertől. A mintából számított $\mathbf{S}$ kovarianciamátrix elemeit behelyettesítve a fenti függvénybe jutunk a likelihood-függvényhez (L). A π maximum likelihood $\widehat{\pi}$ becslésének a π-nek azt az értékét tekintjük, amely mellett a likelihood-függvény felveszi a maximumát. Más szóval, adott minta esetén π becsléséül azt fogadjuk el, amely mellett éppen ennek a mintának a bekövetkezése a legvalószínűbb.

A likelihood-függvény (L) helyett a gyakorlatban annak logaritmusát maximalizáljuk ($\log L$). A $\widehat{\pi}$ értékeit szintén nem változtatja, ha $\log L$ helyett a $(c_1 \log L + c_2)$-t maximalizáljuk, vagy az $F = -(c_1 \log L + c_2)$ függvényt minimalizáljuk, ahol c_1 és c_2 konstansok.

Feltételezve, hogy a megfigyelt változók eloszlása normális, a likelihood-függvény (Anderson, L. (1958) p. 159) logaritmusa egy konstans elhagyásával a következő:

$$\log L + -\frac{1}{2}(N-1)[\log|\boldsymbol{\Sigma}| + tr(\mathbf{S}\boldsymbol{\Sigma}^{-1})], \tag{16.17}$$

ahol $|\boldsymbol{\Sigma}|$: a $\boldsymbol{\Sigma}$ determinánsát jelöli, $tr(\mathbf{S}\boldsymbol{\Sigma}^{-1})$: az $(\mathbf{S}\boldsymbol{\Sigma}^{-1})$ mátrixszorzat diagonális elemeinek összegét jelöli, N: a megfigyelési egységek száma.

Jöreskog javasolta az F függvény minimalizálását π paraméter becslésére:

$$F = \log|\boldsymbol{\Sigma}| + tr(\mathbf{S}\boldsymbol{\Sigma}^{-1}) - \log|\mathbf{S}| - (p+q). \tag{16.18}$$

Az F függvény minimalizálása ugyanazt a $\widehat{\pi}$ értéket adja, mint L maximalizálása, mivel

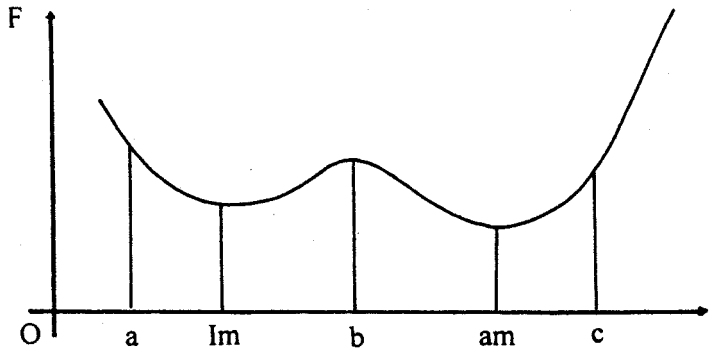
$$F = -(c_1 \log L + c_2),$$

ahol

$$c_1 = \frac{2}{N-1}$$

$$c_2 = \log|\mathbf{S}| + (p+q) \qquad \text{konstansok.}$$

Az F függvény minimalizálását a Fletcher–Powell (1963)-féle eljárással végezzük. Az eljárás F első- és másodrendű deriváltjait használja, és egy önkényes kezdőpontból elindulva gyors konvergenciát biztosít egy lokális minimumhoz. Ha az F függvénynek több minimumhelye létezik, akkor nincs garancia, hogy a globális optimumot megtaláljuk. A következő ábrából látszik, hogy ha a kezdőpontunk a és b közé esik, akkor az eljárás lokális optimumhoz konvergál, míg ha a kezdőpont a b és c értékek közé esik, akkor az eljárás az abszolút minimumot fogja megtalálni.



A paraméterek standard hibáját az ún. információs mátrixból számítjuk. Invertáljuk az információs mátrixot, és az inverz diagonális elemei adják a paraméterek standard hibáit. A standard hibák ismeretében a paraméterekre konfidenciaintervallumot szerkesztünk. Ha $\widehat{\pi}_i$ a π_i paraméter becslését jelöli, és $\widehat{\sigma}_{\widehat{\pi}_i}$ jelöli a $\widehat{\pi}_i$ standard hibáját (a $\widehat{\pi}_i$ mintabeli szórását), akkor a becslés valódi értéktől való eltérése osztva a standard hibával nagy minta esetén standard normális eloszlású:

$$z = \frac{\pi_i - \widehat{\pi}_i}{\widehat{\sigma}_{\widehat{\pi}_i}}.$$

Ennek alapján a π_i paraméter értéke $1 - \alpha$ valószínűséggel a következő intervallumba esik:

$$P\left(\widehat{\pi}_i - z_\alpha \widehat{\sigma}_{\widehat{\pi}_i} \leq \pi_i \leq \widehat{\pi}_i + z_\alpha \widehat{\sigma}_{\widehat{\pi}_i}\right) = 1 - \alpha. \tag{16.19}$$

Ha például $\widehat{\pi}_i = 3$ és $\widehat{\sigma}_{\widehat{\pi}_i} = 0{,}2$ és a paraméter 95%-os konfidenciaintervallumát szeretnénk tudni, akkor $z = 1{,}96$, és így a 95%-os konfidenciaintervallum $(3 - 1{,}96 \cdot 0{,}2,\ 3 + 1{,}96 \cdot 0{,}2)$, vagyis a kérdéses paraméter a (2,608, 3,392) intervallumba esik 95%-os valószínűséggel.

16.6. A modell tesztelése

A paraméterek maximum likelihood becslése után a modell illeszkedésének a jóságát a likelihood-hányados technikával tesztelhetjük. Legyen H_0 az adott modell nullhipotézise.

Tekintsük először azt az alternatív hipotézist H_1, hogy a Σ mátrix tetszőleges pozitív definit mátrix. Ekkor mínusz kétszer a likelihood-hányados logaritmusa egyenlő lesz $(N/2)F_0$-val, ahol F_0 az F függvény minimuma, és N a minta elemeinek száma. Nagy minta esetén ez χ^2 eloszláshoz vezet

$$d = \frac{1}{2}(p+q)(p+q+1) - t \quad \text{szabadságfokkal,}$$

ahol: t: a paraméterek száma a H_0 hipotézisben.

Az adott modellben (amelyre a H_0 hipotézist fogalmaztuk meg) a paraméterek száma legyen egyenlő u-val. Az adott modell paraméterei (γ) részhalmazát adja π-nek, és $u < t$.

Legyen F_0 az F függvény minimuma a H_0 hipotézis mellett, és a H_1 hipotézisnek megfelelő modell F függvényének minimumhelyét jelölje F_1. Ekkor $(N/2)(F_0 - F_1)$ közelítőleg χ^2 eloszlású $(t - u)$ szabadságfokkal.

Az illeszkedés jóságát kifejező χ^2 értékét a következőképpen értelmezzük. Az adott modellre kapott χ^2 értéket a megfelelő szabadságfok mellett összehasonlítjuk az elméleti (táblázatból kiolvasható) χ^2 értékekkel, és megkeressük a szignifikanciaszinteknek azt a tartományát, amely mellett a táblázatbeli (elméleti) χ^2 érték nagyobb az adott modell mintából számított χ^2 értékénél. Ezen tartományon belül a modellt elfogadjuk, ezen kívül pedig elvetjük. Általában a $p = 0{,}05$ szignifikanciaszinthez (ez 95%-os valószínűségi állítást jelent) meghatározzuk a χ^2 kritikus (elméleti) értékét, és ha az adott modell számított értéke ennél nagyobb, a hipotézist (az adott modellt) elvetjük, és megvizsgáljuk a mintából számított kovarianciák ($\mathbf{S}$) és a modell által becsült Σ kovarianciák eltéréseit, valamint az F függvény első deriváltjait. Ezek alapján módosítjuk a modellt, általában úgy, hogy több paramétert engedünk meg a modellben, és újra elvégezzük az illesztést. Így általában kisebb χ^2 értékhez jutunk. Ha ez a javulás (a két χ^2 érték különbsége) a megfelelő szabadságfok (a két szabadságfok különbsége) mellett szignifikáns, akkor a modell változtatása jelentős, helyes volt, egyébként a modell illeszkedését jelentősen nem javítottuk.

16.7. Az általános modell speciális esetei

A következőkben az általános modell speciális eseteit vesszük sorra.

16.7.1A klasszikus mérési modell

A klasszikus mérési modell szerint a megfigyelt (manifeszt) változók mérési értékeit a valódi érték (ξ) és a mérési hiba (δ) összegeként kapjuk meg, más szóval a mérés két hatás eredője, az egyik hatás a valódi (a hipotetikus hatás), a másik a hibahatás. Formalizálva ezt a következőképpen írhatjuk fel:

$$x = \xi + \delta, \tag{16.20}$$

ahol $E(\delta) = 0$, $E(x) = 0$, $E(\xi) = 0$, $E(\xi\delta) = 0$.

Mivel a regressziós együttható egyenlő eggyel, a valódi értéket a megfigyelt értékkel azonos skálán fejezzük ki.

Ha a valódi és a megfigyelt értéket standardizáljuk, a fenti mérési modellt a következőképpen írhatjuk fel:

$$x^* = \lambda^*\xi^* + \delta^*, \tag{16.21}$$

ahol

$$x^* = \frac{x}{\sigma_x}, \qquad \lambda^* = \frac{\sigma_\xi}{\sigma_x},$$
$$\delta^* = \frac{1}{\sigma_x}\delta, \qquad \xi^* = \frac{\xi}{\sigma_\xi},$$

és

$E(x^*) = 0$, $E(\xi^*) = 0$, $E(\delta^*) = 0$, $E(\xi^*\delta^*) = 0$,
$E(\xi^*\xi^*) = 1$, $E(x^*x^*) = 1$.

A valódi és a megfigyelt értékek közötti korreláció négyzete a mérés megbízhatóságát fejezi ki, és megbízhatósági együtthatónak nevezzük (ρ^2):

$$\rho_{x\xi} = E(x^*\xi^*) = E[(\lambda^*\xi^* + \delta^*)(\xi^*)] = \lambda^* = \frac{\sigma_\xi}{\sigma_x},$$

és

$$\rho^2_{x\xi} = \lambda^{*2} = \frac{\sigma^2_\xi}{\sigma^2_x} = \frac{\Phi}{\sigma^2_x}. \tag{16.22}$$

A megbízhatósági együttható egyenlő a standardizált regressziós együttható négyzetével, vagy egyenlő a valódi értékek varianciája és a megfigyelt értékek varianciája hányadosával. Ha legalább két manifeszt változó hipotetikus komponense megegyezik (a valódi értékük ugyanaz), és csak a hibakomponensben különböznek, a következő mérési modellt írhatjuk fel:

$$x_i = \xi + \delta_i \qquad \forall i\text{-re}, \tag{16.23}$$

ahol

$$E(x_i) = 0, \qquad E(\xi) = 0, \qquad E(\delta_i) = 0, \qquad \forall i\text{-re},$$
$$E(\xi\delta_i) = 0, \qquad \forall i\text{-re},$$
$$E(\delta_i\delta_j) = \Theta_{\delta_{ij}} = 0, \qquad \forall i\text{-re},$$
$$\Theta_{\delta_{ii}} = \Theta_{\delta_{jj}}, \qquad \forall i, j\text{-re}.$$

Azokat a mérőeszközöket, amelyek eleget tesznek a fenti modell követelményeinek, „parallel instruments”, „párhuzamos mérőeszközöknek” nevezzük.

A modell standardizált változókkal:

$$x_i^* = \lambda_i^* \xi^* + \delta_i^*, \tag{16.24}$$

ahol

$$E(x_i^*) = 0, \qquad E(\xi^*) = 0, \qquad E(\delta_i^*) = 0, \quad \forall i\text{-re},$$

$$\sigma^2_{x_i^*} = 1, \qquad \sigma^2_{\xi x} = \Phi = 1, \qquad \forall i\text{-re},$$

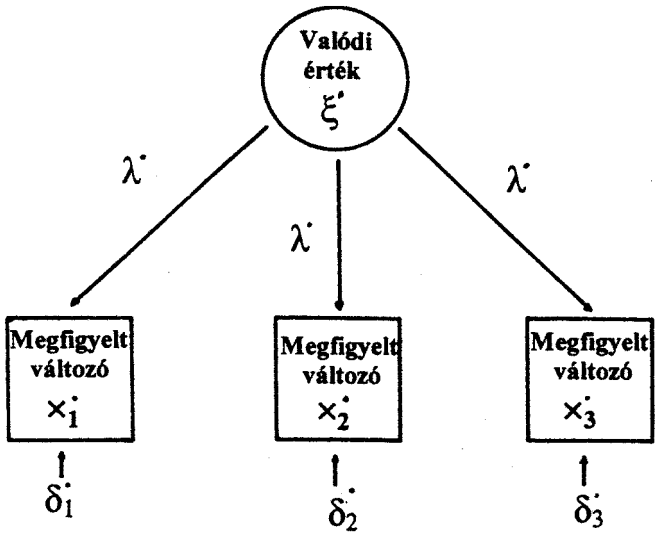
$$E(\xi^* \delta_i^*) = 0, \qquad \forall i\text{-re},$$

$$E(\delta_i^* \delta_j^*) = \Theta_{\delta_{ij}} = 0, \qquad \forall i, j\text{-re},$$

$$\Theta_{\delta_{ii}} = \Theta_{\delta_{jj}}, \qquad \forall i, j\text{-re},$$

$$\lambda_i^* = \lambda_j^*, \qquad \forall i, j\text{-re}.$$

Az utolsó feltétel azt jelenti, hogy a megbízhatósági együttható (λ^{*2}) minden párhuzamos mérőeszköz esetén azonos.



16.1. ábra. A „párhuzamos mérőeszközök” modell útdiagramja

A párhuzamos mérésnél feltételezzük, hogy a megfigyelt (manifeszt) változók valódi értéke és a hibavarianciája minden változónál azonos. Gyakran előfordul, hogy a megfigyelt változók ugyanazt az elméleti változót, kategóriát mérik, de éppen az eltérő mérőeszköz (annak eltérő minősége) miatt különböző lesz a hibavariancia. Ha a mérésnél az utóbbi feltételt (a hibavariancia állandó) nem tekintjük érvényesnek, akkor a mérőeszközöket „tau ekvivalens”-nek nevezzük.

Más esetekben a különböző mérőeszközöknél a valódi értékek nem azonosak, de közöttük lineáris összefüggés van, valamint a mérőeszközök varianciája is különbözik. Ebben az esetben a mérőeszközöket rokonnak, „congeneric”-nek nevezzük, és a következőképpen írhatjuk fel:

$$x_1 = \xi_1 + \delta_1$$

$$x_2 = \xi_2 + \delta_2 \qquad (16.25)$$
$$\xi_2 = \lambda_{21}\xi_1,$$

vagy

$$x_1 = \xi_1 + \delta_1$$
$$x_2 = \lambda_{21}\xi_1 + \delta_2,$$

ahol

$E(x_i) = 0, \qquad E(\delta_i) = 0, \qquad E(\xi_i) = 0, \quad \forall i$-re,

$E(\xi\delta_i) = 0, \qquad \forall i$-re,

$E(\delta_i\delta_j) = 0.$

A „rokon" modell általánosabb, mint az előző két modell, mivel:

a) a „tau-ekvivalens" modelleknél:

$$\lambda_{ij} = 1 \qquad \forall i, j\text{-re},$$

b) a „párhuzamos" mérési modellnél:

$$\lambda_{ij} = 1 \qquad \text{és} \qquad \Theta_{\delta_{ii}} = \Theta_{\delta_{jj}} \qquad \forall i, j\text{-re}.$$

A megbízhatósági együttható nem lesz egyenlő a „tau-ekvivalens" és a „rokon" mérésnél. A korrelációmátrixból azonban becsülhetjük a λ^*-t, és így ebből a megbízhatósági együtthatót is megkaphatjuk $\lambda_i^{*2} = \rho^2_{x_i\xi}$.

16.7.2A többjellemzős-többmódszeres modell (Multitrait-multimethod model)

Az előző részben a klasszikus méréselmélet szerinti mérési modelleket tárgyaltuk. A következőkben a mérési modell megbízhatósága helyett a mérés érvényességének (validity) problémáját a Campbell és Fiske (1959) által bevezetett többjellemzős-többmódszeres modell alapján elemezzük.

A többjellemzős-többmódszeres modell egyenletei és feltételei:

$$x_i = \tau_i + \varepsilon_i \qquad (16.26)$$
$$\tau_i = \xi_j + \xi_k,$$

ahol τ_i: az i-edik megfigyelt változó valódi értéke, ε_i: az i-edik hibatag, ξ_j: a j-edik elméleti változó, ξ_k: a k-adik módszer hatását kifejező változó, valamint:

$E(\varepsilon_i) = 0,$

$E(\xi_j) = 0, \qquad \forall j$-re,

$E(\xi_j\xi_k) = 0, \quad \forall j, k$-ra.

A fentiekből következik, hogy a valódi érték varianciája az elméleti változó és a módszer varianciájának összege:

$$\sigma_{\tau_i\tau_i} = \Phi_{jj} + \Phi_{kk},$$

ahol $\Phi_{jj} = E(\xi_j^2)$ az elméleti változó varianciája, Φ_{kk} a módszer hatását képviseli.

Ha valamennyi változó standardizált, a fenti modellt a következőképpen írhatjuk fel:

$$x_i^* = \lambda_{ij}^*\xi_j^* + \lambda_{ik}^*\xi_k^* + \varepsilon_i^*, \qquad (16.27)$$

ahol

$E(x_i^*) = E(\xi_i^*) = E(\varepsilon_i^*) = 0, \qquad \forall i$-re,

$E(x_i^{*2}) = E(\xi_i^{*2}) = 1, \qquad \forall i$-re,

$E(\varepsilon_i^* \xi_j^*) = 0$, és $E(\varepsilon_i^* \varepsilon_j^*) = 0, \quad \forall i, j$-ra.

A fenti modellből következik:

$$\sigma_{x_i^* x_i^*} = \lambda_{ij}^* \lambda_{ij}^* + \lambda_{ik}^* \lambda_{ik}^* + \Theta_{\varepsilon_{ii}}^*,$$

vagyis a megfigyelt változók varianciája egyenlő az elméleti változó (ξ_i^*) varianciája, a módszer (ξ_k^*) varianciája és a hibavarianciája ($\Theta_{\varepsilon_{ii}}^*$) összegével.

A megfigyelt változók kovarianciái:

$$\sigma_{x_i^* x_j^*} = \lambda_{ij}^* \Phi_{jj}^* \lambda_{\ell j}^* + \lambda_{ik}^* \Phi_{kk}^* \lambda_{\ell k}^*,$$

ami azt jelenti, hogy a megfigyelt változók kapcsolódásaira nemcsak az elméleti változó, hanem a módszer is hat.

16.7.3A variancia-kovariancia komponens modell

A variancia-kovariancia modell a megfigyelt változók mintabeli megfigyelt értékeit a valódi értékek és a hibaösszegeként állítja elő:

$$\mathbf{y} = \mathbf{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\epsilon}, \tag{16.28}$$

ahol $E(\mathbf{y}) = \mathbf{0}$, $\mathbf{E}(\eta) = \mathbf{0}$, $\mathbf{E}(\epsilon) = \mathbf{0}$, $E(\eta\epsilon') = \mathbf{0}$, $\mathbf{\Lambda}_y$ diagonális mátrix.

A valódi értékek azonban bizonyos latens változók függvényei:

$$\boldsymbol{\eta} = \mathbf{\Gamma}\boldsymbol{\xi}, \tag{16.29}$$

ahol $E(\xi\epsilon') = \mathbf{0}$, $E(\xi) = \mathbf{0}$.

A modellben a $\mathbf{\Lambda}_y$ és Γ paramétermátrixok speciális felépítésűek. A $\mathbf{\Lambda}_y$ mátrix diagonális, a diagonális elemei akkor különböznek 1-től, amikor a megfigyelt változók mértékegységei különbözőek. A Γ mátrix (amit design mátrixnak nevezünk) elemei 0 és 1 értéket vehetnek fel.

16.7.4. A faktorelemzés

A faktorelemzésben a megfigyelt változók (x) valódi értékeit (τ) a latens változók függvényeként fejezzük ki. A latens változók közül vannak, amelyek több megfigyelt változó előállításában is szerepelnek, ezeket közös faktoroknak (ξ) nevezzük, és vannak olyanok, amelyek csak egy-egy változó reprodukálásában játszanak szerepet, ezeket egyedi faktoroknak (u) nevezzük.

A faktorelemzés matematikai modellje:

$$x_i = \tau_i + e_i, \tag{16.30}$$

és

$$\tau_i = \lambda_{i1}\xi_1 + \lambda_{i2}\xi_2 + \ldots + \lambda_{im}\xi_m + u_i,$$

ahol e_i: az i-edik megfigyelt változó mérési hibája, τ_i: az i-edik megfigyelt változó valódi értéke, ξ_j: a j-edik közös faktor (latens változó), u_i: az i-edik egyedi faktor, λ_{ij}: a j-edik közös faktor regressziós együtthatója az i-edik megfigyelt változóra, $\delta_i = u_i + e_i$: az i-edik mérési hiba és az i-edik egyedi faktor összege.

Feltételezzük, hogy: $E(x_i) = 0$, $E(\xi_j) = 0$, $E(e_i) = 0$, $E(u_i) = 0$, $\forall i, j$-re, valamint $E(\xi_j u_i) = 0$, $E(\xi_j e_j) = 0$, $\forall i, j$-re.

A faktorelemzés modellje mátrixaritmetikai jelölésekkel:

$$\mathbf{x} = \mathbf{\Lambda}_x \boldsymbol{\xi} + \boldsymbol{\delta}, \tag{16.31}$$

ahol $\delta = \mathbf{u} + \mathbf{e}$, $E(\mathbf{x}) = \mathbf{0}$, $E(\mathbf{\Sigma}) = \mathbf{0}$, $E(\delta) = \mathbf{0}$, $E(\xi\delta') = \mathbf{0}$, $E(\delta\delta') = \Theta_\delta$ diagonális.

A modellben feltételezzük, hogy (1) a mérési hiba változói korrelálatlanok egymással $E(e_i e_j)=0$, (2) az egyedi faktorok korrelálatlanok egymással $E(u_i u_j) = 0$, (3) a közös faktorok korrelálatlanok δ-val. $E(\xi\delta') = \mathbf{0}$.

A modell paramétermátrixai: Λ_x, Φ és Θ_δ.

A variancia-kovarianciamátrix a paraméterek függvényében:

$$\mathbf{\Sigma} = \mathbf{\Lambda}_x \mathbf{\Phi} \mathbf{\Lambda}_x' + \mathbf{\Theta}_\delta. \tag{16.32}$$

A gyakorlatban az exploratív elemzésekben legtöbbször feltételezzük, hogy $\Phi = \mathbf{I}$, vagyis hogy a közös faktorok korrelálatlanok. Az exploratív faktorelemzés során általában a következő kérdéseket kell megválaszolnunk:

1) mennyi közös faktor szükséges a megfigyelt változók közötti korreláció magyarázatához,

2) a faktorsúlyok között melyek szignifikánsak, és melyek nem,

3) mi a faktorok elméleti tartalma.

Az első kérdés könnyen megválaszolható, mivel a különböző faktorszámú modellek szignifikanciáját χ^2 próbával tesztelhetjük, és így megtalálhatjuk a faktoroknak azt a számát, amelyet tovább növelve a modell illeszkedése szignifikánsan tovább nem javítható.

A második kérdés megválaszolására Archer és Jennrich (1973) a rotált faktorsúlyok standard hibájának számítására adott eljárást. A módszer azonban elég bonyolult, így a gyakorlati alkalmazása korlátozott.

Ennél egyszerűbb eljárást javasol Jöreskog (1978), amelyet „legjobban illeszkedő egyszerű struktúrá”-nak nevezett el.

A harmadik kérdés megválaszolásához a szignifikáns faktorsúlyok mellett a szakmai ismeretek kapnak nagy szerepet. A konfirmatív faktorelemzés alapvetően abban különbözik az exploratív faktorelemzéstől, hogy az előbbieknél a faktorstruktúrát *a priori* elméleti vagy korábbi empirikus vizsgálatok alapján ismertnek tételezzük fel, míg az utóbbinál a faktorstruktúrát az elemzés során, a vizsgált adatokból származtatjuk.

16.7.5. A másodrendű faktorelemzés

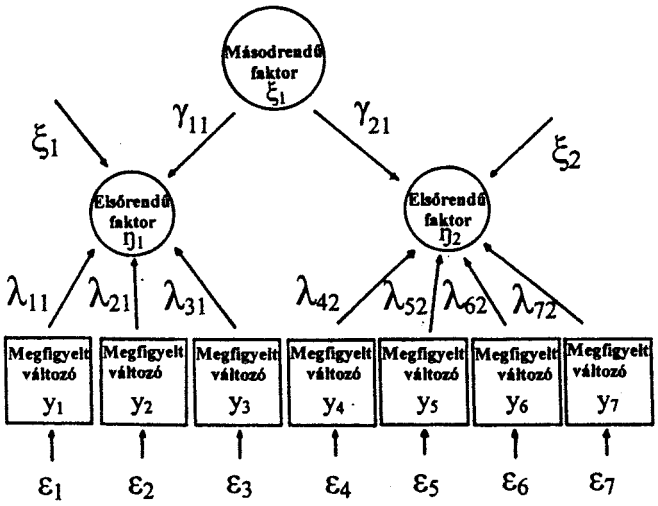
A másodrendű faktorelemzés hipotézise szerint a megfigyelt változók közös faktorai kifejezhetők további latens változók (faktorok) függvényeiként, amelyeket másodrendű faktoroknak nevezünk. A másodrendű faktorelemzés egyenletei:

$$\mathbf{y} = \mathbf{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\epsilon} \tag{16.33}$$

$$\boldsymbol{\eta} = \mathbf{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta},$$

ahol $E(\eta) = \mathbf{0}$, $E(\xi) = \mathbf{0}$, $E(\mathbf{y}) = \mathbf{0}$, $E(\zeta) = \mathbf{0}$, $E(\epsilon) = \mathbf{0}$, $E(\eta\epsilon') = \mathbf{0}$, $E(\xi\zeta') = \mathbf{0}$, $E(\xi\epsilon') = \mathbf{0}$, $E(\zeta\epsilon') = \mathbf{0}$.

Egy ilyen modellt mutat a következő ábra:



16.2. ábra. Másodrendű faktorelemzés diagramja

16.7.6A regresszióelemzés

Az eddig tárgyalt speciális modellekben nem tettünk különbséget a megfigyelt változók között. A következő részben a megfigyelt változók között megkülönböztetjük a függő, okozati változókat (y) és a független, az ok szerepét betöltő változókat (x). A modell szerint előállítani, magyarázni vagy becsülni a függő változót akarjuk a független változók segítségével, amelyekről feltételezzük, hogy hatással vannak a függő változóra.

A független változók lehetnek rögzítettek, vagy véletlen hatást is tartalmazók.

A regressziós modell egyenlete rögzített $\mathbf{x}$ esetén:

$$y = \gamma'\mathbf{x} + z, \tag{16.34}$$

ahol $E(y) = 0$, $E(\mathbf{x}) = \mathbf{0}$, $E(z) = 0$, $E(\mathbf{x}z) = \mathbf{0}$.

A modellben feltételezzük, hogy a reziduális (vagy véletlen) tag korrelálatlan az $\mathbf{x}$ vektorváltozóval.

A modell kovarianciamátrixa:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \gamma'\boldsymbol{\Sigma}_{xx}\gamma + \sigma_z^2 & \\ \boldsymbol{\Sigma}_{xx}\gamma & \boldsymbol{\Sigma}_{xx} \end{pmatrix}. \tag{16.35}$$

Ha $\mathbf{x}$ értékei rögzítettek, akkor $\boldsymbol{\Sigma}_{xx}$ egyenlő $\mathbf{S}_{xx}$-vel, a kovarianciamátrix közvetlenül $\mathbf{x}$ rögzített értékeiből számítható.

Ha $\mathbf{x}$ valószínűségi változó, véletlen hatásokat is tartalmaz, $\boldsymbol{\Sigma}_{xx}$ becslése az $\mathbf{S}_{xx}$ (mintából számított) kovarianciamátrix lesz.

Mindkét esetben az együtthatókat (γ) az $\mathbf{S}$ mátrix első sora alapján (az y és az $\mathbf{x}$ változók közötti kovarianciákból) becsüljük.

A reziduális tag varianciájának (σ_z^2) becslése $s_{yy} - \gamma'\mathbf{S}_{xx}\gamma$ lesz.

Ha egy vagy több független változó ($\mathbf{x}$) és a függő változó megfigyelt értékei mérési hibát is tartalmaznak, akkor a (16.34) egyenlet helyett a valódi regressziót keressük:

$$\eta = \gamma'\xi + \zeta, \tag{16.36}$$

ahol ζ: a függő latens változó reziduális komponense, amiről feltételezzük, hogy korrelálatlan ξ-vel: $E(\xi\zeta) = \mathbf{0}$.

Az η és ξ a függő és független változók valódi értékei, vagy más szóval közös faktorai:

$$y = \lambda\eta + \varepsilon, \tag{16.37}$$

és

$$\mathbf{x} = \mathbf{\Lambda}\boldsymbol{\xi} + \boldsymbol{\delta}, \tag{16.38}$$

ahol feltételezzük, hogy $E(y) = 0$, $E(\eta) = 0$, $E(\varepsilon) = 0$, $E(\eta\varepsilon) = 0$, és $E(\mathbf{x}) = \mathbf{0}$, $E(\xi) = \mathbf{0}$, $E(\delta) = \mathbf{0}$, $E(\xi\delta)' = \mathbf{0}$.

A modell kovarianciamátrixa:

$$\mathbf{\Sigma} = \begin{pmatrix} \lambda(\gamma'\Phi\gamma + \Psi^2)\lambda + \Theta_\varepsilon & \\ \mathbf{\Lambda}\Phi\gamma\lambda & \mathbf{\Lambda}\Phi\mathbf{\Lambda}' + \Theta_\delta \end{pmatrix},$$

ahol $\Phi = \text{cov}(\xi)$, $\Psi^2 = \text{cov}(\zeta)$, $\Theta_\varepsilon = \text{cov}(\varepsilon)$, $\Theta_\delta = \text{cov}(\delta)$.

16.7.7A z útelemzés

Az útelemzés modelljében az ok szerepét játszó változók közvetlen hatásait vizsgáljuk az okozat szerepét betöltő változókra. Ezeket a közvetlen hatásokat lineáris strukturális egyenletekkel fejezzük ki, és a feladata hipotetikusan felállított rekurzív modell paramétereinek a becslése.

Tegyük fel például, hogy két változót két időpontban figyeltünk meg, és a két változó ugyanazon latens változó függvénye. Az y_1 és y_2 az η_1 latens változót mérik az első időpontban, az y_3 és y_4 változók pedig az η_2 latens változót mérik a második időpontban. A mérési egyenletek ekkor:

$$\begin{aligned} y_1 &= \eta_1 + \varepsilon_1 \\ y_2 &= \lambda_1\eta_1 + \varepsilon_2 \\ y_3 &= \eta_2 + \varepsilon_3 \\ y_4 &= \lambda_2\eta_2 + \varepsilon_4. \end{aligned}$$

A fő kérdés az η latens változó időbeli stabilitása, amit a következő regressziós egyenlet fejez ki:

$$\eta_2 = \beta\eta_1 + \zeta.$$

Ha Φ az (η_1, η_2) latens változók kovarianciamátrixa, a Θ a $(\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4)$ mérési hibák kovarianciamátrixa (amelyről feltételezzük, hogy diagonális), akkor a megfigyelt változók kovarianciamátrixa:

$$\mathbf{\Sigma} = \begin{pmatrix} \Phi_{11} + \Theta_{11} & & & \\ \lambda_1\Phi_{11} & \lambda_1^2\Phi_{11} + \Theta_{22} & & \\ \Phi_{22} & \lambda_1\Phi_{21} & \Phi_{22} + \Theta_{33} & \\ \lambda_2\Phi_{21} & \lambda_1\lambda_2\Phi_{21} & \lambda_2\Phi_{22} & \lambda_2^2\Phi_{22} + \Theta_{44} \end{pmatrix}.$$

16.7.8A MIMIC-modell (Multiple Indicators Multiple Causes Model)

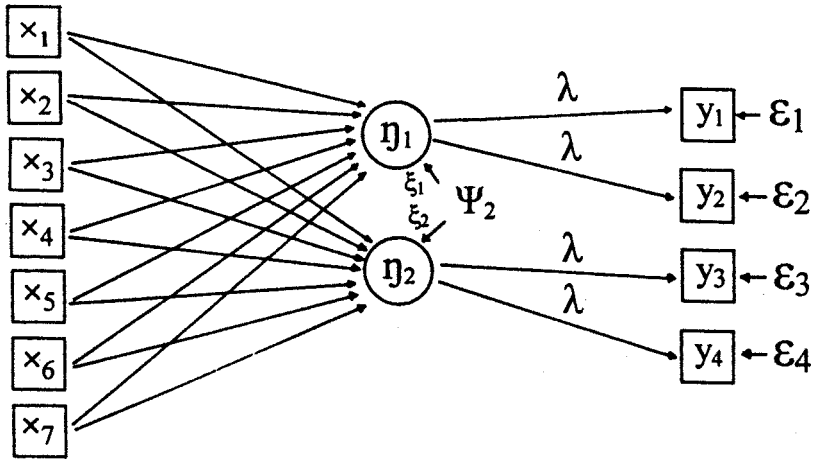
A MIMIC-modell két egyenletből áll. A strukturális egyenletből, amelyben az ok szerepét megfigyelt változók ($\mathbf{x}$), az okozat szerepét pedig latens változók (η) töltik be:

$$\beta\eta = \Gamma\mathbf{x} + \zeta, \tag{16.39}$$

ahol $E(\eta) = \mathbf{0}$, $E(\mathbf{x}) = \mathbf{0}$, $E(\zeta) = \mathbf{0}$, $E(\mathbf{x}\zeta') = \mathbf{0}$, és a mérési hibát tartalmazó egyenletből:

$$\mathbf{y} = \Lambda\eta + \epsilon, \tag{16.40}$$

ahol $E(\mathbf{y}) = \mathbf{0}$, $E(\epsilon) = \mathbf{0}$, $E(\eta\epsilon') = \mathbf{0}$, $E(\mathbf{x}\epsilon') = \mathbf{0}$, $E(\epsilon\zeta') = \mathbf{0}$.



16.3. ábra. A MIMIC-modell diagramja:

16.8. Szimultán elemzés több csoportban

Tekintsük egy vizsgált sokaság g számú csoportját. Ezek a csoportok lehetnek nemzetek egy nemzetközi vizsgálatban, lehetnek kulturális vagy szocioökonomikusan elkülönülő csoportok egy nemzeti vizsgálatban, vagy bármely, megfigyelési egységnek bizonyos változók alapján elkülönített csoportjai.

Feltételezzük, hogy minden csoportban a vizsgálat változóit véletlen mintán mértük. Feltételezzük továbbá, hogy minden csoportban a latens változók strukturális modelljét és a megfigyelt változók mérési modelljeit a (16.1), (16.2) és (16.3) egyenletek formájában adjuk meg, vagyis:

$$\mathbf{B}^{(k)}\eta = \Gamma^{(k)}\xi + \zeta^{(k)}, \tag{16.41}$$

$$\mathbf{y} = \Lambda_y^{(k)}\eta + \epsilon^{(k)}, \tag{16.42}$$

$$\mathbf{x} = \Lambda_x^{(k)}\xi + \delta^{(k)}, \tag{16.43}$$

formában, ahol k a csoportot jelölő index

$$k = 1, 2, \ldots, g.$$

A k-adik csoportban a modellt a nyolc paramétermátrix megadásával definiáljuk ($\mathbf{\Lambda}_y^{(k)}$, $\mathbf{\Lambda}_x^{(k)}$, $\mathbf{B}^{(k)}$, $\mathbf{\Gamma}^{(k)}$, $\mathbf{\Phi}^{(k)}$, $\mathbf{\Phi}^{(k)}$, $\mathbf{\Theta}_\varepsilon^{(k)}$, $\mathbf{\Theta}_\delta^{(k)}$). A paramétermátrixok elemei lehetnek rögzítettek (fixed), szabadok (free) vagy kötöttek (constrained). Ha a csoportok között a paraméterek egyezőségére nem írunk elő feltételt, akkor a csoportokat egymástól függetlenül elemezhetjük. Azonban ha a csoportok paraméterei között feltételezünk egyezőséget, akkor a csoportokat csak szimultán elemezhetjük, csak így kaphatunk efficiens becsléseket.

Ha például a megfigyelt változók mérési tulajdonságai azonosak a különböző csoportokban, akkor a következő paraméterek egyezőségét tételezhetjük fel:

$$\mathbf{\Lambda}_y^{(1)} = \mathbf{\Lambda}_y^{(2)} = \ldots = \mathbf{\Lambda}_y^{(g)}$$
$$\mathbf{\Lambda}_x^{(1)} = \mathbf{\Lambda}_x^{(2)} = \ldots = \mathbf{\Lambda}_x^{(g)}$$
$$\mathbf{\Theta}_\varepsilon^{(1)} = \mathbf{\Theta}_\varepsilon^{(2)} = \ldots = \mathbf{\Theta}_\varepsilon^{(g)}$$
$$\mathbf{\Theta}_\delta^{(1)} = \mathbf{\Theta}_\delta^{(2)} = \ldots = \mathbf{\Theta}_\delta^{(g)}.$$

Ekkor a csoportok a latens változók eloszlásában különböznek egymástól.

A strukturális kapcsolódások invarianciáját tesztelhetjük a következő azonosságok feltételezésével:

$$\mathbf{B}^{(1)} = \mathbf{B}^{(2)} = \ldots = \mathbf{B}^{(g)}$$
$$\mathbf{\Gamma}^{(1)} = \mathbf{\Gamma}^{(2)} = \ldots = \mathbf{\Gamma}^{(g)}.$$

A csoportok függetlenségétől a paraméterek teljes azonosságáig terjedően az invariancia bármely fokát tesztelhetjük. A szimultán modellt a következő függvény minimalizálásával becsüljük:

$$F = \sum_{k=1}^{g} (N_k/N)[\log|\mathbf{\Sigma}^{(k)}| + \mathrm{tr}(\mathbf{S}^{(k)}\mathbf{\Sigma}^{(k)-1}) - \log|\mathbf{S}^{(k)}| - (p+q)], \tag{16.44}$$

ahol N_k: a megfigyelések száma a k-adik csoportban, $N = N_1 + N_2 + \ldots + N_g$.

Ha a megfigyelt változók együttes eloszlása normális, az F függvény a likelihood-függvény logaritmusának mínusz $(N/2)$-szeresével lesz egyenlő. A modell illeszkedését a χ^2 próbával mérjük, hasonlóan ahhoz, amikor csak egy csoportot vizsgáltunk.

A szabadságfok:

$$d = (1/2)g(p+q)(p+q+1) - t,$$

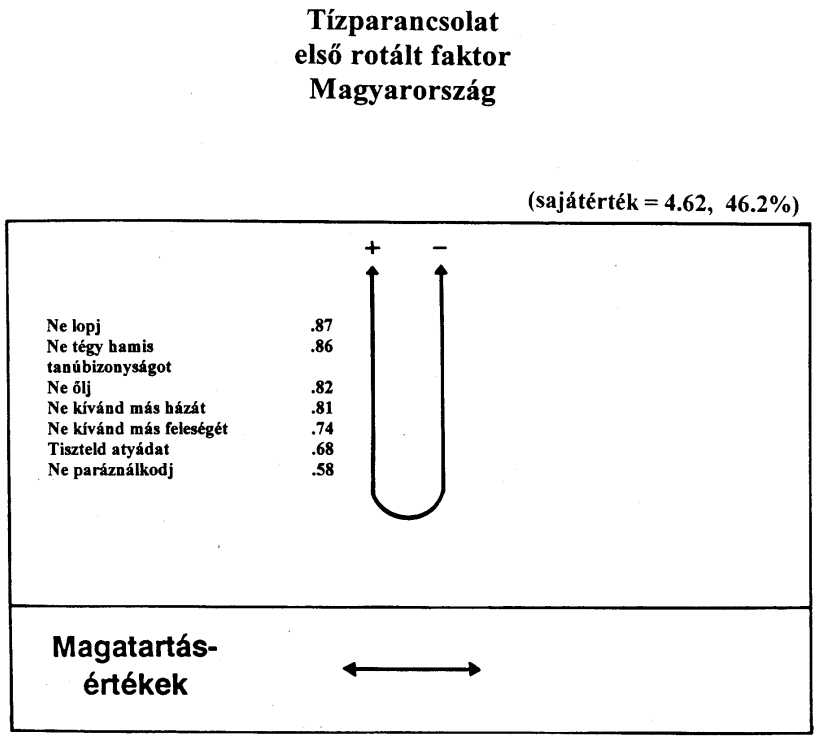
ahol t a független paraméterek száma az összes csoportban.

16.9. Példa a LISREL-modellre (*Az értékrendszer szűrő szerepének modellezése*)

A következő példában azt vizsgáltuk, hogy az emberek társadalmi háttere, a társadalmi státus mennyire befolyásolja az emberek társadalmi tudatát, azt, ahogyan vélekednek a társadalom különböző dolgairól (a példában két társadalmi problémát ragadtunk ki, az egyik a társadalmi intézményekbe vetett bizalom, és az életvitel-változásról alkotott vélemény), és hogyan módosítja, erősíti, csökkenti ezt a hatást az emberek értékrendje, világnézete. Azt a hipotézist vizsgáltuk, hogy az emberek értékrendszere szűrő szerepet

játszik-e a társadalomban elfoglalt hely és a szociális tudat között. A modellezés lépéseit most nem követjük végig, csupán a végső modellt prezentáljuk; méghozzá az Életvitel-változás és Bizalom endogén változó-blokkot tartalmazó modelleket. (Az adatok forrása: MTA Szociológiai Kutatóintézet Értékszociológiai Műhelye.)

Hogy a két modell ábráit olvasni tudjuk, a Gyermeknevelés, Tízparancsolat, Vallás blokkok indikátor-változóinak az értelmezéséhez – amelyek a manifeszt változók közös faktorai – közöljük a megfelelő faktorstruktúrákat.



16.4. ábra. Tízparancsolat (első, rotált faktor)

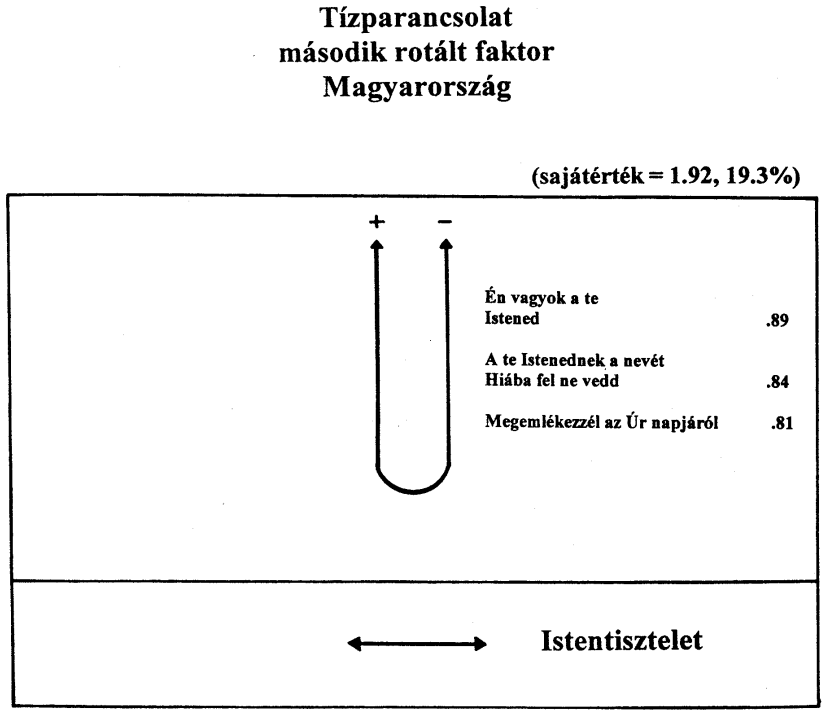
Gyermeknevelés

A 17 gyermeknevelési elv három közös faktorát az exploratív faktorelelemzés alkalmazásával állítottuk elő (PC-elemzés).

Az első rotálatlan faktor, a legfőbb értékpolaritást fejezi ki:

Modern személyiségértékek ⟷ Hagyományos közösségi értékek

Ezt az alapvető értékdimenziót figyeltük meg a Roceach-értékteszt elemzése során 1980-ban és 1982-ben is (intellektuális, autonómia értékei álltak szemben a hagyományos közösségi és öröm-értékekkel). Lásd Hankiss, Manchin, Füstös, Szakolczai: *Kényszerpályán.* MTA Szociológiai Intézet, 1983. Ezen a dimenzión a „belülről irányított ember" típusa áll szemben a „tradícióktól irányított" ember tulajdonságaival.



16.5. ábra. Tízparancsolt (második, rotált faktor)

A második rotálatlan faktor:

Munka, biztonságh it ⟷ Emberi kapcsolatok, autonómia

A pozitív oldalon kétfajta érték keveredik; a munkával kapcsolatos és a vallásos értékek, ezen belül az elsőknek van nagyobb, meghatározóbb súlya. Ebben az összefüggésben a vallásos hit a kötelesség, felelősség, becsület fogalmaihoz kapcsolódik. A munka értéke bizonyos fajta biztonságra való törekvéssel fonódik össze: anyagi szempontból a takarékossággal, lelki szempontból a vallásos hittel és a hűséggel. A másik, negatív oldalon az emberi kapcsolatokat szabályozó tulajdonságok szerepelnek. Olyan értékek, amelyek a személyes autonómia fenntartása mellett a másik ember számára is könnyebbé, zökkenőmentesebbé teszik az együttélést.

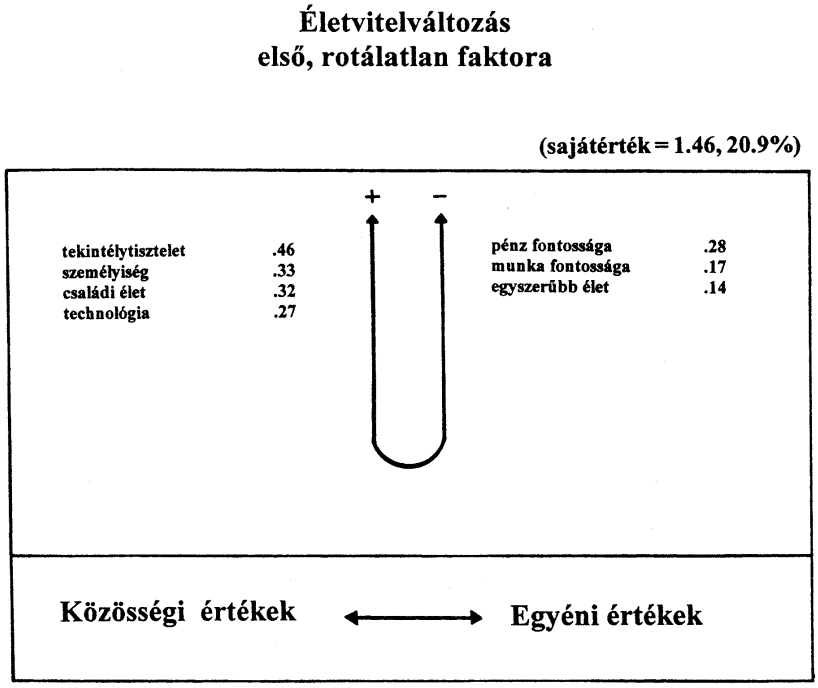
A harmadik rotálatlan faktor dichotómiája:

Hagyományos keresztény értékek ⟷ Protestáns értékek

A pozitív póluson a hagyományos keresztény értékek szerepelnek, középpontban mások tisztelete, az alkalmazkodás értékei, mindössze a határozottság keveredik közéjük. Ennek a dimenziónak az ellentétes oldala az evilági élet, ahol az anyagi, protestáns értékek a fontosak.

Tízparancsolat

A Tízparancsolat latens változó manifeszt változói és a faktorelemzés eredményei a következők:



16.6. ábra. Életvitel-változás (első, rotálatlan faktor)

Kérem, mondja meg mindegyik parancsolatról, hogy mennyire érvényes Önre: teljesen, bizonyos mértékig vagy nem érvényes!

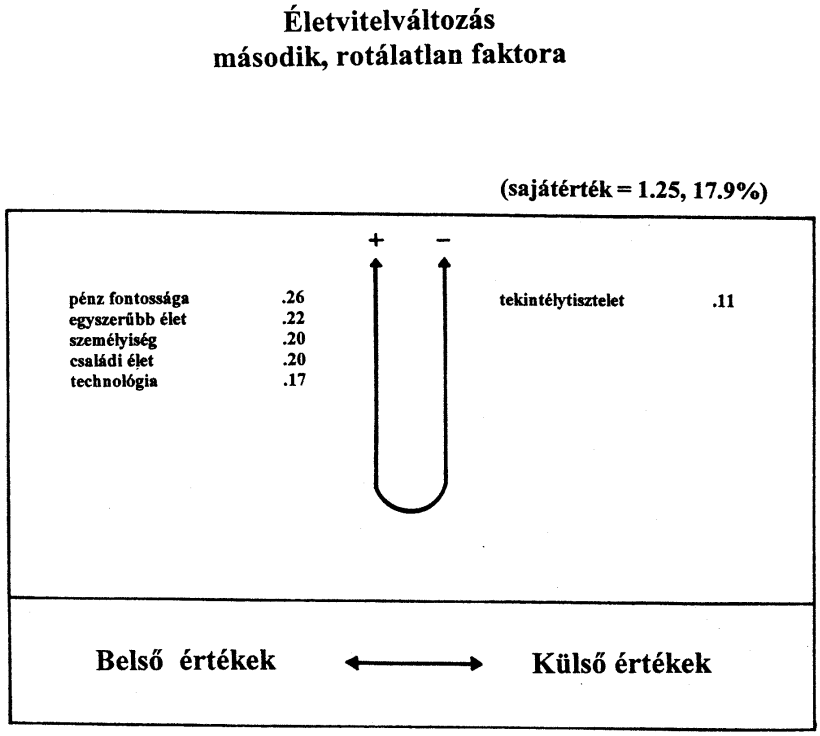
1. Én vagyok a te Istened, ne legyenek néked idegen isteneid előttem!
2. A te Istenednek a nevét hiába fel ne vedd!
3. Megemlékezzél az Úr napjáról, hogy megszenteljed azt!
4. Tiszteld atyádat és anyádat!
5. Ne ölj!
6. Ne paráználkodj!
7. Ne lopj!
8. Ne tégy felebarátod ellen hamis tanúbizonyságot!
9. Ne kívánd a te felebarátod felségét!
10. Se házát, se mezejét, se másféle jószágát!

Életvitel-változás

A kérdés, amit a vizsgált személyeknek feltettünk, a következő volt:

Felsorolok Önnek olyan életvitel-változásokat, amelyek a közeli jövőben bekövetkezhetnek. Kérem, mondja meg mindegyikről, hogy jó, rossz dolognak tartaná vagy nem törődne vele!

1. visszaszorulna a pénz és anyagi javak fontossága
2. csökkenne a munka fontossága életükben



16.7. ábra. Életvitel-változás (második, rotálatlan faktor)

3. a technológia fejlesztése nagyobb hangsúlyt kapna
4. fontosabbá válna az emberi személyiség fejlesztése
5. növekedne a tekintélytisztelet
6. nagyobb hangsúlyt kapna a családi élet
7. egyszerűbb és természetesebb lenne az életvitel

Az első faktor pozitív oldalán a *belső boldogságra törekvés*, a családi egyensúly, a személyiség fejlesztése és a tekintély növekedése, a negatív – bár gyenge oldalon – a *külső boldogságra* törekvés, a pénz, a technológia, az anyagi, hatalmi megelégedettség változói jelennek meg. Az első faktorban megjelenik – a máshol is tapasztalt – belső kontra külső irányultság dichotómiája.

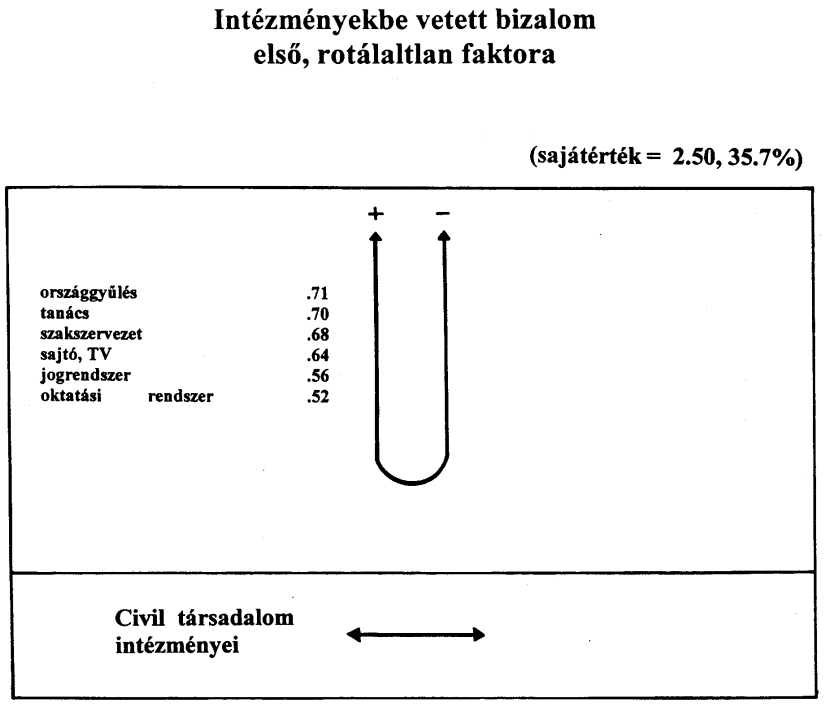
A második faktor lényegében egy posztmodern vágyódás faktorának tekinthető, az egyszerűbb életvitel, a személyiség fejlesztésének a pozitív súlya áll szemben a materiális változókkal.

Intézményekbe vetett bizalom

Felsorolok néhány intézményt. Kérem, mondja meg mindegyikről, hogy mennyire van bizalma benne!

1. az egyházban
2. az oktatási rendszerben
3. a jogrendszerben

4. a sajtóban, televízióban, rádióban
5. a szakszervezetben
6. a vállalatokban
7. az országgyűlésben
8. a tanácsokban



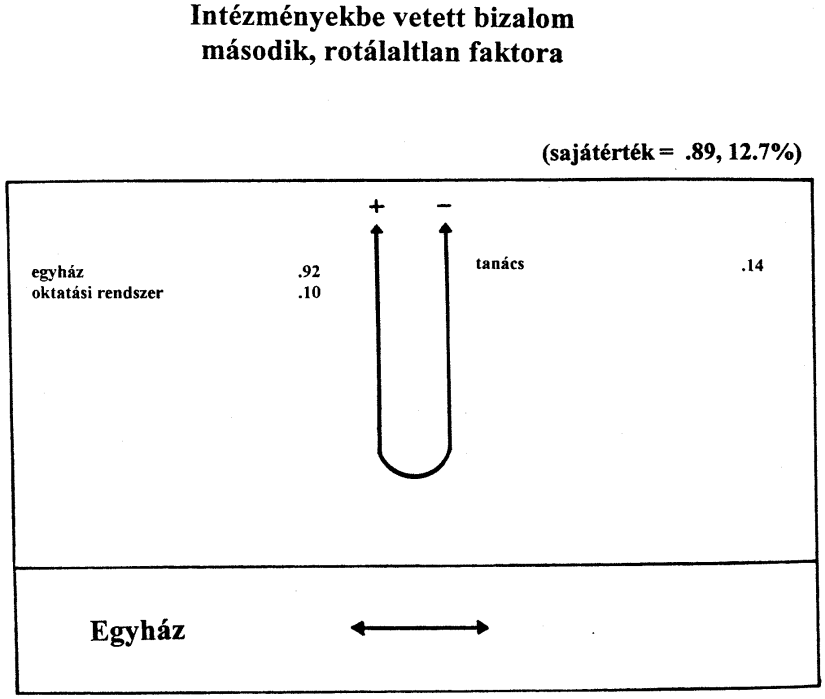
16.8. ábra. Intézményekbe vetett bizalom (első, rotálatlan faktor)

Az intézményekbe vetett bizalom első faktora főfaktornak tekinthető, azzal a lényeges megjegyzéssel, hogy a politikai intézményrendszer elemeihez képest az egyház intézménye nem bipolarizálódik, hanem egy az első civil társadalom intézményeitől független, második faktorban jelenik meg.

A komplex modell vizsgálata

Itt és most csak a végső, statisztikailag (0,01 elsőfajú hibával) illeszkedő modellt ismertetjük, és nem térünk ki a modellváltozatokra. A vizsgálat során azt a determinisztikus modellt elemeztük, hogy a múltbeli társadalmi pozíció és a jelenlegi pozíció hogyan befolyásolja az emberek értékrendjét, és ez az értékrend milyen hatással van az emberek nézetrendszerére, a társadalmi problémákra adott válaszaikra. Ehelyütt csupán két endogén változócsoportot emeltünk ki, az Életvitel-változásra, és az értékrendnek az Intézményekbe vetett bizalomra gyakorolt befolyásoló és/vagy szűrő szerepét modelleztük.

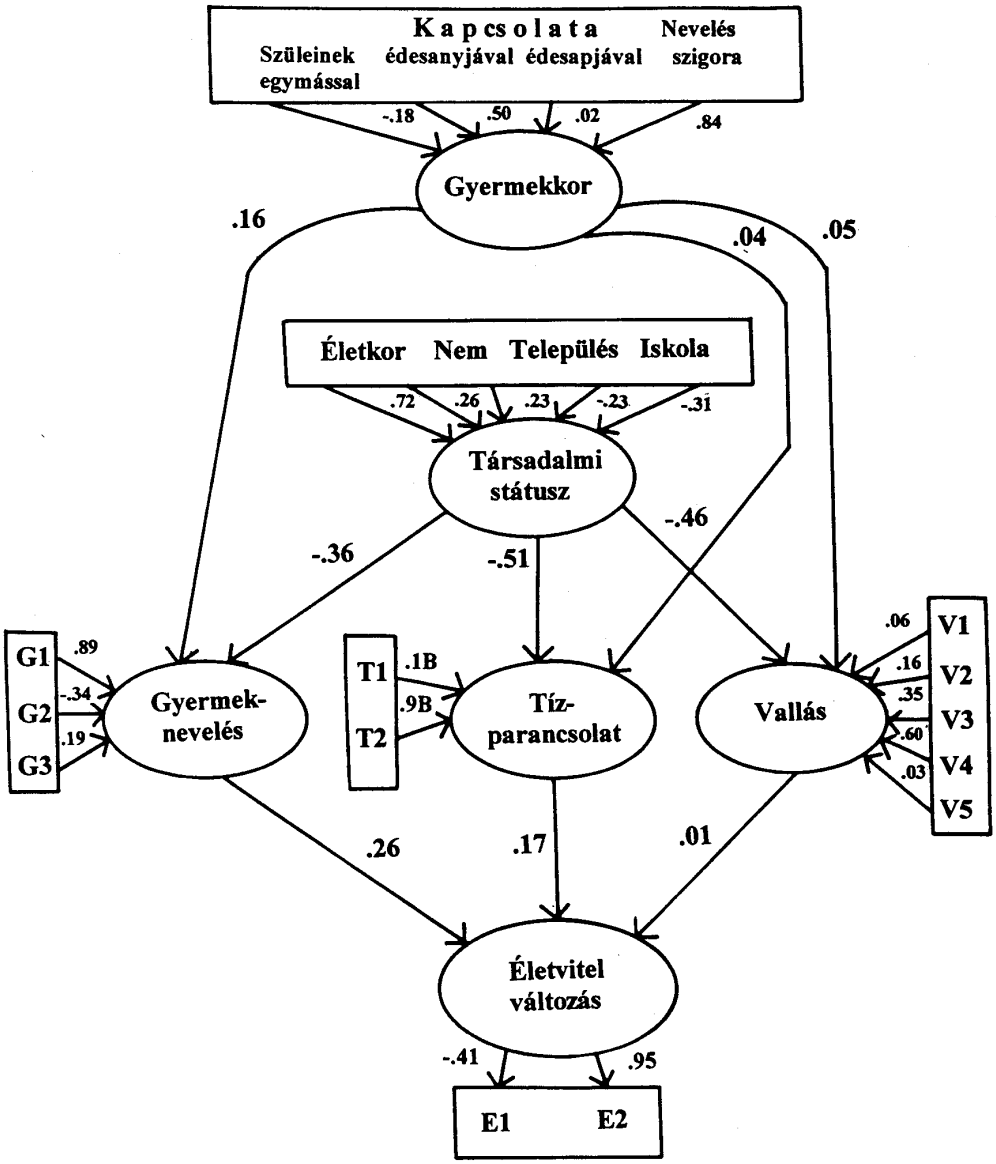
A két LISREL-modell ábráján látható útegyütthatók alapján az exogén latens változók szerepét, hatását érthetjük meg. A múltbeli társadalmi pozíció hatását az emberek



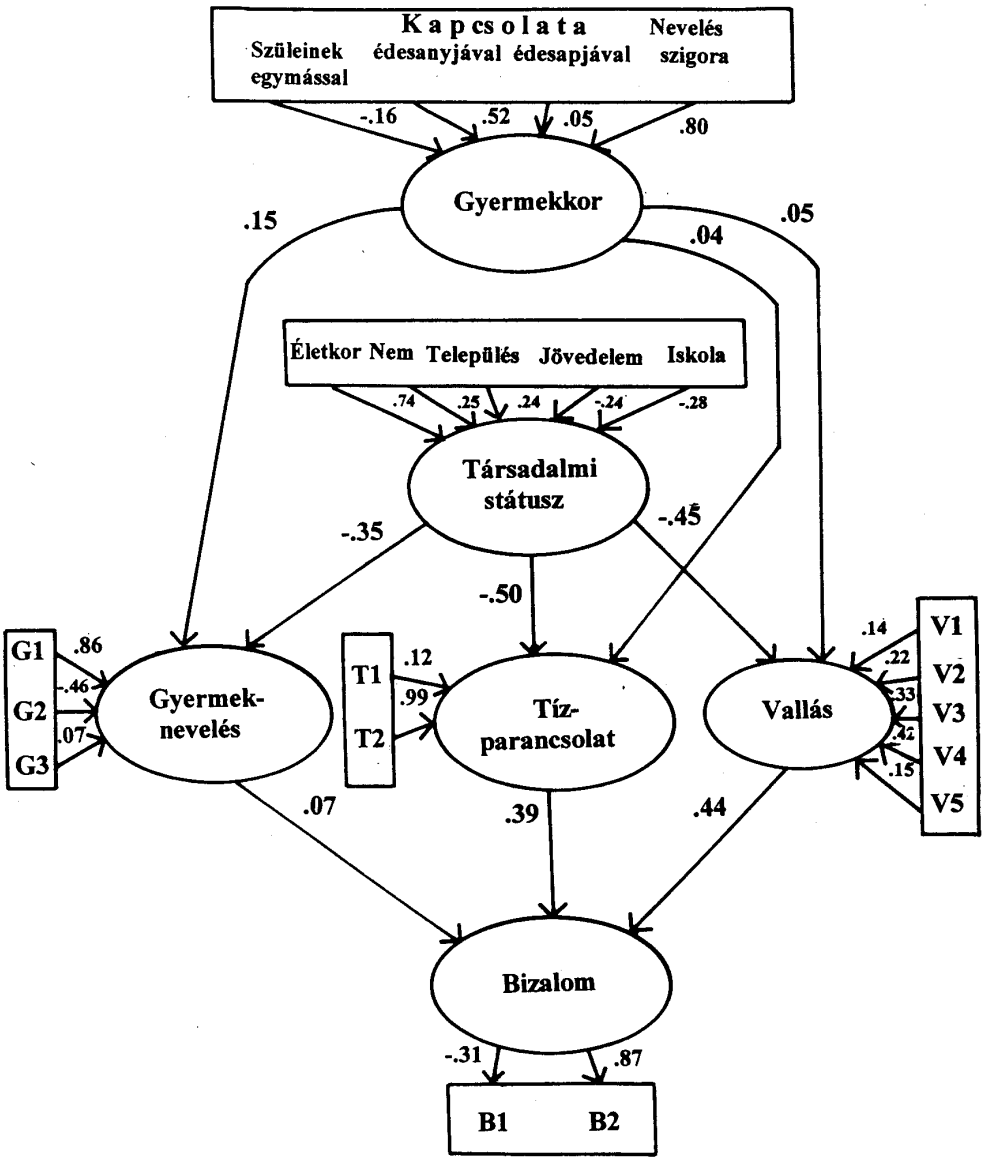
16.9. ábra. Intézményekbe vetett bizalom (második, rotálatlan faktor)

értékrendje jelentősen felerősítette, amikor az Életvitel-változásokra adott válaszokat akartuk becsülni, míg az Intézményekbe vetett bizalom esetén az értékrendnek éppen hogy szűrő szerepe volt, és a Gyermekkori pozíció a vallásos és hagyományos hit és erkölcsi elveken keresztül vezető úton befolyásolta a vizsgált nézetrendszert.

A jelenlegi társadalmi pozíció hatása mindegyik értékrendi blokkra jelentősnek ítélhető, és az értékrend erősítő, fokozó, vagy éppen szűrő szerepe hasonlóképpen működött, mint a múltbeli társadalmi pozíció esetében.



16.10. ábra. LISREL-modell (Életvitel-változás)



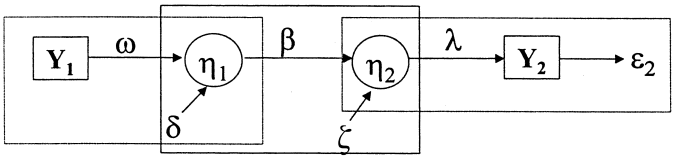
16.11. ábra. LISREL-modell (Bizalom)

17. fejezet

LVPLS-modell

Latent Variables Path Analysis with Partial Least-Squares Estimation)

A latens változók út-modelljének szkémája:



ahol Y_1, Y_2: a megfigyelt, manifeszt változók halmaza, η_1, η_2: nem megfigyelt, latens változók halmaza, β: útegyüttható(k), λ: az endogén manifeszt változók faktorsúlya(i), ω: az exogén manifeszt változók regressziós súlya(i), ζ: a latens endogén változó sztochasztikus reziduális tagja(i), δ: az exogén latens változó reziduális tagja(i), ε: az endogén manifeszt változó mérési hibája(i).

A modell három egyenletből áll.

1. Az első a latens változók közötti utakat írja le, ez a modell strukturális egyenlete:

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\zeta}, \tag{17.1}$$

ahol $E(\eta) = \mathbf{0}$, $E(\zeta) = \mathbf{0}$, $E(\eta\zeta') = \mathbf{0}$.

Az i-edik latens változót előállító strukturális egyenlet:

$$\eta_i = \sum_{j<i} (\beta_{ij}\eta_j) + \zeta_i. \tag{17.2}$$

Feltételezzük, hogy az endogén latens változó (η_i) az ok szerepét betöltő latens változók feltételes várható értéke:

$$E(\eta_i \mid \eta_j) = \sum_{j<i} (\beta_{ij}\eta_j). \tag{17.3}$$

2. A második egyenlet a manifeszt változók mérési modellje:

$$\mathbf{y} = \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\epsilon}, \tag{17.4}$$

ahol Λ: általános eleme λ_{kj} a j-edik latens változó (η_j) regressziós együtthatója a k-adik megfigyelt változó (y_k) előállításában, vagy másképpen a k-adik megfigyelt változónak a j-edik latens változóra vonatkozó faktorsúlya,

ϵ: a megfigyelt változók mérési hibáit, a megfigyelt változók reziduális tagjait tartalmazza.

Feltételezzük, hogy

$$E(\mathbf{y}) = \mathbf{0}, \quad E(\eta) = \mathbf{0}, \quad E(\epsilon) = \mathbf{0}, \quad E(\eta\epsilon') = \mathbf{0}.$$

Az utolsó feltétel azt jelenti, hogy a mérési hibák korrelálatlanok a latens változókkal.

Az egyenlet mérési hiba nélküli része feltételezés szerint a megfigyelt változók latens változóra vonatkozó feltételes várható értékét adja. A k-adik megfigyelt változó feltételes várható értéke:

$$E(y_k \mid \boldsymbol{\eta}) = \sum_j (\lambda_{kj}\eta_j). \tag{17.5}$$

3. A harmadik egyenlet a súlymodell. Ebben a latens változókat a megfigyelt változók regressziós függvényeként állítjuk elő. A súlymodell egyenlete a következő:

$$\boldsymbol{\eta} = \boldsymbol{\Omega}\mathbf{y} + \boldsymbol{\delta}, \tag{17.6}$$

ahol Ω: a súlyegyüttthatókat tartalmazó mátrix, δ: a latens változók reziduális tagjait tartalmazza.

A modellben feltételezzük, hogy

$$E(\boldsymbol{\eta}) = \mathbf{0}, \quad E(\mathbf{y}) = \mathbf{0}, \quad E(\boldsymbol{\delta}) = \mathbf{0}, \quad E(\mathbf{y}\boldsymbol{\delta}') = \mathbf{0}.$$

A j-edik latens változó feltételes várható értéke:

$$E(\eta_j \mid \mathbf{y}) = \sum_k (\omega_{jk} y_k). \tag{17.7}$$

17.1. A strukturális egyenlet redukált formája

A strukturális egyenlet a latens változók kapcsolódásait írja le:

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\zeta}. \tag{17.8}$$

A $\mathbf{B}$ mátrix latens változók egymásra gyakorolt közvetlen hatásait, a közvetlen kapcsolódások útegyütthatóit tartalmazza. Ha ezt az egyenletet átrendezzük úgy, hogy a bal oldalra kerüljenek a latens változók, akkor a strukturális egyenlet redukált formájához jutunk:

$$\boldsymbol{\eta} = (I - \mathbf{B})^{-1}\boldsymbol{\zeta} = \mathbf{B}^*\boldsymbol{\zeta}. \tag{17.9}$$

A redukált forma $\mathbf{B}^*$ mátrix a latens változók egymásra gyakorolt teljes hatásait tartalmazza. A teljes és a közvetlen hatások különbsége a közvetett hatást mutatja, azt a hatást, amelyet a direkt utak mellett az indirekt utakon az egyes latens változók más változókra kifejtenek:

$$\mathbf{B}^+ = \mathbf{B}^* - \mathbf{B}. \tag{17.10}$$

Ehhez hasonlóan kifejezhetjük a latens változónak a megfigyelt változókra vonatkozó teljes és közvetett hatásait is:

$$\mathbf{y} = \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\epsilon} = \boldsymbol{\Lambda}\mathbf{B}^*\boldsymbol{\zeta} + \boldsymbol{\epsilon} = \boldsymbol{\Lambda}^*\boldsymbol{\zeta} + \boldsymbol{\epsilon}, \tag{17.11}$$

ahol Λ: a közvetlen hatásokat, Λ^*: a teljes hatásokat tartalmazza.

Így

$$\boldsymbol{\Lambda}^+ = \boldsymbol{\Lambda}^* - \boldsymbol{\Lambda} \tag{17.12}$$

a latens változók manifeszt változókra vonatkozó közvetett hatásait tartalmazza, amelyek egyenlők nullával az exogén változók esetén.

17.2. A modellben szereplő változók és paraméterek

$\mathbf{y}' = [y_1, y_2, \ldots, y_p]$: a megfigyelt, manifeszt változók p-elemű vektora,

$\eta' = [\eta_1, \eta_2, \ldots, \eta_m]$: a latens változók m-elemű vektora,

$\zeta' = [\zeta_1, \zeta_2, \ldots, \zeta_m]$: a latens változók reziduális komponense (sztochasztikus reziduális tag),

$\epsilon' = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p]$: a manifeszt változók reziduális komponense (mérési hiba),

$\delta' = [\delta_1, \delta_2, \ldots, \delta_m]$: a latens változók reziduális komponense (a súlyegyenletek reziduális tagja),

$\mathbf{B}$: a latens változók útegyüttthatóinak mátrixa, $(m \times m)$ típusú, általános eleme β_{ij} a j-edik latens változónak az i-edik latens változóra kifejtett közvetlen hatását mutatja,

Λ: a megfigyelt (manifeszt) változók faktorsúlymátrixa, $(m \times p)$ típusú, általános eleme λ_{kj} a j-edik latens változó közvetlen hatását fejezi ki a k-adik megfigyelt változóra,

Ω: a latens változók súlyegyüttthatói (regressziós együtthatók), $(m \times p)$ típusú, általános eleme ω_{jk} a k-adik megfigyelt változó közvetlen hatását fejezi ki a j-edik latens változóra,

Ψ: a latens változók sztochasztikus reziduális tagjai között számított variancia-kovarianciamátrix, $(m \times m)$ típusú, általános eleme ψ_{ij} az i-edik és j-edik reziduális komponens kovarianciája ($E(\zeta_i \zeta_j)$, és ψ_{ii} jelöli az i-edik reziduális komponens varianciáját ($E(\zeta_i \zeta_i)$),

Θ: a megfigyelt változók mérési hibáinak variancia-kovarianciamátrixa, $(p \times p)$ típusú, általános eleme $\theta_{k\ell}$ a k-adik és ℓ-edik megfigyelt változó mérési hibája között számított kovarianciát, a θ_{kk} a k-adik mérési hiba (reziduális tag) varianciáját jelöli.

17.3. A modellben szereplő változók variancia-kovarianciamátrixai

Mivel a megfigyelt változókat a mérési modellben a latens változók függvényeiként fejeztük ki, nézzük először a latens változók kovarianciamátrixát:

$$\mathbf{C} = E(\eta\eta') = \text{cov}(\eta) = (\mathbf{I} - \mathbf{B})^{-1}\Phi(\mathbf{I} - \mathbf{B}')^{-1} = \mathbf{B}^*\Phi\mathbf{B}^{*'}, \tag{17.13}$$

ahol Φ: a sztochasztikus reziduális tag variancia-kovarianciamátrixa.

A paraméter becslésénél feltételezzük, hogy a latens változók standardizáltak, így a kovariancia egyenlő lesz a korrelációval. A fentieket felhasználva a megfigyelt változók variancia-kovarianciamátrixa:

$$\mathbf{\Sigma} = \text{cov}(\mathbf{y}) = \mathbf{\Lambda C \Lambda}' + \mathbf{\Theta}_\varepsilon. \tag{17.14}$$

A megfigyelt változók variancia-kovarianciamátrixa a Λ, $\mathbf{C}$ és Θ_ε paraméterek függvénye.

Számíthatjuk még a megfigyelt változók és a latens változók közötti páronkénti kovarianciákat (standardizált változók esetén korrelációkat), a latens változók struktúráját:

$$\mathbf{T} = \text{cov}(\mathbf{y}\boldsymbol{\eta}') = \boldsymbol{\Lambda}\mathbf{C} = \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Phi}(\mathbf{I} - \mathbf{B}')^{-1}. \tag{17.15}$$

A sztochasztikus reziduális tag (ζ) variancia-kovarianciamátrixa:

A mérési hiba variancia-kovarianciamátrixa:

$$\boldsymbol{\Theta}_{\varepsilon} = \mathbf{S} + \boldsymbol{\Lambda}\mathbf{C}\boldsymbol{\Lambda}' - \boldsymbol{\Lambda}\mathbf{T}' - \mathbf{T}\boldsymbol{\Lambda}'. \tag{17.17}$$

A mérési hiba és a megfigyelt változók variancia-kovarianciamátrixa:

$$\text{cov}(\boldsymbol{\epsilon}\mathbf{y}) = \mathbf{T} - \boldsymbol{\Lambda}\mathbf{C}. \tag{17.18}$$

A mérési hiba és a sztochasztikus reziduális tag variancia-kovarianciamátrixa:

$$\text{cov}(\boldsymbol{\epsilon}\boldsymbol{\zeta}) = (\mathbf{T} - \boldsymbol{\Lambda}\mathbf{C})(\mathbf{I} - \mathbf{B}'). \tag{17.19}$$

A latens változó ($\widehat{\eta}$) magyarázott részének ($\mathbf{B}\widehat{\eta}$) kovarianciái:

$$\mathbf{C}^* = \text{cov}(\mathbf{B}\widehat{\boldsymbol{\eta}}) = \mathbf{B}\mathbf{C}\mathbf{B}'. \tag{17.20}$$

A $\mathbf{C}^*$ mátrix diagonálisában az endogén latens változók többszörös korrelációi négyzetét tartalmazza, ami a latens változók varianciáinak az a része, amit a becslő latens változók magyarázni tudnak.

A $\mathbf{C}^*$ diagonális elemei:

$$\mathbf{R}^2 = \mathbf{I} \times (\mathbf{B}\mathbf{C}\mathbf{B}'), \tag{17.21}$$

ahol $\times$ az ún. Hadamard-szorzatot jelöli.[1]

A megfigyelt változók mérési modellje a megfigyelt változókat két részre bontja: a közös faktorok hatására ($\boldsymbol{\Lambda}\boldsymbol{\eta}$) és a mérési hibára ($\boldsymbol{\epsilon}$). A közös, a latens változók (faktorok) által magyarázott rész kovarianciái:

$$\mathbf{H}^2 = \text{cov}(\boldsymbol{\Lambda}\widehat{\boldsymbol{\eta}}) = \boldsymbol{\Lambda}\mathbf{C}\boldsymbol{\Lambda}'. \tag{17.22}$$

A $\mathbf{H}$ diagonális elemei a megfigyelt változók varianciáival standardizálva:

$$\mathbf{H}^2 = (\mathbf{I} \times (\boldsymbol{\Lambda}\mathbf{C}\,\boldsymbol{\Lambda}'))(\mathbf{I} \times \mathbf{S})^{-1} \tag{17.23}$$

a megfigyelt változók kommunalitásait adják. A h_k^2 a k-adik megfigyelt változó (y_k) varianciájának az a része, amelyet a megfigyelt változóhoz közvetlenül kapcsolódó latens változók reprodukálnak.

Kapcsoljuk össze a mérési modellt és a strukturális egyenletet:

$$\mathbf{y} = \boldsymbol{\Lambda}\boldsymbol{\beta}\boldsymbol{\eta} + \boldsymbol{\Lambda}\boldsymbol{\zeta} + \boldsymbol{\epsilon}. \tag{17.24}$$

Azt látjuk, hogy a manifeszt változók $\mathbf{y}$ közös része két részre bontható: az egyedi részre $\boldsymbol{\Lambda}\boldsymbol{\zeta}$ és a redundáns részre $\boldsymbol{\Lambda}\boldsymbol{\beta}\boldsymbol{\eta}$.

Az egyedi rész a megfigyelt változónak az a része, amit a vele közvetlenül összekötött latens változók önmagukban reprodukálnak. Az y_k redundáns részét akár a vele közvetlenül összekötött latens változók, akár ezek prediktorai reprodukálhatják.

A megfigyelt változók redundáns részeinek kovarianciamátrixa:

$$\mathbf{F} = \text{cov}(\boldsymbol{\Lambda}\mathbf{B}\boldsymbol{\eta}) = \boldsymbol{\Lambda}\mathbf{B}\mathbf{C}\mathbf{B}'\boldsymbol{\Lambda}'. \tag{17.25}$$

[1] A Hadamard-szorzatot a következőképpen definiáljuk: $\mathbf{A} = \mathbf{B} \times \mathbf{C}$ ha $a_{jk} = b_{jk}c_{jk}$ bármely j, k-ra.

Az $\mathbf{F}$ mátrix diagonális elemeit standardizáljuk a megfigyelt változók varianciáival:

$$\mathbf{F}^2 = (\mathbf{I} \times (\mathbf{\Lambda B C B}'\mathbf{\Lambda}'))(\mathbf{I} \times \mathbf{S})^{-1}. \tag{17.26}$$

Az $\mathbf{F}^2$ mátrix diagonális elemei a megfigyelt változók redundanciaegyütthatói. A redundancia-együttható a megfigyelt változó varianciájának az a része, amit a megfigyelt változóval közvetetten kapcsolódó latens változók magyaráznak. Az exogén megfigyelt változók redundanciája nulla.

A (17.24) egyenletbe behelyettesítjük a latens változók becsléseit ($\widehat{\eta} = \mathbf{\Omega y}$):

$$\mathbf{y} = \mathbf{\Lambda B \Omega y} + \mathbf{\Lambda}\zeta + \epsilon, \tag{17.27}$$

ahol a megfigyelt változók redundáns része:

$$\widehat{\mathbf{y}} = \mathbf{\Lambda B \Omega y} = \mathbf{My}. \tag{17.28}$$

A redundáns rész kovarianciamátrixa:

$$\begin{aligned} \operatorname{cov}(\widehat{y}) &= \mathbf{\Lambda B \Omega S \Omega}'\mathbf{B}'\mathbf{\Lambda}' \\ &= \mathbf{\Lambda B C B}'\mathbf{\Lambda}' \\ &= \operatorname{cov}(\mathbf{\Lambda B}\eta) = \text{ a (17.25) egyenlet szerint.} \end{aligned}$$

17.4. A paraméterek becslése a parciális legkisebb négyzetek módszerével

A legkisebb négyzetek módszere a statisztikában jól ismert eljárás. A parciális jelzővel ellátott becslési módszer a klasszikus kritérium kiterjesztését jelöli. A parciális legkisebb négyzetek módszer lényege, hogy a paramétereket részhalmazokra bontja, particionálja, az egyes partíciókat a legkisebb négyzetek módszere kritériuma szerint becsüli, miközben a többi paraméter kötött értékkel szerepel. Az eljárás iteratív, az iterációs ciklus mindegyikében a paraméterek egy-egy részhalmazát becsüli a többi paraméter értékét ismertnek feltételezve.

A parciális legkisebb négyzetek módszerénél a következő előfeltételezéseket teszszük:

1) A manifeszt változók egymást át nem fedő (diszjunkt) blokkokra vannak particionálva.

2) A latens változók is egymást át nem fedő blokkokra vannak particionálva, és egy latens változó-blokk csak egy manifeszt változó-blokkhoz kapcsolódhat, vagyis a manifeszt és latens változók kapcsolódásait egy blokk-diagonális mátrix írja le.

3) Az út-modell (a strukturális egyenletrendszer) rekurzív.

A becslés technikai feltételei:

1) A latens változókat a manifeszt változók lineáris kombinációjával becsüljük:

$$\widehat{\eta}_j = \sum_k (w_{jk} y_k), \tag{17.29}$$

ahol w_{jk} a ω_{jk} becslése.

2) A latens változók strukturális egyenleteit (útegyenleteket) a reziduális tag varianciájának minimalizálásával becsüljük:

$$\mathrm{var}(\zeta_k) = \Psi_k^2 \longrightarrow \min . \tag{17.30}$$

3) A modell strukturális egyenleteken kívüli részét blokkonként vagy

a) a súlymodell szerint $E(\eta_j \mid \mathbf{y}) = \sum_k (\omega_{jk} y_k)$, a súlyegyütthatókat a reziduális tag varianciájának minimalizálásával becsüljük:

$$\mathrm{var}(\delta_j) \longrightarrow \min, \tag{17.31}$$

vagy

b) a mérési modell szerint $E(y_k \mid \eta) = \sum_j (\lambda_{kj} \eta_j)$, a faktorsúly-együtthatókat a reziduális tag varianciájának minimalizálásával becsüljük:

$$\mathrm{var}(\varepsilon_k) \longrightarrow \min . \tag{17.32}$$

A parciális legkisebb négyzetek módszer algoritmusa iteratív eljárással ad becslést a paraméterekre. Minden iterációs ciklus három lépésből áll.

Az iterációs eljárás lépései:

1. lépés: becsüljük a latens változókat vagy a súlymodell (17.18) kritériuma, vagy a mérési modell (17.19) kritériuma szerint: $\widehat{\boldsymbol{\eta}} = \mathbf{W}\mathbf{y}$.

Kiszámítjuk a latens változók kovarianciamátrixának becslését:

$$\widehat{\mathbf{C}} = \mathbf{W}\mathbf{S}\mathbf{W}', \tag{17.33}$$

ahol $\mathbf{S}$ a manifeszt változók variancia-kovarianciamátrixa, valamint a latens változók és a megfigyelt változók páronkénti kovarianciáit tartalmazó struktúramátrix becslése:

$$\widehat{\mathbf{T}} = \mathbf{S}\mathbf{W}'. \tag{17.34}$$

Az eljárás legelső lépésében a súlymátrix ($\mathbf{W}$) elemeit pszeudo-véletlen számokkal feleltetjük meg.

2. lépés: amennyiben a latens változók egymás közti kapcsolódásaira feltételekkel élünk (közöttük korrelálatlanságot írunk elő), a második lépésben a latens változókat transzformáljuk a „patterned orthonormalizing rotation” eljárás szerint.

3. lépés: becsüljük a latens változók belső súlyegyütthatóit:

$$\boldsymbol{\eta}^* = \mathbf{V}\widehat{\boldsymbol{\eta}}.$$

A belső súlyegyütthatókat a következő módokon adhatjuk meg:

a) *útegyüttható*ként, amikor a belső strukturális modell út-modell. Ebben az esetben az exogén latens változó optimális becslő, az endogén latens változó optimális becsült, és a predeterminált latens változó optimális közvetítő változó.

$$v_{hh'} = \begin{cases} b_{hh'}, & \text{ha } \eta_{h'} \text{ megelőzi } \eta_h\text{-t} \\ r_{hh'}, & \text{ha } \eta_{h'} \text{ követi } \eta_h\text{-t} \\ 0, & \text{ha } \eta_{h'} \text{ és } \eta_h \text{ nincs közvetlenül összekötve.} \end{cases} \tag{17.35}$$

b) *centroid együttható*ként, amikor azt figyeljük, hogy a latens változókhoz mely latens változók kapcsolódnak közvetlenül. A latens változót az útdiagramban szomszédos latens változók súlyozatlan összegével becsüljük. Hogy az egymással negatívan korreláló szomszédos latens változók ne semlegesítsék egymás hatását, az előjelfüggvényt alkalmazzuk:

$$v_{hh'} = \begin{cases} \mathrm{sign}(r_{hh'}), & \text{ha } \eta_h \text{ és } \eta_{h'} \text{ közvetlenül össze van kötve} \\ 0 & \text{egyébként.} \end{cases} \tag{17.36}$$

Ezzel a feltétellel a latens változók úgy korrelálnak egymással, hogy a centroid faktornak (amit másodrendű faktornak tekinthetünk) maximális legyen a varianciája. (A centroid faktor faktorsúlyai 1, 0, –1 értéket vehetnek fel.)

c) Faktorsúlyként, amikor az a célunk, hogy a latens változók főkomponenseinek varianciái maximálisak legyenek. Ekkor az együtthatók:

$$v_{hh'} = \begin{cases} r_{hh'}, & \text{ha } \eta_h \text{ és } \eta_{h'} \text{ közvetlenül össze van kötve} \\ 0 & \text{egyébként.} \end{cases} \tag{17.37}$$

4. lépés: Becsüljük a belső súlyegyütthatókkal a latens változókat:

$$\boldsymbol{\eta}^* = \mathbf{V}\widehat{\boldsymbol{\eta}}.$$

5. lépés: Ortogonálisan transzformáljuk a becsült latens változókat (η^*), ha a latens változók páronkénti kapcsolódásaira feltételezésekkel élünk.

6. lépés: Becsüljük a külső súlyokat ($\mathbf{W}$) a reziduális tag minimalizálásával:

$$\widehat{\boldsymbol{\eta}} = \mathbf{W}\mathbf{y}.$$

Az iteráció a 6 lépés ismétléséből áll, és akkor áll le, ha a két egymás utáni sorozatban a súlyegyütthatókra a következő teljesül:

$$|w^{(s)} - w^{(s+1)}| < 10^{-5},$$

vagy

$$|(w^{(s)} - w^{(s+1)})/w^{(s)}| < 10^{-5}.$$

17.5. A becslés illeszkedésének mérése

A becslés illeszkedésére nincs egyetlen indexünk, de a becslés jóságát megítélhetjük a következő mennyiségekkel, és így dönthetünk az érvényességéről:

$$h^2 = \text{trace}(\mathbf{H}^2)/p \quad \text{kommunalitás}, \tag{17.38}$$

$$f^2 = \text{trace}(\mathbf{F}^2)/p \quad \text{redundancia}, \tag{17.39}$$

$$c^2 = \text{trace}(\mathbf{C}^2)/m, \tag{17.40}$$

$$s^2 = \text{trace}(\mathbf{S})/p, \tag{17.41}$$

$$r^2 = \text{trace}(\mathbf{C})/m, \tag{17.42}$$

$$\Theta^2 = \text{trace}(\boldsymbol{\Theta})/m, \tag{17.43}$$

$$\Psi^2 = \text{trace}(\boldsymbol{\Phi})/m, \tag{17.44}$$

$$s = \text{rms}\mathbf{S}, \tag{17.45}$$

$$c = \text{rms}\mathbf{C}, \tag{17.46}$$

$$\Theta = \text{rms}\boldsymbol{\Theta}, \tag{17.47}$$

$$\Psi = \text{rms}\boldsymbol{\Phi}, \tag{17.48}$$

$$\text{ceh} = \text{rms cov}(\mathbf{e}\boldsymbol{\eta}), \tag{17.49}$$

$$\text{ceh} = \text{rms cov}(\mathbf{e}\boldsymbol{\zeta}), \tag{17.50}$$

$$\overline{c} = \sum_{h}^{-} \sum_{h<h'}^{-} |c_{hh'}|, \tag{17.51}$$

$$t = \text{rms}\mathbf{T}. \tag{17.52}$$

A fenti mennyiségekből[2] h^2, f^2 és c^2 arányszámok a variancia arányait jelölik. Ha a megfigyelt változók standardizáltak, akkor $\Theta^2 = s^2 - h^2 = 1 - h^2$ is arányszám. A paraméterek maximum likelihood becslésének veszteségfüggvényét

$$L = \log|\mathbf{\Sigma}| + \text{tr}(\mathbf{S}\mathbf{\Sigma}^{-1}) - \log|\mathbf{S}| - p \tag{17.53}$$

használhatjuk a modell tesztelésére. Ha a megfigyelt változók együttes eloszlása normális, az $(n-1)L$ eloszlása közelítőleg χ^2 eloszlású (n a megfigyelések száma)

$$f_t = (p^2 + p)/2 - t \tag{17.54}$$

szabadságfokkal, ahol t a modell független paramétereinek a számát jelöli.

Tekintsük a legegyszerűbb modellt:

$$H_d : \mathbf{\Sigma} = \mathbf{S} \times \mathbf{I}, \tag{17.55}$$

ahol a megfigyelt változókról feltételezzük, hogy páronként korrelálatlanok. A veszteségfüggvény egyszerűsödik, mivel $\mathbf{S}\mathbf{\Sigma}^{-1}$ mátrix egységmátrix, aminek a nyoma (trace) egyenlő p-vel, így

$$L_d = \log|\mathbf{I} \times \mathbf{S}| - \log|\mathbf{S}|. \tag{17.56}$$

Ha a változók standardizáltak is, akkor $(\mathbf{I} \times \mathbf{S}) = \mathbf{I}$, és így $\log|\mathbf{I} \times \mathbf{S}| = 0$, és $L_d = -\log|\mathbf{S}|$. A szabadságfok $f_d = (pp - p)/2$. Tekintsük továbbá a korlátozott faktoranalitikus modellt:

$$H_r : \mathbf{\Sigma} = \mathbf{\Sigma}(\mathbf{C}, \mathbf{\Lambda}, \mathbf{\Theta} = \mathbf{I} \times \mathbf{\Theta}) \tag{17.57}$$

az L_r veszteségfüggvénnyel, ahol a latens változók korrelációi és a faktorsúlyok közül rögzíthetünk értékeket (feltételezhetjük, hogy egyenlőek nullával), és feltételezzük, hogy a reziduális kovarianciamátrix diagonális.

A H_r és H_s hipotézissel megfogalmazott modellek illeszkedésének különbözőségére Tucker és Lewis (1973) a következő megbízhatósági indexet definiálta:

$$r_{TL} = \frac{L_s/f_s - L_r/f_r}{L_d/f_d - 1/N}. \tag{17.58}$$

Vagy a Bentler és Bonnet (1980)-féle megbízhatósági index:

$$r_{BB} = \frac{L_s - L_r}{L_d}. \tag{17.59}$$

Mindkét megbízhatósági indexnél a modellek különbségeit a legegyszerűbb (a legkorlátozottabb) modellhez viszonyítottuk.

A latens útmodell mérési és súlymodell részét (a modell külső részét, egyenleteit) a mérési adatokhoz kielégítően illeszkedőnek akkor tekintjük,

[2] Megjegyzés: az rms**A** az **A** mátrix diagonális nélküli elemeinek négyzetes átlagát jelöli (root mean squared (rms)), ahol a szumma jel feletti vonás az átlag operátor jele.

$$\text{rms}\ \mathbf{A} = \left[\sum_{i}^{-} \sum_{\neq j}^{-} a_{ij}^2 \right]$$

– ha a manifeszt változók varianciáinak nem megmagyarázott része $(1 - h^2)$ elég kicsi,

– ha a manifeszt változók kovarianciáinak nem megmagyarázott része (Θ/s) elég kicsi,

– ha a h_k^2 (az y_k változók kommunalitása) elég nagy, és a Θ_k^2 reziduális kovarianciák elég kicsik a mérési modellben,

– ha a blokkok közötti reziduális kovarianciák $(\Theta_{kk'})$ a nullához közelítenek,

– ha a mérési hiba (a reziduális tag) és a latens változók közötti kovarianciák cov($\mathbf{e}\eta$) közelítőleg nullával egyenlőek,

– ha egy blokkban csak egy manifeszt változó van, akkor a reziduális tag varianciája egyenlő nullával,

– ha egy blokkban csak két manifeszt változó van, akkor a reziduális változók tökéletesen korreláltak lesznek,

– ha egy blokkban csak kevés latens változó szerepel, a reziduális változók negatívan korrelálnak.

A modell sturkturális (belső) részének illeszkedése kielégítő,

– ha a latens változók nem megmagyarázott része $(1 - r^2)$ elég kicsi,

– ha a latens változók kovarianciáinak nem megmagyarázott része (Ψ/c) elég kicsi.

A teljes modell illeszkedését kielégítőnek tekintjük,

– ha a redundancia-együttható (f^2) elég nagy,

– ha a belső és külső reziduális tagok közötti kovarianciák (cov($\mathbf{e}\zeta$)) elég kicsik.

17.6. Kategorikus változók

Ismeretes, hogy a dichotom változó két kategóriájához bármely értéket (természetesen két különbözőt) hozzárendelhetünk. Ha a dichotom változó két lehetséges értékéhez a 0 és az 1 értékeket rendeljük, akkor a változót Boole-változónak nevezzük. A Boole-változót formálisan a következőképpen definiáljuk:

$$P(x = 1) = 1 - P(x = 0) = \mu. \tag{17.60}$$

Ebből a definícióból következik, hogy

$$E(x) = \Sigma P(x = i)i = \mu 1 + (1 - \mu)0 = \mu, \tag{17.61}$$

$$E(x^2) = \Sigma P(x = i)i^2 = \mu 1^2 + (1 - \mu)0^2 = \mu, \tag{17.62}$$

és általában

$$E(x^r) = \Sigma P(x = i)i^r = \mu 1^r + (1 - \mu)0^r = \mu, \quad \text{ha} \quad r > 0. \tag{17.63}$$

Az x változó varianciája:

$$\sigma^2 = \mu(1 - \mu) = \mu - \mu^2. \tag{17.64}$$

Egy kategorikus változó kategóriái legfeljebb megszámlálhatóan végtelen számosságú halmazt alkotnak. A k-adik kategória bekövetkezésének valószínűsége:

$$P(x = k) = \mu_k. \tag{17.65}$$

A kategorikus változót helyettesítjük Boole-változók egy halmazával a következő előírás szerint:

$$x = k \Longleftrightarrow x_k = 1. \tag{17.66}$$

A fentiekből következik, hogy

$$E(x_k) = E(x_k^2) = E(x_k^\nu) = P(x = k) = \mu. \tag{17.67}$$

vagyis a várható értéket mint valószínűséget értelmezhetjük. Boole-változók bármely párjának, hármasának stb várható értéke:

$$E(x_k x_{k'} x_{k''} \ldots) = \begin{cases} \mu, & \text{ha } k = k' = k'' = \ldots \\ 0 & \text{különben.} \end{cases} \tag{17.68}$$

Ha az $\mathbf{x}$ vektor tartalmazza a Boole-változókat, akkor a (17.67) a következőképpen írható fel:

$$E(\mathbf{x}) = \boldsymbol{\mu}. \tag{17.69}$$

Tekintsünk most két Boole vektor változót, $\mathbf{x}$ és $\mathbf{y}$-t az x és az y kategorikus változók helyettesítéseiként. Az $\mathbf{x}$ és és $\mathbf{y}$ valószínűségeloszlása:

$$P(x_k = 1,\ y_\ell = 1) = \mu_{k\ell}, \tag{17.70}$$

ahol $k = 1, 2, \ldots q$, $\ell = 1, 2, \ldots, p$.

A várható értékek:

$$E(x_k y_\ell) = \mu_{k\ell} \quad \text{és} \quad E(\mathbf{x}\mathbf{y}') = \boldsymbol{\mu}_{xy}. \tag{17.71}$$

A $\boldsymbol{\mu}_{xy}$ egy $(q \times p)$ típusú mátrix, a kétváltozós valószínűségek táblázata vagy más néven kontingenciatáblázat. Helyezzük az $\mathbf{x}$ és $\mathbf{y}$ vektorokat az $\mathbf{u}$ vektorba:

$$\mathbf{u}' = [\mathbf{x}', \mathbf{y}'] = [\ldots x_k \ldots, \ldots y_\ell \ldots].$$

Az $\mathbf{u}$ vektor várható értéke:

$$E(\mathbf{u}) = \boldsymbol{\mu}_u = [\mathbf{u}'_x, \boldsymbol{\mu}'_y] = [\ldots \mu_k \ldots, \ldots \mu_\ell \ldots].$$

Ezek után definiáljuk a szuper-kontingenciatáblázatot:

$$E(\mathbf{u}\mathbf{u}') = E\begin{pmatrix} \mathbf{x}\mathbf{x}' & \mathbf{x}\mathbf{y}' \\ \mathbf{y}\mathbf{x}' & \mathbf{y}\mathbf{y}' \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_{xx} & \boldsymbol{\mu}_{xy} \\ \boldsymbol{\mu}_{yx} & \boldsymbol{\mu}_{yy} \end{pmatrix}. \tag{17.72}$$

A szuper-kontingenciatáblázat tartalmazza a kontingenciatáblázatot, a $\boldsymbol{\mu}_{xy}$-t, ennek transzponáltját, $\boldsymbol{\mu}_{yx}$-t, és a $\boldsymbol{\mu}_{xx}$ és $\boldsymbol{\mu}_{yy}$ diagonális mátrixokat. A $\boldsymbol{\mu}_{xx}$ diagonális mátrix diagonális elemei az egyváltozós valószínűségek ($\mu_{kk} = \mu_k$ a (17.68) szerint). Általában q kategorikus változó Boole-változóinak szuper-kontingenciamátrixa q^2 szubmátrixot (blokkot) tartalmaz.

Az x és y kategorikus változók megfigyeléseit x_i és y_i jelölje. A helyettesítő Boole-változók megfigyelt értékeit ekkor x_{ki} és $y_{\ell i}$ jelöli. A Boole-változók összege:

$$n_k = \sum_i x_{ki}, \quad n_\ell = \sum_i y_{\ell i}, \tag{17.73}$$

és átlaga:

$$m_k = \overline{\Sigma}_i x_{ki}, \quad m_\ell = \overline{\Sigma}_i y_{\ell i}, \tag{17.74}$$

ahol $\overline{\Sigma}$: az átlag műveletet jelöli.

A relatív szorzat momentummátrixa:

$$\mathbf{M} = \frac{1}{n}\begin{pmatrix} \mathbf{x}\mathbf{x}' & \mathbf{x}\mathbf{y}' \\ \mathbf{y}\mathbf{x}' & \mathbf{y}\mathbf{y}' \end{pmatrix}. \tag{17.75}$$

Definiáljuk még a normált szorzatmátrixot:

$$\mathbf{Q} = \left[q_{k\ell} = \frac{m_{k\ell}}{\sqrt{m_k m_\ell}} \right], \tag{17.76}$$

a kovarianciamátrixot:

$$\mathbf{C} = [c_{k\ell} = m_{k\ell} - m_k m_\ell], \tag{17.77}$$

a normált eltérésmátrixot:

$$\mathbf{D} = \left[d_{k\ell} = \frac{m_{k\ell} - m_k m_\ell}{\sqrt{m_k m_\ell}} \right], \tag{17.78}$$

és a korrelációmátrixot:

$$\mathbf{R} = \left[r_{k\ell} = \frac{m_{k\ell} - m_k m_\ell}{\sqrt{(m_k - m_k^2)(m_\ell - m_\ell^2)}} \right]. \tag{17.79}$$

A normált eltérés mátrix felhasználásával számíthatjuk a Yule-féle (phi) együttható négyzetét:

$$\Phi^2 = \frac{1}{n}\chi^2 = \sum_k \sum_\ell d_{k\ell}^2, \tag{17.80}$$

és a kontingenciaegyütthatókat:

$$C_{\text{Cramer}} = \sqrt{\Phi^2/(\min(pq) - 1)} \tag{17.81}$$

$$C_{\text{Pearson}} = \sqrt{\Phi^2/(1 + \Phi^2)} = \sqrt{\chi^2/(n + \chi^2)}. \tag{17.82}$$

17.6.1. Kanonikus korreláció

Legyen adva két változóhalmaz, $x_k \quad (k = 1, 2, \ldots, q)$ és $y_\ell \quad (\ell = 1, 2, \ldots, p)$. Adottak továbbá a szorzatmátrixok $\mathbf{M}_{xx}$ és $\mathbf{M}_{yy}$, valamint a várható értékek vektorai: $\mathbf{m}_k = \mathbf{0}$ és $\mathbf{m}_y = \mathbf{0}$.

A kanonikus korreláció modellje a következő egyenletekből áll:

$$\xi = \sum_k w_k x_k, \quad \text{var}(\xi) = \mathbf{w}'\mathbf{M}_{xx}\mathbf{w} = 1, \tag{17.83}$$

$$\eta = \sum_\ell v_\ell y_\ell, \quad \text{var}(\eta) = \mathbf{v}'\mathbf{M}_{yy}\mathbf{v} = 1, \tag{17.84}$$

$$\eta = \rho\xi + \zeta, \quad E(\eta|\xi) = \rho\xi, \tag{17.85}$$

ahol ξ és η a kanonikus latens változók egységnyi varianciával, $\mathbf{w} = [w_k]$ és $\mathbf{v} = [v_\ell]$ a kanonikus súlyegyütthatók, és ρ az első kanonikus korreláció.

A kanonikus korreláció modell skálafüggetlen, ami azt jelenti, hogy ha például x_k-t megszorozzuk $1/c$ konstanssal és a w_k értékét c-vel, akkor a (17.85) változatlan marad.

A kanonikus korreláció modellben annyi kanonikus változópár állítható elő, amennyi az x_k és y_k változóhalmaz elemeinek a minimuma ($\min(p, q)$).

Bizonyítható (lásd Lancaster, 1969), hogy a kanonikus korrelációs együtthatók négyzetösszege megegyezik a Yule-féle együtthatóval:

$$\Phi^2 = \sum_i^{\min(p,q)} \rho_i^2, \tag{17.86}$$

ami azt is jelenti, hogy a kanonikus korrelációs együtthatók négyzetes átlagának négyzetgyöke egyenlő a Cramer-féle kontingenciaegyütthatóval.

17.6.2. Főkomponens-elemzés

Legyen adva a dichotom változók normált vektora:

$$\mathbf{u}' = [u_k] = [\ldots x_{k(Q)} \ldots, \ldots y_{\ell(Q)} \ldots],$$

ekkor a dichotom változók lehetséges értékei: $\left(0, \frac{1}{\sqrt{m_k}}\right)$.

Adottak továbbá a modell egyenletei:

$$u_k = p_k \eta + e_k \quad E(u_k|\eta) = p_k \eta, \tag{17.87}$$

$$\eta = \sum_k w_k u_k \qquad \text{var}(\eta) = \mathbf{w}'\mathbf{Q}\mathbf{w}. \tag{17.88}$$

A $\mathbf{Q}$ mátrix sajátértéke

$$d_{(Q)} = \sum_k p_k^2,$$

és az $\mathbf{x}_Q$ és $\mathbf{y}_Q$ közötti kanonikus korreláció között az alábbi összefüggés létezik:

$$d_{(Q)} = 1 + \rho_{(Q)}, \tag{17.89}$$

valamint, hogy a főkomponens (FK) súlyok $w_{k(FK)}$ és a kanonikus súlyok (KK) $w_{k(KK)}$ és $v_{\ell(KK)}$ úgy viszonylanak egymáshoz, hogy a kanonikus változók értékeit közvetlenül a főkomponens súlyokból számíthatjuk:

$$\begin{aligned} \xi_{(KK)} &= f \sum_k w_{k(FK)} x_{x(Q)}, \\ \eta &= g \sum_\ell v_{\ell(FK)} y_{\ell(Q)}, \end{aligned} \tag{17.90}$$

ahol f és g a latens változók standardizálásához szükségesek.

17.6.3A kategóriák skálázása

Egy kategorikus változót – láttuk az előzőekben – helyettesíthetünk Boole-változókkal úgy, hogy minden kategóriához hozzárendelünk egy $(0, 1)$ értéket felvevő változót. A k-adik Boole-változó akkor veszi fel az 1-et $(y_k = 1)$, ha a kérdéses megfigyelés (i) az y változó k-adik kategóriába esik $(y_i = k)$. Ebből következően felhasználva az előző két alfejezet latens változó-képzését:

$$\sum_k w_k y_{ki} = w_k, \tag{17.91}$$

vagyis a latens változó értéke az i-edik megfigyelésnél (ha a k-adik kategóriába tartozik) egyenlő lesz a w_k súlyegyütthatóval, ami pedig a k-adik kategóriához rendelt Boole-változók súlya. Így a latens változó a megfigyelt változó intervallummérési szintű változójának tekinthető.

Több, mint két kategorikus változó esetén a kategóriák skálázásához a szuperkontingenciamátrixot $(\mathbf{M})$ normalizáljuk $(\mathbf{Q})$, meghatározzuk a főkomponens súlyokat

$(x_{k(Q)})$ és ezután a kapott súlyokat visszatranszformáljuk:

$$w_{k(M)} = w_{k(Q)}/\sqrt{m_k}. \tag{17.92}$$

A **Q** mátrix első komponense triviális, a főkomponens együtthatói egyenlőek m_k-val, a súlyok pedig 1-gyel. A második (és a további főkomponensek) pedig egyenlőek lesznek a **D** (normált eltérés) mátrix első (és többi) főkomponenseivel, így a gyakorlati megoldásoknál a **D** mátrixot használjuk.

17.7. Háromdimenziós útelemzés

Az eddigi modellek kétdimenziós mátrixok elemzését végezték. A háromdimenziós általánosításukhoz nézzük először az ún. Kronecker-szorzatot, amit mátrixok között értelmezünk.

Az **A** mátrix a **B** és **C** mátrix Kronecker-szorzata:

$$\underset{(JK \times PQ)}{\mathbf{A}} = \underset{(J \times P)}{\mathbf{B}} \otimes \underset{(K \times Q)}{\mathbf{C}} = [a_{(jk)(pq)}] = [b_{jp}c_{kq}]. \tag{17.93}$$

Bevezetünk még két jelölést:

– a vektorfüggvényt, jele: vec **A**,

$$\text{vec}\mathbf{A} = (a_{11}, a_{12}, \ldots, a_{1K}, a_{21}, \ldots, a_{jk}, \ldots, a_{JK})' = a_{(jk)}.$$

A vec**A** függvény az **A** mátrix sorait sorban belehelyezi egy oszlopvektorba. A zárójelben lévő dupla index kombinált indexet jelöl: $(jk) = (j-1)J+k$, ahol a második index (k) gyorsabban fut, mint az első index (j).

– egy paramétermátrix (**A**) modell (pattern) mátrixát, **M** -et vagy más jelöléssel $\mathbf{M}_\mathbf{A}$, $M(\mathbf{A})$ vagy pattern (**A**). Az **M** mátrix azt jelöli, hogy a paramétermátrix mely eleme a szabad paraméter, és melyet rögzítettünk 0-hoz:

$$\mathbf{A} = \mathbf{M}_\mathbf{A}\mathbf{A}.$$

Ha $m_{jk} = 1$, akkor bármely valós a_{jk}-ra $(m_{jk}a_{jk} = a_{jk})$. Ha $m_{jk} = 0$, akkor az $a_{jk} = 0$ kötött elem. Ha a modell mátrix egységmátrix, akkor a paramétermátrix diagonális:

$$\mathbf{A} = \mathbf{I}\mathbf{A}.$$

Az indexek az eddigiektől eltérnek, így összefoglaljuk őket:

$$\begin{aligned} i &= 1, 2, \ldots, I, \\ j &= 1, 2, \ldots, J, \\ k &= 1, 2, \ldots, K, \\ r &= 1, 2, \ldots, R, \\ q &= 1, 2, \ldots, Q. \end{aligned}$$

A Kronecker-főkomponensek

Ha egy **A** mátrixnak Kronecker-struktúrája van és elég nagy mátrix, akkor jól jellemezhetjük a viszonylag kicsi **B** és **C** paramétermátrixokkal. A modellben

$$\mathbf{A} = \mathbf{B} \otimes \mathbf{C} + \mathbf{A}_r, \tag{17.94}$$

ahol **B** és **C** mátrixokat az **A** mátrix Kronecker-főkomponenseinek nevezzük. A modell becslését a legkisebb négyzetek módszerének kritériuma szerint a következő függvény minimalizálásával kapjuk meg:

$$\Phi = \text{trace}(\mathbf{A}_r'\mathbf{A}_r) \Longrightarrow \min . \tag{17.95}$$

Ha az **A** mátrix szimmetrikus (multitrait-multimethod) korrelációmátrix, akkor a reziduális $\mathbf{A}_r$ mátrix diagonális elemeit elhagyhatjuk a minimalizáláskor. Ekkor a reziduálismátrix:

$$\mathbf{A}_s = \mathbf{A}_r(\mathbf{1} - \mathbf{I}), \tag{17.96}$$

és a **B** és **C** mátrixokat Kronecker–Minres faktornak nevezzük, amelyek a

$$\Phi = \text{trace}(\mathbf{A}_s'\mathbf{A}_s) \longrightarrow \min \tag{17.97}$$

függvényt minimalizálják.

A Kronecker-főkomponensek illeszkedésének jóságát a következő index mutatja:

$$fitkpc(\mathbf{A}) = 1 - \text{trace}(\mathbf{A}_r'\mathbf{A}_r)/\text{trace}(\mathbf{A}'\mathbf{A}). \tag{17.98}$$

Háromdimenziós modellek

A következő táblázatban Lohmöller és Wold nyomán összefoglalunk különböző kétdimenziós modelleket és a háromdimenziós megfelelőiket.

Modell	Kétdimenziós	Háromdimenziós
Komponens-elemzés	$\underset{(J\times I)}{\mathbf{Y}} = \underset{(J\times P)}{\mathbf{A}}\,\underset{(P\times P)}{\mathbf{D}}\,\underset{(P\times I)}{\mathbf{X}} + \underset{(J\times I)}{\mathbf{E}}$	$\underset{(JK\times I)}{\mathbf{J}} = (\mathbf{A}_1 \otimes \mathbf{A}_2)\,\underset{(PQ\times M)}{\mathbf{D}}\,\underset{(M\times I)}{\mathbf{X}} + \underset{(JK\times I)}{\mathbf{E}}$ Háromdimenziós komponens-modell Tucker, 1966
Faktorelemzés	$\underset{(J\times 1)}{\mathbf{y}} = \underset{(J\times P)}{\mathbf{A}}\,\underset{(P\times 1)}{\eta} + \underset{(J\times 1)}{\epsilon}$	$\underset{(K\times I)}{\mathbf{y}} = (\mathbf{A}_1 \otimes \mathbf{A}_2)\,\underset{(PQ\times 1)}{\eta} + \underset{(JK\times 1)}{\epsilon}$ Háromdimenziós faktormodell, Bloxom (1968), Bentler and Lee (1978)
Útmodell manifeszt változókkal (MVPM)	$\underset{(J\times 1)}{\mathbf{y}} = \underset{(J\times J)}{\mathbf{B}}\,\underset{(J\times 1)}{\mathbf{y}} + \underset{(J\times 1)}{\epsilon}$	$\underset{(JK\times 1)}{\mathbf{y}} = (\mathbf{B}\otimes\mathbf{B}_2)\,\underset{(JK\times JK)}{\mathbf{y}} + \underset{(JK\times 1)}{\epsilon}$ Háromdimenziós útmodell (MVP3M) Lohmöller (1983)
Útmodell latens változókkal (LVPM)	$\underset{(J\times 1)}{\mathbf{y}} = \underset{(J\times P)}{\mathbf{A}}\,\underset{(P\times 1)}{\eta} + \underset{(J\times 1)}{\epsilon}$ $\underset{(P\times 1)}{\eta} = \underset{(P\times P)}{\mathbf{B}}\,\underset{(P\times P)}{\mathbf{B}}\,\underset{(P\times 1)}{\eta} + \underset{(P\times 1)}{\delta}$	$\underset{(JK\times 1)}{\mathbf{y}} = (\mathbf{A}_1\otimes\mathbf{A}_2)\,\underset{(JK\times PQ)}{\eta} + \underset{(JK\times 1)}{\epsilon}$ $\underset{(PQ\times 1)}{\eta} = (\mathbf{B}_1\otimes\mathbf{B}_2)\,\underset{(PQ\times PQ)}{\eta} + \underset{(PQ\times 1)}{\delta}$ LVP3M, Lohmöller (1983)
Komponens-elemzés kevert mérési skálával	$\underset{(K\times I)}{\mathbf{Z}} = \text{scaled}\,\underset{(J\times I)}{(\mathbf{Y})}\,\underset{(J\times I)}{(\mathbf{F})}$ $\underset{(J\times I)}{\mathbf{Y}} = \underset{(J\times P)}{\mathbf{A}}\,\underset{(P\times P)}{\mathbf{D}}\,\underset{(P\times I)}{\mathbf{X}} + \underset{(J\times I)}{\mathbf{E}}$ PRINCIPALS (Young, Takane and de Leeuw 1978)	$\underset{(JK\times I)}{\mathbf{Z}} = \text{scaled}\,\underset{(JK\times I)}{(\mathbf{Y})}\,\underset{(JK\times I)}{\mathbf{F}}$ $\underset{(JK\times I)}{\mathbf{Y}} = (\mathbf{A}_1\otimes\mathbf{A}_2)\,\underset{(JK\times RR)}{\mathbf{D}}\,\underset{(R\times I)}{\mathbf{X}} + \underset{(JK\times I)}{\mathbf{E}}$ ALSCOM3 (Sands and Young 1980)

A táblázatban szereplő jelölések:

A: faktorsúlymátrix, $(J \times P)$ jelöli a mátrix méretét, vagyis az **A** mátrixnak J sora és P oszlopa van, $(J \times P)$ típusú mátrix,

B: az útegyüttható-mátrixa,

E, **F**, ϵ, δ: a reziduálisokat tartalmazzák,

Z, **Y**: adatmátrixok,

X: főkomponensek értékei,

y: manifeszt változók vektora,

η: latens változók vektora,

D: sajátértékek mátrixa.

17.8. Példa latens változók útelemzésére

Társadalmi státus – értékrend – magatartás út-modellek

A latens változókat a megfigyelt (manifeszt) változók mérési modelljeiként írjuk le. A manifeszt változók három szintjét különítjük el. Az első szinten az egyén családi hátterét írjuk le: a Család társadalmi státusát a Nagyapa, Apa és Anya iskolai végzettségével mérve, a Gyermekkort a különböző településkategóriákban az első 14 életévben eltöltött idővel, Gyermekkori boldogsággal, Jóléttel és Zaklatottsággal kifejezve. A második szint a jelen jellemzését adja: a Társadalmi státus latens változót a Nem, Lakóhely, Életkor, Iskolai végzettség, Jövedelem megfigyelt változók fejezik ki, a Vallásosságot pedig a „Vallásos szellemben nevelték-e Önt a szülei?" és a „Vallásos embernek tartja-e magát?" kérdésekkel mértük. A Család társadalmi státusa, a Gyerekkor, a Társadalmi státus és a Vallásosság latens változók és az őket meghatározó megfigyelt változók kapcsolatát a modellben belsőnek (inwards) tekintjük. Ez alatt azt értjük, hogy a manifeszt változók, mint az ismeretlen dimenzióról összegyűjtött megfigyelt változók, a hatásukat kifejező súlyegyütthatókkal előállítják a latens változót anélkül, hogy azt gondolnánk, hogy a latens változó a felelős a megfigyelt változók varianciáiért. A modellben ezt úgy jelöljük, hogy a megfigyelt változókból húzzuk a nyilat a latens változó felé. A manifeszt változók harmadik szintje az endogén változók halmaza. Egy vagy két blokkot különítünk el. Az egyik – a modellekben állandó – az értékrendszert kifejező blokk. Megfigyelt változói valójában már önmagukban is képzett változók, a MINISSA tér három dimenziója (az első kettő 45 fokkal elforgatva). A modellek abban különböznek, hogy az értékrendszer mellé milyen output változókat jelölünk második blokként. Jellegében ezek az emberek magatartásait, vélekedéseit és értékeléseit fejezik ki. A két endogén blokkban a latens és megfigyelt változók kapcsolata külső (outwards) jellegű, feltételünk szerint a latens változó, mint egy közös faktor, dimenzió felelős a manifeszt változók variancia-kovarianciáiért. Jelölésben a latens változóból a megfigyelt változóba mutató nyillal jelezzük ezt.

17.8.1. A modell latens változói

A Család társadalmi státusa

A Nagyapa, Nagyanya, Apa és Anya iskolai végzettsége közül a legnagyobb súlya (és ez átlagosan kétszeres az őt követő legfontosabb súlynak) az Apa iskolai végzettségének van. (Ezek a súlyok a különböző modellekben 0,53-tól 0,68-ig terjednek.) Másodsorban határozza meg a Család társadalmi státusát az Anya iskolai végzettsége. A súlyok a modellben 0,33-tól 0,38-ig terjednek. A nagyszülők közül a Nagymama iskolai végzettsége játszik nagyobb szerepet, a súlyok 0,19 és 0,26 között mozognak.

(A Bizalom endogén változónál ez a súly megugrik és a 0,36 értéket veszi fel.) A Nagyapa iskolai végzettségének kicsi az együtthatója, de ez az együttható negatív, ami a társadalom nagymértékű átrétegződésére utal. Ha generációnként is megvizsgáljuk a Család társadalmi státus latens változót, azt találjuk, hogy a Nagyapa iskolai végzettsége a 40–59 éves generációnál jelenik meg negatív együtthatóval. És ez a súly nagyobb, mint az Anya iskolai végzettségének a súlya. Ugyanennél a generációnál a Család társadalmi státus latens változót legnagyobb mértékben a Nagyanya iskolai végzettsége határozza meg.

A 60 éves és idősebbeknél a Nagyanya iskolai végzettségének van negatív súlya, és ennél a generációnál az Apa iskolai végzettségének a szerepe a meghatározó. (0,81-től 0,96-ig terjedő együtthatókkal.)

A 20–39 éves generációnál átlagosan kisebbek a faktorsúlyok, kiegyenlítettebb az Apa, Anya és Nagyanya együtthatójának értéke, és nincs köztük negatív súly.

Gyermekkor

A Gyermekkor latens változó részben a gyermekkorban (első 14 év) különböző településeken eltöltött idővel, részben kilenc fokú létrán mért boldogsággal és jóléttel, és ötfokú skálán mért zaklatottsággal függ össze. A Gyermekkor latens változó pozitívan kapcsolódik Budapesttel és a Nagyvárosban eltöltött gyermekkori idővel, negatívan a Falun és Tanyán töltött gyermekkorral, pozitívan a Boldogsággal, negatívan (kicsi súllyal) a Zaklatottsággal, és átlagosan kicsi a súlya a Jólétnek.

A modellek Gyermekkori latens változója az őt meghatározó manifeszt változók alapján a boldogabb, urbanizált településen eltöltött gyermekkori évek kontra a falun, tanyán töltött boldogtalan évek dichotómiát fejezi ki.

A különböző generációk közötti alapvető különbség, hogy a fiatal 20–39 éves generációnál lecsökken a Tanya és a Falu települések súlya, és ezzel párhuzamosan megnő a városi kategóriák együtthatója (Kisváros, Nagyváros, Budapest), az idősebb generációknál ez éppen ellenkezőleg változik, a falu-város dichotómiában a falusi kategóriáknak van nagyobb szerepe. Még egy; nem jelentős súllyal, de a 20–39 évesek Zaklatottság emlékezete kap a negatív irányba nagyobb együtthatót (−0,10).

A gyermekkori Jólét csak a Bizalom, Egyenlőtlenség, Pozíció és a Tudat endogén blokkok esetén játszik nagyobb szerepet a Gyermekkor latens változó meghatározásában (0,21 és 0,28 faktorsúlyokkal).

Társadalmi státus

A Társadalmi státus latens változót öt manifeszt változóval fejeztük ki: Nem, Lakóhely, Életkor, Iskolában töltött évek, Jövedelem (egy főre jutó jövedelem). Ezek közül

alapvetően az Iskola (0,56) és a Lakóhely (0,46) határozza meg a társadalmi státus értékét. A modellekben (átlagosan) egységnyi emelkedés a társadalmi hierarchiában a latens változó értékelése szerint elérhető egységnyivel urbanizáltabb településekre költözéssel és egységnyivel több iskolával.

Az életkor, mint ezt láttuk már korábban is, a társadalmi hierarchiában eltöltött hellyel párhuzamosan mozog a különböző modellekben. A Társadalmi státus latens változót −0,26-tól −0,38-ig terjedő súllyal határozza meg.

A Jövedelemnek kicsi a szerepe a társadalmi státus emelkedésében (0,07) az is inkább az idősebb generációknál jellemzőbb (0,10-től 0,27-ig).

A Társadalmi státus emelkedése – gyenge összefüggésben (0,05) – inkább a férfiakhoz kapcsolódik.

Vallásosság

A vallás latens változót legnagyobbrészt (0,80 feletti súlyokkal) az ateista vagyoktól a rendszeresen járok templomba kategóriákat tartalmazó kérdésekre adott válaszok határozzák meg. Kisebb súlya van a Vallásos szellemű nevelésnek (0,25-től 0,44-ig terjedő együtthatókkal).

17.8.2A modellek endogén változói

Az alapmodell: Mi határozza meg az emberek értékrendjét?

Mit is nevezünk értékrendnek? A Rokeach-értékek axiológiai tere (MINISSA-tér) három dimenziója (az első kettő 45 fokkal elforgatva) a Jólét–Eszme, Közösség–Autonómia és az Indusztriális-Posztindusztriális dichotómiával jellemezhető. Ezek közös, a társadalmi helyzettel magyarázható dimenziója a Jólét, Közösség, Indusztriális értékektől az Eszme, Autonómia, Posztindusztriális értékekig terjed. Ezen belül is legnagyobb súlya a Közösség–Autonómia tengelynek van, amíg az Indusztriális–Posztindusztriális tengely szerepe csekély. Azt mondhatjuk tehát, hogy az Értékrend latens változó negatív oldalán a hagyományos értékek választását, pozitív oldalán pedig a modernizáció értékeinek választását méri. A modellben meghatározott Értékrend latens változót a Család társadalmi státusa, Gyermekkor, Társadalmi státus, Vallás latens változók 26%-ban (R^2) tudják meghatározni. Ezen belül a Társadalmi státusnak a direkt hatása 0,40 (útegyüttható), teljes hatása a Vallásosságon keresztül 0,50, vagyis azt mondhatjuk, hogy egységnyi emelkedés a társadalmi státuson (amit elérhetünk például úgy, hogy a településrendszerben egységnyivel urbánusabb településre költözünk és eggyel növeljük iskolai végzettségünket) értékrendünk modernizációjában 0,50 egységnyi fejlődést idéz elő. A Család társadalmi státusának direkt hatása csekély, de teljes hatásban az iskolázottabb családi háttérkörnyezet 0,27 egységgel modernizáltabbá teszi értékrendünket. (Megjegyezzük, hogy a Család társadalmi státusában legnagyobb súlya az apa iskolai végzettségének van, kisebb, de jelentős szerepe van az anya iskolai végzettségének is, a nagyanya iskoláinak nagyobb a jelentősége, mint a nagyapáé). Általában, a modell szerint a családi és gyermekkori háttérnek a közvetlen hatása az emberek értékválasztásaira csekély. Azonban azzal, hogy az emberek társadalmi státusát jelentős súllyal meghatározzák, ezen keresztül azt mondhatjuk, hogy iskolázottabb szülők és nagyszülők, több gyermekkorban jólétben eltöltött boldog (kevésbé zaklatott) városi év az embereket a hagyományos értékek felől az autonóm, modern értékek választása felé segíti.

Modernizáció

Az előző modell értékdimenzióit kicseréltük három, az értékekből választott dichotómiával: Érzelem – Pragmatikus (Megbocsátó, Szeretettel teljes – Alkotó szellemű, Hatékony), Érzelem – Racionalitás (Megbocsátó, Szeretettel teljes – Értelmes, Logikus gondolkodású), Hagyományos – Önálló (Engedelmes, Segítőkész, Tiszta, Udvarias – Bátor, Önálló).

A modell latens változóját nevezhetjük a Modernizációnak, mivel az Érzelem, Hagyományos értékektől a Pragmatikus, Racionális és Önálló értékekig méri az értékválasztást. Ez a modernizációs latens változó hasonlóan működik, mint az előbbi modellben az Értékrend latens változó, amelyik szintén egy modernizációs tengelyként értelmezhető.

A társadalmi státuson felfelé haladva csökken a vallásosság (−0,50), amely pedig a Modernizáció ellen hat (−0,40), így a társadalmi státus direkt hatása (0,32) a Vallás latens változón keresztül még felerősödik, a teljes hatás így 0,43. A Család és Gyermekkori háttér ebben a modellben is a Társadalmi státuson keresztül hat a Modernizációra (a totális együtthatók: 0,18 és 0,16).

Anómia

A modellben bevezettük az Értékrend mellé endogén változónak az Anómiát. Bár a modell tanúsága szerint az anomikusság nem függ túlságosan a társadalmi státustól és a háttértől, még kevésbé az értékrendtől, a latens változók teljes hatása nem elhanyagolható:

Család társadalmi státusa:	−0,19
Gyermekkor:	−0,08
Társadalmi státus:	−0,27

Vagyis a társadalmi hierarchián felfelé haladva kevésbé anomikusak az emberek Magyarországon. Ahhoz mondjuk, hogy valaki egységnyivel csökkentse anómiáját, nem kell mást tennie, mint hogy például elköltözik faluról nagyvárosba, és képezi magát, mondjuk a szakma után elvégzi a középiskolát és az egyetemet. De ha a társadalmi hierarchiában az előbb említett mozgást megteszi és úgy dönt, anomikus marad, célszerű, ha nem enged a szekularizációnak és megmarad a hagyományos értékrendnél.

Három generációra külön-külön is megbecsültük a modell paramétereit. A 20–39 éves generációnál nagyon lecsökken a modell státus direkt hatása (−0,09), és a teljes hatás is csak akkora, mint a teljes mintában a közvetlen hatás (−0,18). Megnőtt viszont az értékek fontossága. Eszerint egységnyi lépés a hagyományos értékek felé, és 0,16 egységgel nő az anómia. Nő az anómia szintje az alacsony családi háttérrel. Nő az anómia a gyermekkorban eltöltött városi élettel is, bár ez a gyermekkor indirekt hatása közvetítőkön keresztül nézve lecsökken, de nem fordul át (míg a többi generációnál és a teljes mintában is ez a gyengébb közvetlen hatás miatt átfordul).

A 40–59 éves generációnál az anómia varianciájának mindössze 6%-át tudjuk az öt latens változóval megmagyarázni. Egyedül a társadalmi státus latens változó közvetlen hatása jelentős (−0,21), vagyis a társadalmi hierarchián egységgel lejjebb menve az anómia 0,21 egységgel nő. Az Értékrend választása ennél a rétegnél nem kapcsolódik szisztematikusan az anómiához.

Társadalmi célok (Inglehart)

A 12 Inglehart-célkitűzés közül a modell latens változója azokat állítja elő pozitív súllyal, amelyeket Inglehart posztmateriálisnak nevezett. Társadalmi célok latens változó negatív együtthatóval fejezi ki a materiális célokat. Ezt a szabályt két cél megszegi, „Az ország gazdasági egyensúlyát biztosítani" és a „Városainkat, falvainkat és tájainkat szebbé tenni" célok, amelyek megcserélődnek, a „gazdasági egyensúly…." a posztmateriális célokhoz kerül, a „városaink szebbé tevése…." pedig materiális céllá válik. A társadalmi célok latens változót azért nevezhetjük posztmateriális – materiális dimenziónak (azért nem fordítva, mert a manifeszt változók mérésekor rangsorolást alkalmaztunk).

Megjegyezzük, hogy ez a latens változó nagyon hasonlít a MINISSA első tengelyére, vagyis jól értelmezhető a modell többi blokkjától függetlenül is mint a célok immanens dimenziója.

A Posztmateriális–Materiális latens változót a családi és gyermekkori háttér nem befolyásolja. A Társadalmi státus közvetlen hatása $-0{,}28$, teljes hatása (az értékrenden és a valláson keresztül) $-0{,}40$, vagyis a társadalmi hierarchián egységgel lejjebb a célok választása a materiális irányban mozog 0,40 egységgel, a társadalmi hierarchián feljebb jutva a célváltás a posztmateriális irányban mozog ugyanilyen mértékben. Ha külön blokkban szerepeltetjük a Materiális és Posztmateriális célokat, az derül ki, hogy a Materiális célok választása a társadalmi hierarchiával ellentétes irányba változik, ennek a kapcsolatnak az együtthatója 0,22, azonban a Posztmateriális célok választását csak gyengén befolyásolja a társadalmi státus.

Generációnként ez a kép különbözik. Míg a 20–39 éveseknél a Materiális dimenzió háttérbe kerül (0,05), fontosabbá válik a Posztmateriális cél ($-0{,}13$).

A 40–59 éveseknél ez megfordul, a társadalmi státus nagyobb súllyal befolyásolja a materiális célokat (0,28).

A 60 éves és ennél idősebbeknél szintén a Posztmateriális célok választása válik fontosabbá a társadalmi státus emelkedésével (0,23).

Bizalom

A modellben szereplő társadalmi pozíciót kifejező blokkok egyáltalán nem, vagy csak gyengén befolyásolják a Bizalom latens változót, de a modernebb értékrend is csak kis súllyal esik latba (0,04) abban, hogy valaki megbízik-e az emberekben vagy sem. Mégis, a magasabb Családi társadalmi státus háttérként inkább az önmagukban való bízást és nem általában az emberekben való bizalmat erősíti ($-0{,}06$). Ez az elért társadalmi státus hatásával ellensúlyozódik, és átfordul egy nagyon gyenge, vagy semmilyen másokban való bízásba. A Bizalmat, ha gyengén is, de erősíti a társadalmi hierarchiában elfoglalt magasabb pozíció (0,10).

Mit is tartalmaz a Bizalom latens változó?

Az egyik oldalán az „emberekben általában meg lehet bízni" kijelentésnek van nagy súlya és kisebb az „emberek többsége szívesen segít másokon" kijelentés fontossága, a dimenzió másik oldalát az „emberek többsége csak csak önmagával törődik" és „ha törésre kerül a sor, nem sok jót lehet várni az emberektől" kijelentésekkel való egyetértés határozza meg. Az értékrend modernizációja alig, csak nagyon-nagyon gyengén (0,04) növeli a Bizalmat.

Más képet kapunk azonban generációnként. A 20–39 éveseknél a Társadalmi státusszal a *Bizalmatlanság* növekszik ($-0{,}10$), a 40–59 éveseknél viszont már a magasabb társadalmi státusszal megjön a Bizalom is (0,05), és a 60 évesek vagy idősebbeknél ez egy kicsit még fokozódik is (0,07).

Megjegyzendő, hogy ezek a súlyok meglehetősen kicsik, így a kapcsolódás csak nagyon gyengének mondható.

Az Értékrend latens változó eltérő tartalmat kap generációnként.

A 20–39 éveseknél megnő az Indusztriális – posztindusztriális tengely fontossága, és ellentétes előjelű lesz a Közösség – Autonómia és Jólét – Eszme tengellyel. Ennél a rétegnél az Eszme, Autonómia értékek együttjárása az Indusztriális értékekkel növeli legjobban a Bizalmat.

A Jólét, Közösség és Posztindusztriális értékek együttes választása növeli legjobban viszont a Bizalmatlanságot, vagyis a csak magunkban bízást. Megjegyzendő, hogy az a furcsa modernizáció ennél a generációnál is csak itt, a Bizalom output változóblokk esetén jelent meg.

Participáció, beleszólás

A részvételt és beleszólást a szűkebb és tágabb környezet döntéseinek meghozatalába a társadalmi háttérrel és státusszal, vallással és értékrenddel a következőképpen tudjuk magyarázni:

	Participáció		
	Direkt hatás	Teljes hatás	Indirekt hatás
Család társadalmi státusa	−0,01	0,13	0,14
Gyermekkor	−0,03	0,11	0,14
Társadalmi státus	0,17	0,32	0,15
Vallás	−0,15	−0,18	−0,03
Értékrend	0,16	0,16	–

Összesen a fenti tényezők a variancia 13%-át tudják reprodukálni. Látható, hogy a családi és gyermekkori háttér közvetlen hatása elhanyagolható, a jelenlegi társadalmi státusra gyakorolt jelentős hatásukkal azonban a súlyuk felemelkedik 0,13, illetve 0,11-ra.

A társadalmi státus egységnyi emelkedésével a Participáció 0,17 egységnyivel emelkedik. De ha azt is figyelembe vesszük, hogy a Vallásosság 0,50 egységnyivel csökken, valamint azt, hogy az Értékrend 0,41 egységnyivel lesz modernebb, és ezekkel együtt azt, hogy a szekularizácó 0,15 egységgel, a Modernizáció pedig 0,16 egységgel növeli a Participációt, a Társadalmi státus egységnyi emelkedése összességében −0,32 egységnyivel növeli az emberek beleszólásait a döntésekbe.

A Társadalmi státusnak a súlya a Participációban a különböző generációknál:

	Útegyüttthatók		
	Direkt hatás	Teljes hatás	Indirekt hatás
20–39 évesek	0,19	0,29	0,10
40–59 évesek	0,22	0,37	0,15
60 évnél idősebb	0,48	0,48	–

Teljesítménymotiváció

A Teljesítménymotiváció latens változó, amely tartalmilag a dolgos életet jelenti és a sikeresen befejezett munkával függ össze nagy súllyal, kisebb mértékben a kitartással, a modell latens változói csak nagyon kevéssé magyarázzák, mindössze a variancia 20%-át.

Az Értékrendben történő elmozdulás az Eszme, Autonómia és Posztindusztriális értékek felé növeli 0,13 egységgel a Teljesítménymotiváltságot. A többi útegyüttható

közül a Társadalmi státusé kicsi, de pozitív 0,05, a Vallásosság is pozitívan hat (0,90), a teljesítménymotivációra.

Tehát óvatosan az mondható, hogy a magasabb Társadalmi státus elősegíti a modernebb Értékrend kialakulását, amely pozitívan hat a Teljesítménymotivációra.

Egyenlőtlenség

Az Egyenlőtlenség latens változó 12, ma még meglévő társadalmi egyenlőtlenség megszűnését, illetve csökkenését fejezi ki.

A társadalmi változókkal tulajdonképpen magyarázni nem tudjuk, egyedül az Értékrend modernizáció súlya nagyobb (0,14), de ezzel együtt is csak 0,03-ra emelkedik fel a többszörös korreláció négyzete.

Nagyon halovány kapcsolódást találunk még a Társadalmi státusnál és a Vallásnál. Eszerint a státus emelkedésével az egyenlőtlenség növekedését látják az emberek. A Vallásosság is az egyenlőtlenség növekedése érzetét kelti. És ez az Értékrenden keresztül már összességében 0,10 útegyütthatót jelent. A magasabb státusszal kapcsolódó modernebb értékrend a társadalmi státus hatását pozitívra változtatja, és teljes hatásában a társadalmi hierarchia 0,09 súllyal az egyenlőtlenség csökkenése irányában alakítja a véleményeket.

Pozíció

A Pozíciót, vagyis azt, hogy az emberek hol képzelik el a maguk életét az elképzelhető legjobb élet és az elképzelhető legrosszabb élet között, a modellbe bevont változóblokkokkal magyarázni nem tudjuk (0,01 a többszörös korreláció).

Sérelem

A 11 igazságtalanságot, méltánytalanságot kifejező Sérelem latens változó érzékenyen reagál az Értékrend változására (0,33).

Vagyis egységnyivel modernebb Autonóm, Eszme (kisebb súllyal Posztindusztriális) értékrendszerrel 0,33 egységnyivel több esetet élünk meg igazságtalanságként, méltánytalanságként. A társadalmi státus emelkedésével viszont csökken a méltánytalanságok súlya 0,16 egységgel, amit viszont a státusnak az értékrendre gyakorolt hatása semlegesít.

A boldogabb, városiasabb környezetben eltöltött gyermekkor is elősegíti, hogy bizonyos eseteket súlyos igazságtalanságként éljünk meg (a súlya 0,09). A Vallásosság viszont, ha nem is nagyon, de segít abban, hogy ne érezzünk nagyon súlyosnak igazságtalanságokat.

Tolerancia

A modellben a Társadalmi státus 0,21-es együtthatóval befolyásolja az emberek toleranciáját. Másként fogalmazva, alacsonyabb iskolai végzettséggel, falusiasabb településen élve és öregedve az emberek intoleranciája 0,21 egységgel fokozódik. Kedvezően hat az emberek toleranciájára az iskolázottabb családi környezet (0,11) és a modernebb értékrend is (0,06). Az Értékrend közvetítésével a Társadalmi státus direkt hatása még növekszik, és totális hatásként a modellben egységnyi státuszemelkedés a Toleranciát 0,30 egységgel növeli.

Tudat

Több, különböző magatartást és percepciót kifejező kérdéscsoport átlagos válaszértékeiből állítottuk össze a Tudat elnevezésű endogén változóblokkot. A Tudat latens változó értelmezéséhez azt vesszük sorra, hogy milyen súllyal tudtuk vele előállítani a manifeszt változók átlagos értékeit. A Tudat latens változó csökkenő mértékű *pozitív súllyal* kapcsolódik: a Participációhoz a Toleranciához, az Egyenlőtlenségek csökkenéséhez, Posztmateriális célok, Vállalkozási hajlam, Teljesítménymotiváció, Bizalom, Fogyasztói elvek, csökkenő fontosságú *negatív súllyal* határozza meg: az Anómiát, a Sérelmet, az Elégedettséget. A Tudat latens változó a társadalomhoz pozitív viszonyulását, a társadalmi változásokhoz való adaptációs készséget fejezi ki.

A társadalmi státusnak ehhez a pozitív Tudatra gyakorolt hatása a modell útegyütthatója szerint elég erős (0,32). Ha a Társadalmi státust meghatározó együtthatókat is figyelembe vesszük, és a státusnak az Értékrenden és Valláson keresztül kifejtett indirekt hatását is, akkor a Tudat egységnyi változtatását a szocializáció irányába elérhetjük úgy, hogy megemeljük az iskolázottságot két egységgel, két egységnyivel urbanizáltabbá tesszük a településeket és egységnyivel növeljük a jövedelmeket, de egységnyi urbanizációs költséget spórolhatunk, ha egységnyivel „fiatalabbá" tesszük az embereket. A legdrágább út a tudati modernizációban, ha a személyes jövedelem emelésével próbálkozunk, mivel egységnyi iskola négy egységnyi jövedelemnek felel meg, egységnyi lakóhelyváltozás az urbanizáció irányában megfelel két és félszeres jövedelememelkedésnek. A Vallás mint visszahúzó erő szerepel a modellben, közvetlenül −0,18 együtthatójával, (teljes hatása a Tudatra −0,20). A társadalmi háttér két latens változójának közvetlen súlya elég kicsi, de a totális hatás már jelentősnek mondható. A Családi státus teljes hatása 0,28, a Gyermekkoré 0,14.

A Tudatot nagyon különbözően befolyásolja generációnként a modell többi változója. Az Értékrend modernizációja (a Posztindusztriális értékek is nagyobb súlyt kapva) legnagyobb mértékben a 20–39 éves generációnál hat (0,22 az útegyüttható). A 40–59 éveseknél ez a hatás 0,14, 60 évesnél öregebbeknél már csak 0,09. Ezzel szinte párhuzamosan nő a Társadalmi státus szerepe a Tudatra. Míg a fiatalabb generációnál a státus közvetlen hatása gyakorlatilag nulla (a teljes hatás is csak 0,16), a 40–59 éveseknél a társadalmi hierarchiában felfelé haladva 0,35 egységgel változik a Participáció, Egyenlőtlenség csökkenése, Teljesítménymotiváció, Bizalom, magasabb Pozíció értékelés, és Posztmateriális értékek irányában az emberek tudata. A legidősebbeknél a társadalmi státus közvetlen hatása 0,42.

Praktikusan a pozitív Tudatra a 20–39 éves generációnál leginkább az Értékrend modernizációjával lehet hatni, az idősebb generációnál pedig egyre inkább a Társadalmi státus növelésével.

17.8.3. Ökológiai blokk felvétele a modellbe

Az eddigi modellekben szereplő társadalmi háttér, társadalmi státus és értékrend, valamint más tudati, viselkedési blokkok mellé most az egyén ökológiai környezetét is felvonultatjuk magyarázó, exogén blokként. A területi adatokat az Akadémia Adatarchívuma Településsoros adatbázisából vettük. Ez az adatállomány az MTA I. Természettudományi Főosztálya és a Városépítési Tudományos és Tervező Intézet között 1980-ban kötött együttműködési megállapodás alapján jött létre, és a KSH különböző adatfelvételei – Népszámlálások, rendszeres településstatisztikai adatfelvétele – nyomtatásban nyilvánosságra hozott településsoros és gépi adathordozókra rögzített adatait tartalmazza. Ebből

az adatállományból 18 mutatót választottunk ki és rendeztünk saját adatbázisunkhoz, amelyek a település infrastruktúráját, foglalkozási szerkezetét, demográfiai összetételét jellemzik.

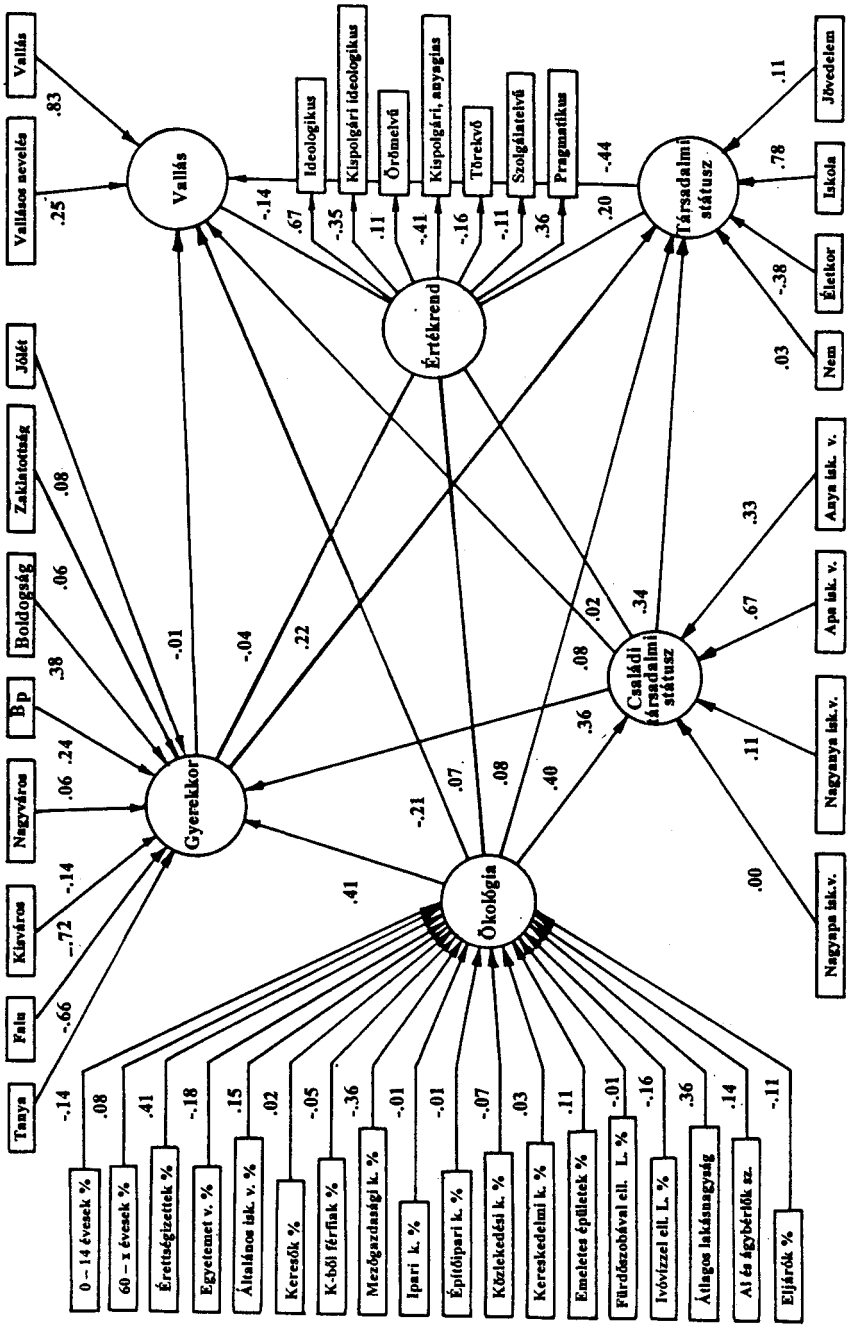
Az Ökológiai blokk latens változóját 18 mutató csökkenő fontosság (súly) szerinti sorrendben a következőképpen állította elő:

Pozitív súllyal	Negatív súllyal
Érettségizettek aránya	
	Mezőgazdasági aktív keresők aránya
Átlagos lakásnagyság	
	Egyetemet végzettek aránya
	Ivóvízzel ellátott lakások aránya
Általános iskolát végzettek aránya	
Al- és ágybérlők száma	
	0–14 évesek aránya
Emeletes épületek aránya	
Fürdőszobával ellátott lakások száma	
	Eljárók aránya

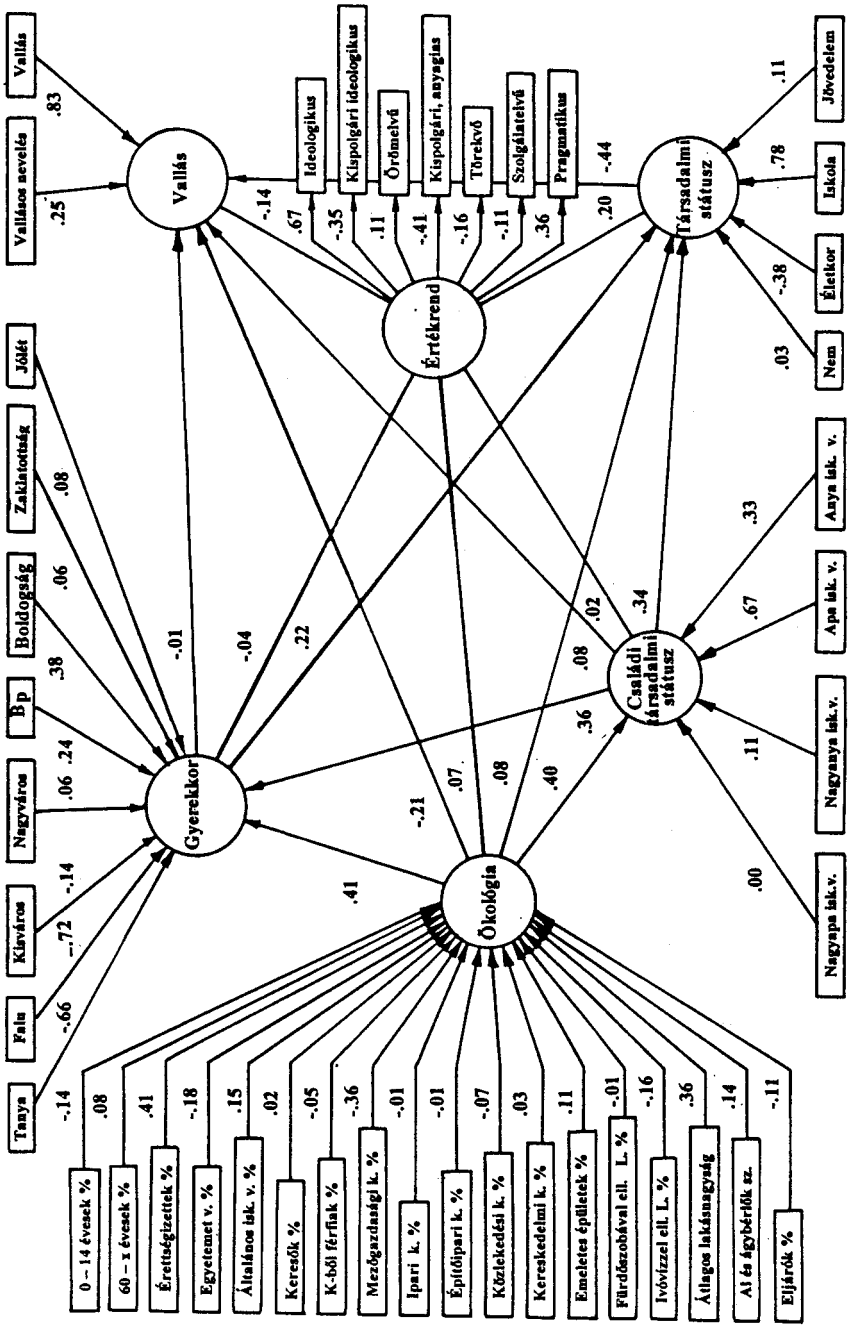
A többi mutató súlya már 0,1 alatt van.

Az Ökológia latens változó pozitív felét nagyobb súllyal a legalább középiskolát végzettek aránya, infrastruktúrában az urbanizáltabb kép (emeletes épületek aránya), fürdőszobával ellátott lakások aránya és a fejlettebb kereskedelmi hálózat határozza meg, míg a negatív oldalon a mezőgazdasági aktív kereső aránya a meghatározó, kisebb mértékű az eljárók aránya és ugyanerre az oldalra esik az egyetemet végzettek aránya és az ivóvízzel ellátott lakások aránya, vagyis a bal oldal önmagában is heterogén, legalábbis két dimenziót is magában foglal. Ez utalhat arra, hogy a modellben az ökológiát két latens változóval is képviseltethettük volna. Így mindenesetre azzal a dichotómiával jellemezzük az ökológia latens változót, hogy foglalkoztatásában a mezőgazdasági terület, amelyik a teljes foglalkoztatást nem tudja biztosítani (vagy ennek másik véglete, magasan kvalifikált népesség, magas gyermekszám és jó ivóvíz), mint lakóhelyi környezet terjed a középiskolai végzettségűek magas arányáig, az ezzel együttjáró városi kultúráig, a zsúfoltsággal a városi településekig, de ezzel a magas lakáshiányig is.

Ez az Ökológiai latens változó a modell egyetlen olyan exogén változója, amelyik egyben predeterminált is. Jelentős a direkt hatása az egyén családjának társadalmi státusára és a Gyermekkor latens változóra. Az Ökológia latens változó, amely az ember társadalmi környezetének viszonyát fejezi ki, a modellünk szerint közvetlenül az ember társadalmi hátterére hat leginkább (0,40 és 0,41 útegyütthatókkal), lényegesen kisebb a közvetlen hatása az egyén társadalmi státusára és értéktudatára (0,10 és 0,14 útegyütthatókkal). De ha nem feledkezünk meg az indirekt hatásokról sem, akkor már a modell szerint azt állíthatjuk, hogy egységnyi emelkedése az ökológiának a társadalmi státuson több, mint egyharmadot emel, és majdnem egyharmaddal segíti az értéktudat modernizálódását. Egy teljes egységnyit az értéktudaton úgy modernizálhatunk, hogy az Ökológia egységnyi emelése mellett a Társadalmi státuson másfél egységnyit emelünk (pl. két egységgel emeljük az iskolázottságot, vagy csak egy egységnyivel, de akkor még sok pénzzel). Ezt a kapcsolódást akkor kaptuk, amikor a modellben az értékrendet a Rokeach-értékteszttel mértük, méghozzá úgy, hogy az értékek három fő dimenziójával, a Jólét–Eszme Közösség–Autonómia, Indusztriális–Posztindusztriális dimenziókkal fejeztük ki az Értékrend latens változót. Amikor a Rokeach-tesztek három kiraga-



17.1. ábra. Társadalmi státus – értékrend útmodell



17.2. ábra. Ökológiai – társadalmi státus – értékrend útmodell

dott (de nem alapvető) dichotómiáját választottuk az értékrend modernizálódásának kifejezésére (Érzelem–Pragmatikus, Érzelem–Racionális, Hagyományos–Önálló), az Ökológia közvetlen és teljes hatása lényegesen kisebb volt (0,20 és 0,13).

Az *Ökológia* latens változó közvetlen hatása a tudat más területeire ellentétes a *Társadalmi státus* hatásával. Míg az *Ökológia* direkt hatásában anomikussá tesz, a *Társadalmi státus* megfordítja ezt a hatást, és teljes hatásában már gyengén ugyan, de csökken az Anómia (−0,70).

Amikor az Anómiához még hozzávettünk más tudati változókat, az Intoleranciát, Participációt, Teljesítménymotivációt, Vállalkozási hajlamot, Bizalmat, Fogyasztói elveket, Pozíciót, Igazságosságot, Elégedettséget és Egyenlőséget, és ezek fejezték ki a *Tudat* latens változót, az Ökológia pozitív (0,12-os) súlya az Anómiához és Intoleranciához kapcsolódik, és csak a Társadalmi státus hatása közvetítésével fordul át a teljes hatása a társadalomhoz való pozitív viszonyt kifejező tudati oldalhoz.

18. fejezet

Többszempontú modellek

A társadalomtudomány komplex fogalmainak kifejezéséhez nagyon gyakran a pszichológiában kidolgozott faktorelemzés módszerét használja. A faktorelemzés segítségével a mögöttes, közvetlenül nem megfigyelhető fogalmakat és összefüggéseket az empirikus, korrelált változókból származtatják. A korreláció matematikailag csupán a kovariancia egy mutatója (standardizált változata). Charles Spearman (1927), a faktorelemzés egyik kidolgozója volt az, aki más értelmezést adott a megfigyelt változók kapcsolódásainak. A korrelált változókat úgy tekintette, hogy azok „egy közös faktortól" függnek (Spearman, C., 1927). A későbbi irodalmakban azután beszéltek latens hatásokról, igazi („true") értékekről vagy forrásokról, jellemző vonásokról („traits"), vagy más terminológiát használtak annak kifejezésére, hogy a korreláció rendszerét egy rejtett, mögöttes ok magyarázza. A faktor a kapcsolatoknak egy magasabb szintű struktúráját jelenti, amely McDonald terminológiájában egy „absztraktív attribútum", amely lehet ok, de nem feltétlenül az, hanem lehet „csupán" mögöttes, latens változó.

A faktorelemzés hasznosnak bizonyult a társadalomtudományok széleskörű alkalmazásaiban a sokváltozós adathalmazok magasabb szintű kapcsolatainak felderítésében. A faktorelemzés mellett a metrikus sokdimenziós skálázás általános módszer a korrelációs együtthatók vagy hasonlósági (illetve különbözőségi) együtthatók struktúrájának kiértékelésére, amikor ezeket a megfigyelési egységeknek egy változó (vagy stimulus) halmazára adott válaszaiból számítjuk.

Abban az esetben, ha a megfigyelt adatok kétutas (kétdimenziós) mátrixát kiegészítjük egy újabb szemponttal, akkor ezek a módszerek már nem használhatók.

A háromutas (háromdimenziós) mátrixok kiértékelésének klasszikus módszere a Tucker-féle háromszempontú faktorelemzés (Tucker, L. R., 1963, 1964, 1966, 1972).

Tucker használta először a szempont (mode) kifejezést (Tucker, 1964, p. 112), ami alatt az indexeknek azt a halmazát értette, amely alapján az adatokat klasszifikálhatjuk. Például ha egy mintát veszünk egy populációból, és valamely változóhalmazt (teszthalmazt) mérünk, megfigyelünk a minta egyedeivel, akkor a megfigyelési egységekhez hozzárendelünk egy indexhalmazt, amely azonosítja, megkülönbözteti a megfigyelési egységeket, klasszifikálja azokat. Ez lesz az adathalmaz egyik szempontja.

A második szempont a teszthalmaz. A megfigyelt értékeket a két indexhalmaz alapján egy kétdimenziós mátrixba rendezzük, ahol a sorok jelölik a megfigyelési egységeket, az oszlopok pedig a változókat (teszteket). Ez a mátrix Tucker terminológiájában kétszempontú mátrix. Ha az adatokat többször lekérdeztük (megismételtük a lekérdezést ugyanazon a mintán), akkor ez az ismétlés egy újabb szempontot, a harmadik szempontot jelenti, így az adatokat egy tömbbe helyezzük, ahol a horizontális réteg a megfigyelési egységeket jelöli, a vertikális réteg a teszteket, a harmadik réteg a különböző alkalmakat. Természetesen további szempontokat is bevezethetünk, és így beszélhetünk általánosan n-szempontú mátrixról. Tucker javaslata és megközelítése a komplex adatok absztrakt kapcsolatainak felderítésére mindmáig a legmeghatározóbb az exploratív strukturális elemzésekben.

Tradicionálisan a faktoranalitikus alkalmazásoknál az entitásokat két csoportba soroljuk – leggyakrabban emberekre és változókra (tesztekre) –, és a kapcsolatot azok az értékek jelzik, amelyeket a személyre és azok minden jellemzőjére, tulajdonságára, minden változójára megfigyeltünk. (Az entitás vonatkozhat dolgokra vagy megfigyelt objektumokra.)

Tehát minden egyes emberhez és változóhoz tartozik egy megfigyelt érték. Tucker (1966) az entitásoknak ezeket a halmazait szempontoknak (modes), a klasszifikáció szempontjainak nevezte. Így a faktorelemzésben a klasszifikáció két szempontja: az emberek és a tesztek. Ha bevezetjük az entitásoknak egy újabb, az előzőektől független halmazát – pl. más szituációt vagy alkalmat –, akkor a szempontok száma emelkedik, és ezek növelésével remélhetjük, hogy komplexebben és pontosabban tudunk klasszifikálni. A többszempontú elemző eljárások az így többszörösen klasszifikált adatok szisztematikus reprezentációját biztosítják.

18.1. Faktoranalitikus megközelítések

A hagyományos faktorelemzésben a közös faktorok terének paramétereit a mintából becsüljük, a megfigyelt változóknak (teszteknek, attribútumoknak) a vizsgálati egységekre (az entitások egy halmazára) – emberekre – vonatkozó megfigyelt értékeiből számított kétváltozós lineáris kapcsolatai alapján (kétszempontú, – vagy másképpen – kétdimenziós mátrixból). Minden faktort a megfigyelt változókból származtatunk, minden faktort latens változónak tekintünk, amely felelős a megfigyelt változók kovarianciáiért. Az egyéni különbségeket faktorértékeknek (factor scores) nevezzük. Eszerint – faktorok a megfigyelt változókból képződnek, a faktorértékek pedig a megfigyelési egységeknek a faktortérben elfoglalt helyét fejezik ki.

További szempontot bevezetve Tucker az Eckart–Young dekompoziciós elméletéből (1936) kiindulva a Kronecker-szorzatoperátor segítségével dolgozta ki eljárását. Tucker

eljárása – a klasszikus faktorelemzéssel ellentétben, ahol a kétszempontú adatmátrixból a latens hatásoknak csak egyetlen halmazát származtathatjuk – lehetővé teszi, hogy mindegyik szempontból különböző mögöttes, latens struktúrát tárjunk fel. Tucker felállítja a belső struktúra (inner core) fogalmát, amely az egyes szempontok külön komponenselemzéssel meghatározott faktorainak a kapcsolatát fejezi ki. Ezek tulajdonképpen olyan komponens-súlyok, amelyek összekapcsolják az eredeti megfigyelt szempontokból származtatott faktorokat.

A Tucker-féle háromszempontú komponenselemzést jól jellemzi a 18.1. ábra (forrás: Research Methods for Multimode Data Analysis, ed. Henry G. Law, Conrad W. Snyder, Jr., John A. Hattie and Roderich P. McDonald, New York: Praeger, 1984). Ez az ábrázolás Kroonenberg TUCKALS3 nevű programján alapul, aki az alternáló legkisebb négyzetek módszerét alkalmazta a teljes modell illesztésére.

Tucker megoldásának érdekes változata Kroonenberg modellje (Kroonenberg and de Leeuw, 1980), amely a belső struktúra mátrixra helyezi a súlyt, de a komponensek –latens változók mellett – megőrzi az adott klasszifikációs egységet. Azok mindegyikére külön számolja a belső struktúra mátrixot, ezért kiterjesztett belső struktúra módszernek (extended core method) is nevezik eljárását.

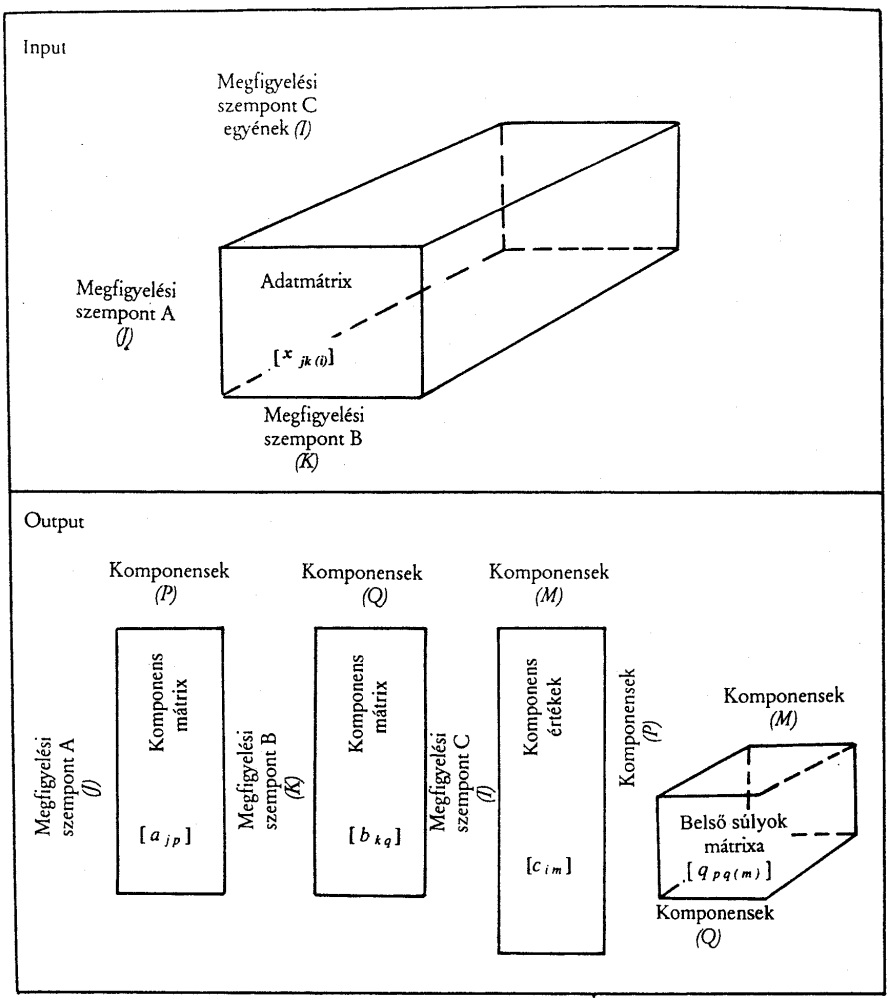
Ez a modell megegyezik Carroll és Chang (1972) IDIOSCAL modelljével abban az esetben, ha a másik két szempont szimmetrikus.

Kroonenberg modellje azoknál az alkalmazásoknál hasznos, ahol az egyik megfigyelt szempont lényegében nem faktorizálható (nincs értelmezhető mögöttes faktor), például az idősoroknál, egyes részleges megfigyelési egységeknél vagy specifikus feltételek figyelembevételénél. A modellt Kroonenberg TUCKALS2 nevű programja alapján a 18.2. ábra reprezentálja.

A kiterjesztett belső struktúra módszert értelmezhetjük a hierarchikus (másodrendű) dekompozíciók terminológiája szerint is. Eszerint az első szempont szerint számítjuk az elsőrendű faktorokat, amelyek azután a másik szempont szerint kapcsolódnak a másodrendű faktorokban, majd a harmadik szempont minden kategóriájában, és így tovább. Ezt szemlélteti a 18.3. ábra.

Harshman (1970) párhuzamos faktorok modellje (PARAFAC) Catell (1944) „arányos profilok elméletéből" (principle of proportional profiles) kiindulva az előzőektől különböző megoldást adott. Catell elmélete azt fejezi ki, hogy a „reális" faktorok és az ezzel összefüggő faktorértékek megőrzik struktúrájukat, csak arányosan változnak a másik szempont, feltétel szerint, akkor, ha változik fontosságuk a kovarianciák magyarázatában az egyes feltételek szerinti kategóriákban. Eljárása matematikailag megegyezik Carroll és Chang (1970) CANDECOMP eljárásának általános háromutas (three-way) esetével. A PARAFAC lényege, hogy háromszempontú (háromdimenziós) adatmátrixból kiindulva mindegyik szempontra megadja a faktorokat (faktormátrixokat), amelyek mindegyike a lehető legjobban becsüli az adott szempont változóit, de a három szempontra szimultán becslést ad a váltakozó (alternáló) legkisebb négyzetek becslési eljárásának a felhasználásával. A PARAFAC eljárása multiplikatív modell, így szigorúan feltételezi az adatok mérési típusát (az adatoknak „arány" típusúaknak kell lenniök).

A PARAFAC módszerét lehet alkalmazni kis minták esetén is, így jól alkalmazható alminták esetén keresztvalidálásra. A PARAFAC, különösen akkor, ha a modell nem illeszkedik az adatokhoz, lehetővé teszi, hogy az adatokon olyan transzformációkat végezzünk, amelyek biztosítják az adatok aránymérési skála típusát. A PARAFAC-eljárást mutatja a 18.4. ábra. Az előző eljárás a modellt közvetlenül a háromszempontú adatmátrixhoz illeszti. Különböző eredményt kaphatunk, ha a modellt a kovariancia struktúrához illesztjük. Ebben az esetben, mivel a kovarianciákat a megfigyelési

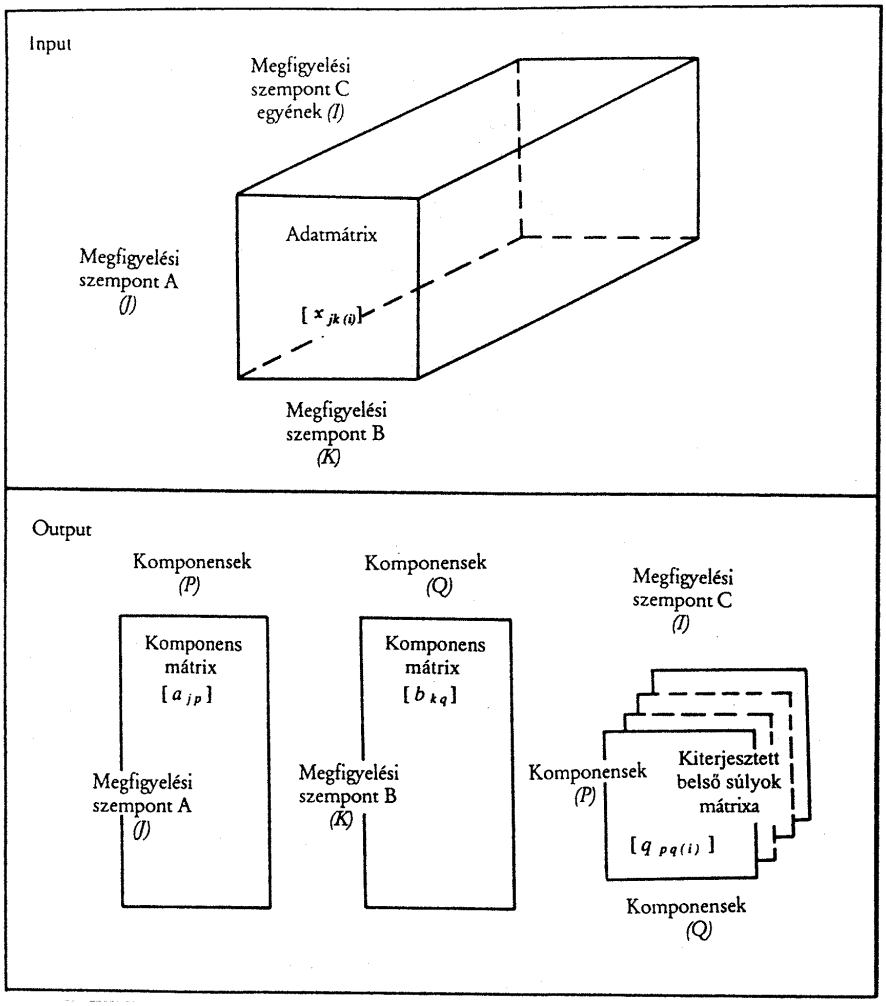


Modell

$$x_{jk(i)} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{m=1}^{M} a_{jp} b_{kq} c_{im} g_{pq(m)} + e_{jk(i)}$$

Számítógépes program: TUCKALS3

18.1. ábra: A háromszempontú főkomponens-elemzés

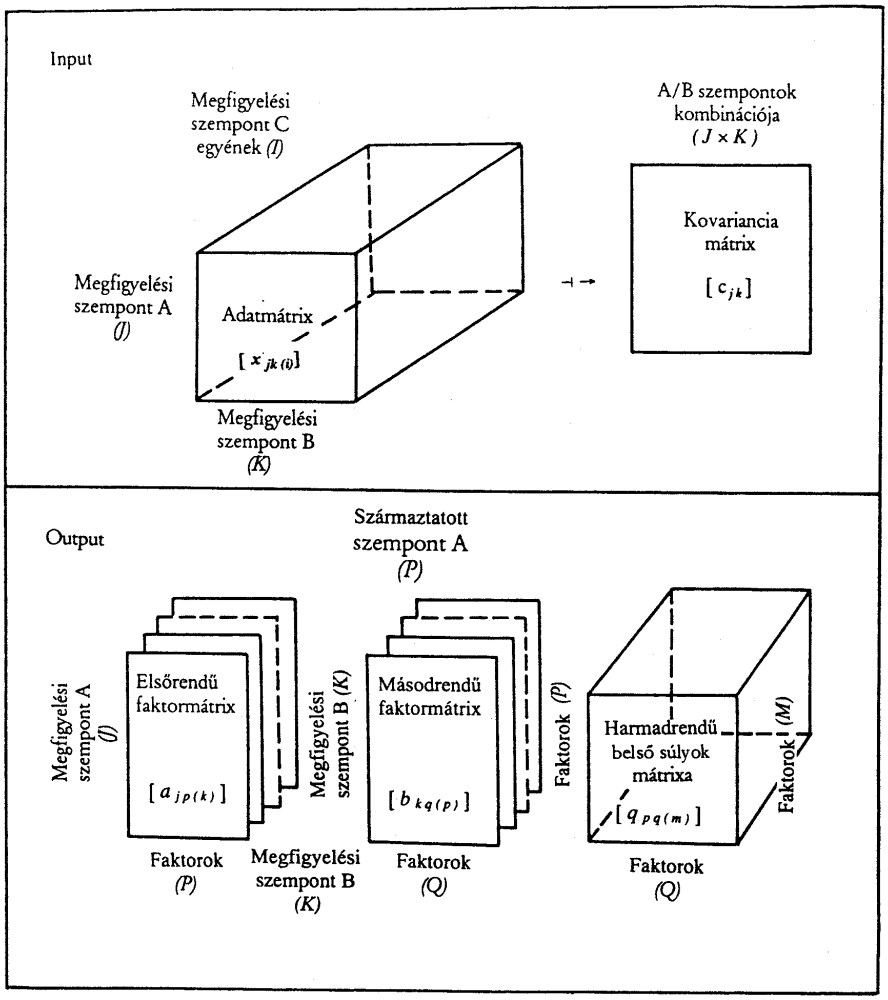


Modell

$$x_{jk(i)} = \sum_{p=1}^{P} \sum_{q=1}^{Q} a_{jp} b_{kq} g_{pq(i)} + e_{jk(i)}$$

Számítógépes program: TUCKALS2

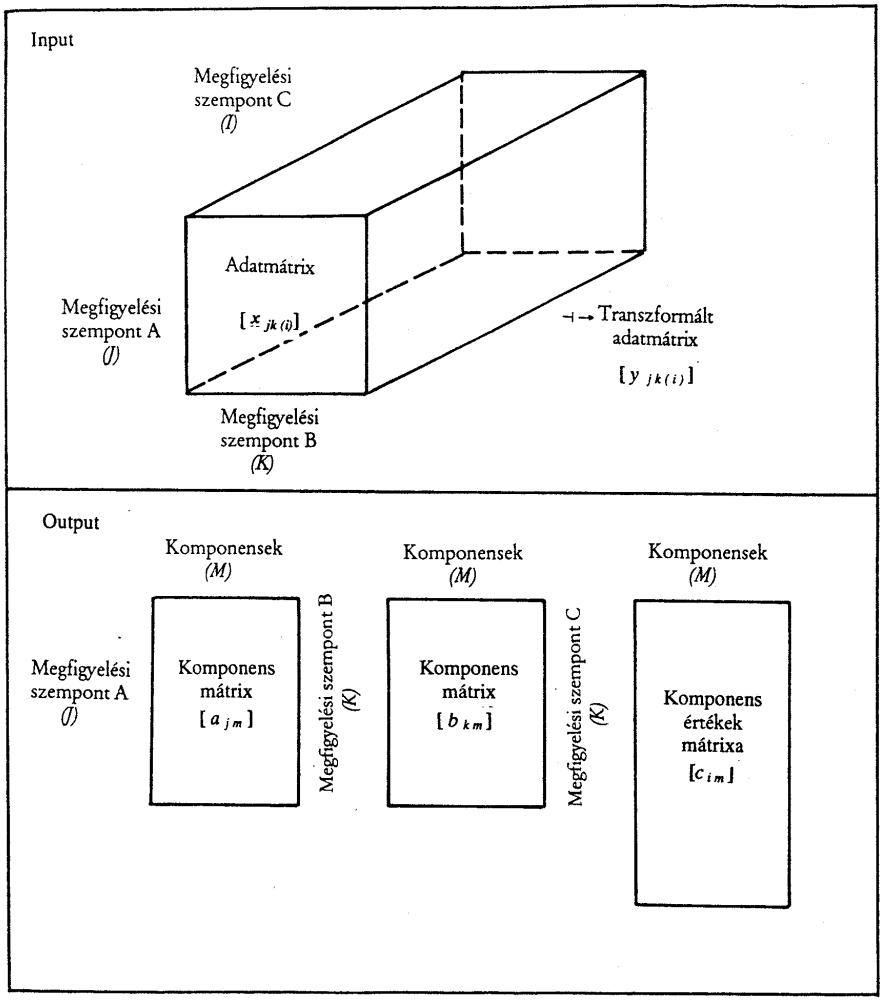
18.2. ábra: Háromszempontú főkomponens-elemzés.
(Kiterjesztett belső struktúra modell)



Modell

$$x_{jk(i)} = \sum_{p=1}^{P} a_{jp(k)} \sum_{q=1}^{Q} b_{kq(p)} \sum_{m=1}^{M} c_{im} g_{pq(m)} + e_{jk(i)}$$

18.3. ábra. Háromszempontú hierarchikus dekompozició.
(A/B szempontok kombinációja)



Modell

$$x_{jk(i)} \dashv\vdash \rightarrow y_{jk(i)} = \sum_{m=1}^{M} \left(a_{jm} b_{km} c_{im}\right) + e_{jk(i)}$$

Számítógépes program: PARAFAC

18.4. ábra. Párhuzamos faktorok elemzése (Direkt módszer)

egységek (pl. személyek) alapján számoljuk, az egyik szempont a három közül e számítás miatt elvész, így a súlymátrixokat csak a másik két szempontra kapjuk meg. Ezt a modellt illusztrálja a 18.5. ábra. A faktorokat ugyan hasonlóan értelmezhetjük, azonban az indirekt megoldás esetén a faktorokban a partikuláris megfigyelések hatása mellett a szisztematikus hatás is benne van. Azokat a faktorokat, amelyeket a megfigyelési egységek egy részhalmazánál azonosíthatunk, sokkal inkább a hatások jellegzetes típusának, mint a latens forrásnak tekinthetjük.

Az ilyen individuális típusú és szisztematikus típusú hatások felderítésére jó stratégia, ha a két módszert egymás után alkalmazzuk.

McDonald (1984) javasolt egy modellt, amely nagyon hasonlít a PARAFAC-modellhez, de sokkal inkább a hagyományos megoldásokon alapul. A számítógépes program neve COSAN (Fraser, 1980). A COSAN a modellek különböző formáit képes kezelni. Egyik előnye, hogy a hibatagot lehet struktúrálni, a másik, hogy maximum likelihood becslést ad, és így aszimptotikus khi-négyzet tesztet is biztosít.

Az előzőektől eltérő megközelítést alkalmaz Catell (1966, 1980) az n-utas faktorelemzéssel. Catell egyik modelljében, az együttes n-utas faktorelemzésben (Conjoint n-Way Factor Analysis) felcseréli az egyes szempontok kombinációit, és ezzel az adatmátrixot kétszempontúvá feszíti ki. Ebből számolja a korrelációkat, és határozza meg a faktormátrixokat. Ezt jól szemlélteti a 18.6. ábra.

Catell másik modelljében, az additív attribútum modellben az adatokat diszjunkt faktorokkal reprezentálja, kétutas multiplikatív tagok összegével fejezi ki. Ebben a modellben a kétutas faktorelemzést alkalmazza az adatok mindegyik felületére (face). A felület az eredeti háromutas adatmátrixokból képzett kétutas mátrix, amelyben a harmadik dimenzió szerint vesszük az elemek átlagát. Ezt mutatja a 18.7. ábra.

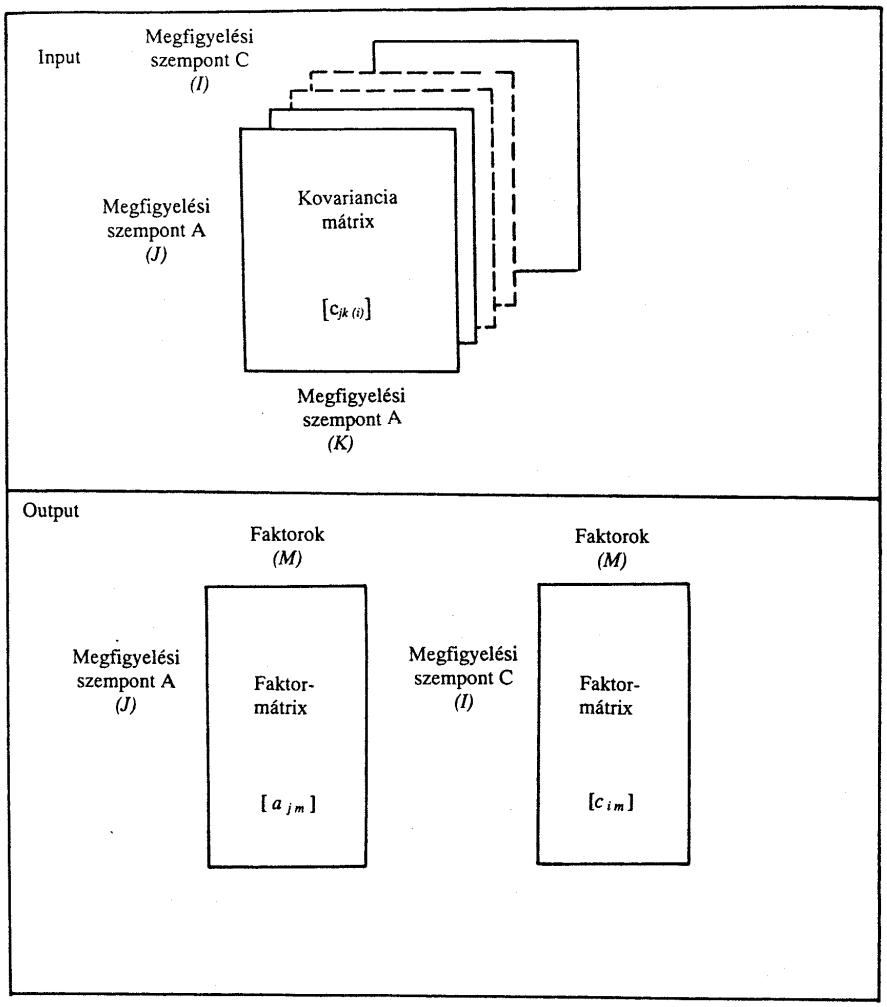
18.2. Skálázás és az ezzel kapcsolatos modellek

A klasszikus sokdimenziós skálázásnál a stimulusok (változók) hasonlósági vagy különbözőségi mátrixaiból (kétutas mátrix) indulunk ki, és a cél megtalálni a többdimenziós térben a stimulusoknak megfelelő pontokat úgy, hogy a konfigurációjuk a lehető legjobban feleljen meg az eredeti hasonlósági vagy különbözőségi mátrixnak (vagyis a leghasonlóbb két stimulus legyen a legközelebb egymáshoz és így tovább). Az eredmények értelmezésénél elsősorban a stimulusok klasztereire, térbeli elrendeződésére koncentrálunk, így az egyéni különbségek, eltérések nem jelennek meg (a klasszikus kétutas faktorelemzésnél éppen az egyéni különbségek struktúráját (pattern of individual differences) vizsgáljuk.

Carroll és Chang (1970) javasolt egy eljárást a sokdimenziós skálázásnál az egyéni különbségek figyelembevételére, amely Eckart-Young CANDECOMP nevű kanonikus dekompozíciós eljárásának n-utas általánosítása.

A CANDECOMP tulajdonképpen a háromszempontú komponenselemzés speciális esete, amikor feltételezzük az azonos dimenzionalitást az egyes megfigyelési egységeknél. A CANDECOMP háromdimenziós változata matematikailag megegyezik a PARAFAC direkt módszerével.

Carroll és Chang módszere (amely tehát a CANDECOMP-eljárást használja) az INDSCAL, az egyéni különbségek skálázása. Az INDSCAL-modellben minden meg-

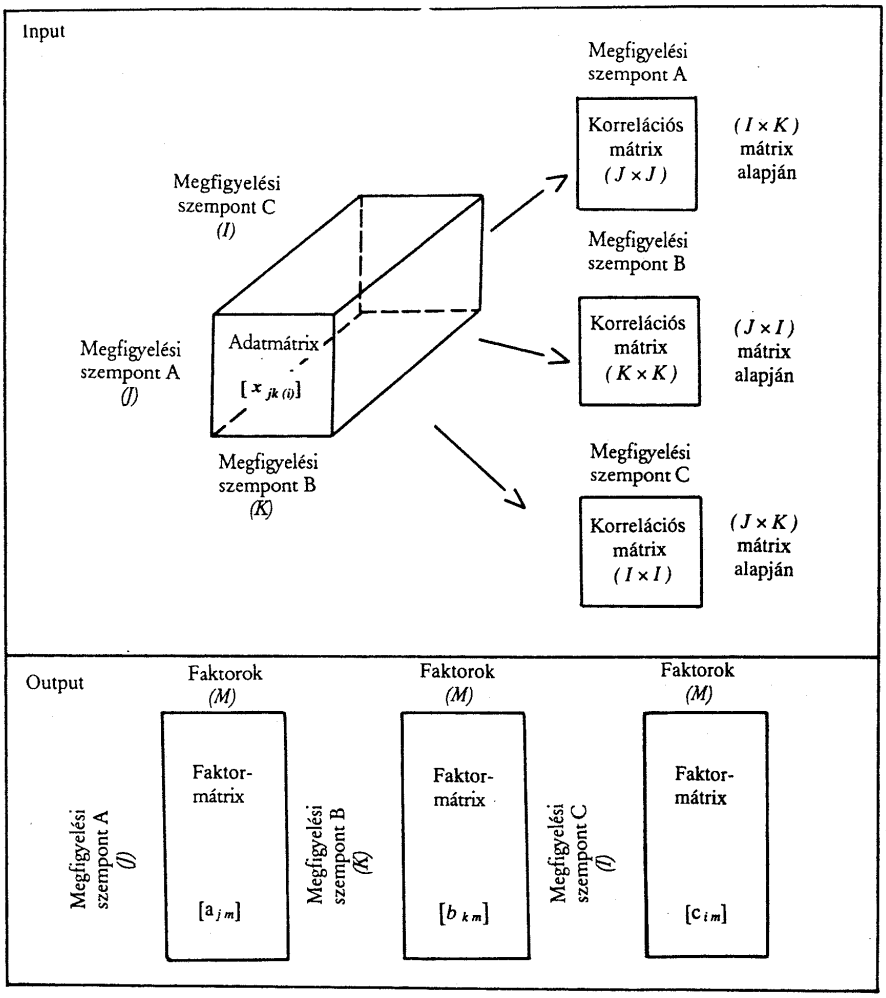


Modell

$$c_{jk(i)} = \sum_{m=1}^{M} \left(a_{jm} a_{km} c_{im} \right) + e_{jk(i)}$$

Számítógépes program: PARAFAC

18.5. ábra. Párhuzamos faktorok elemzése (Indirekt módszer)

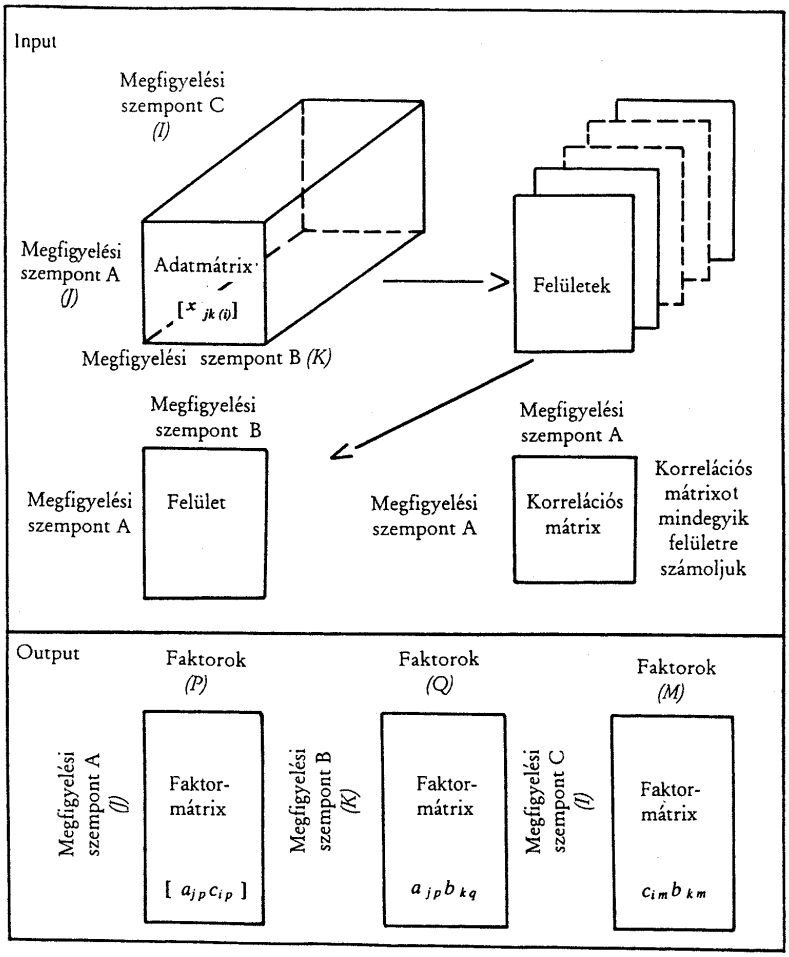


Modell

$$x_{jk(i)} = \sum a_{jm} b_{km} c_{im} + u_{jk(i)} + e_{jk(i)}$$

Számítógépes program: Standard Faktor Analízis

18.6. ábra. Együttes n-utas faktorok elemzése



$$x_{jk(i)} = \sum_{p}^{P} a_{jp}c_{ip} + \sum_{q}^{Q} a_{jq}b_{kq} + \sum_{m}^{M} c_{im}b_{km} + u_{jk(i)} + e_{jk(i)}$$

Számítógépes program: Standard Faktor Analízis

18.7. ábra. Diszjunkt (Face) *n*-utas faktorelemzés

figyelési egységre rendelkezünk a stimulusok (változók) hasonlósági (vagy különbözőségi) mátrixával. Eredményül pedig részben a stimulusok közös terét, részben a közös tér dimenzióira vonatkozó egyedi súlyokat kapjuk. A közös tér (group stimulus space) dimenziói a megfigyelések közös „faktorai".

A megfigyeléseknek (egyedeknek) a közös tér dimenzióira vonatkozó súlyai szerint az egyéni térben az egyén a közös, csoport stimulus tér egyes dimenzióit különféleképpen súlyozhatja (megnyújthatja vagy zsugoríthatja). Ezt mutatja a 18.8. ábra.

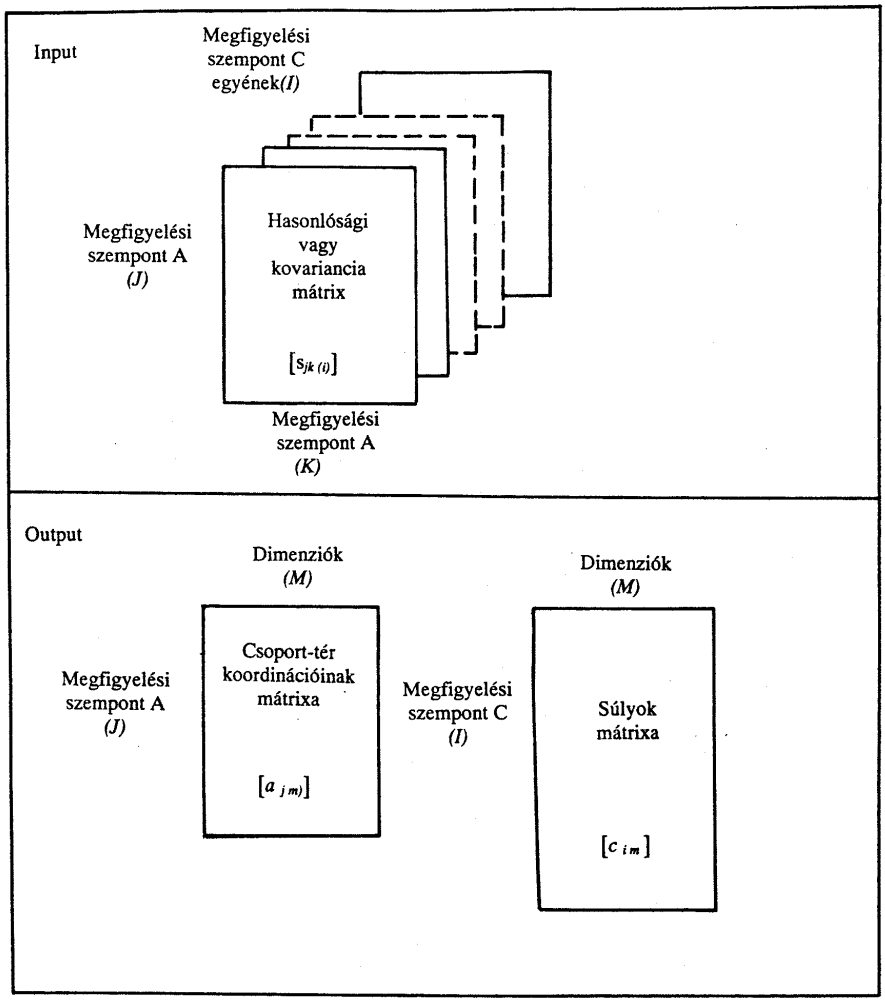
A CANDELINC (CANonical DEcomposition with LINear Constraints) eljárásban (Carroll, Green and Carmone, 1976) lehetőség van arra, hogy a modell paramétereire lineáris feltételeket tegyünk. Ha a megfigyeléseket csoportosítjuk valamilyen külső változó(k) szerint, pl. nem, életkor, iskolai végzettség stb., a paraméterekre megfelelő feltételeket téve, megvizsgálhatjuk, hogy illeszthető-e egy, csak fő hatásokat tartalmazó MANOVA-modell az adatokhoz.

A CANDECOMP-eljárás ugyancsak illeszthető Lazarsfeld (1960) latens-osztály modelljéhez.

Az INDSCAL és a vele kapcsolatos modellek nagyon hasonlítanak a faktoranalízishez: egy redukált, releváns latens teret határoznak meg, amelyben a lehető legjobban reprodukálódnak az empirikus relációk. A faktoranalízisnél a faktorokra *a priori* feltételezéseket teszünk, a faktortér, amelybe a megfigyelt változóvektorokat illesztjük, szemantikusan korlátozottan értelmezhető. A skálázásnál közvetlenül a páronkénti hasonlósági együtthatókból indulunk ki, és a közös tér dimenziói a megfigyelési egységeknek egyfajta átlagos latens tulajdonságait fejezi ki, és az egyes megfigyelések ezt a teret a saját súlyaik alapján transzformálják az egyéni térbe. Az INDSCAL újabb változata, az IDIOSCAL (Individual DIfference in Orientation SCALing) a csoport-tér ortogonális transzformációját is lehetővé teszi.

Lingoes és Borg (1976) PINDIS (Procrustean INdividual DIfferences Scaling) biztosítja mind az INDSCAL, mind az IDIOSCAL transzformációs lehetőségét, ahogyan ezt a 18.9. ábra is mutatja. Carroll és Arabie (1984) a folytonos latens dimenziók helyett diszkrét struktúrát kereső modellt javasoltak, amelyben az egyének különbözően súlyozhatják a klaszterek közös halmazát. Az INDCLUS (INdividual Differences CLUStering) az INDSCAL-modell diszkrét (klaszter) változata. Ezt a modellt mutatja a 18.10. ábra. Young (1984) a GEMSCAL (General Euclidean Model SCALing) modellben a megfigyelési egységeket (az egyéneket) a közös térbe illeszti, és minden egyént egy vektorral reprezentál a közös térben. Ennek az irányvektornak a helyzete jellemzi az egyént, részben iránya jelzi, hogy az egyén milyen stimulusokat preferál, a hossza pedig az egyes dimenziók relatív fontosságát mutatja. Ezt a modellt mutatja be a 18.11. ábra. Az INDSCAL-, PINDIS- és GEMSCAL-típusú elemzésekkel szimmetrikus mátrixoknak egy adott halmazát elemezhetjük. Amikor a háromdimenziós adatok nemszimmetrikus mátrixokból állnak, akkor a megfelelő modell a súlyozott egyéni különbségek unfolding modellje lesz.

De Sarbo és Carroll (1981) Coombs (1976) unfolding modelljét általánosította. A modell feltételezi, hogy minél kisebb a többdimenziós térben a stimulus pontok és egy egyén ideális pontja között a súlyozott euklideszi távolság, annál jobban preferálja az egyén ezeket a stimulusokat. Ezt a modellt mutatja be a 18.12. ábra. Az eddig tárgyalt modellek metrikus adatokhoz kapcsolódtak. Az ALSCAL (Alternating Least-squares SCALing) (Tokane, Young és de Leeuw, 1977) számítógépes program lehetőséget biztosít nemmetrikus adatok kezelésére is. Tokane és munkatársai fejlesztették ki a varianciaelemzés típusú additív modelleket nemnumerikus adatok elemzésére. Ezek a modellek egy függő változót magyaráznak két vagy több faktor hatásával. Ezeket a faktorokat

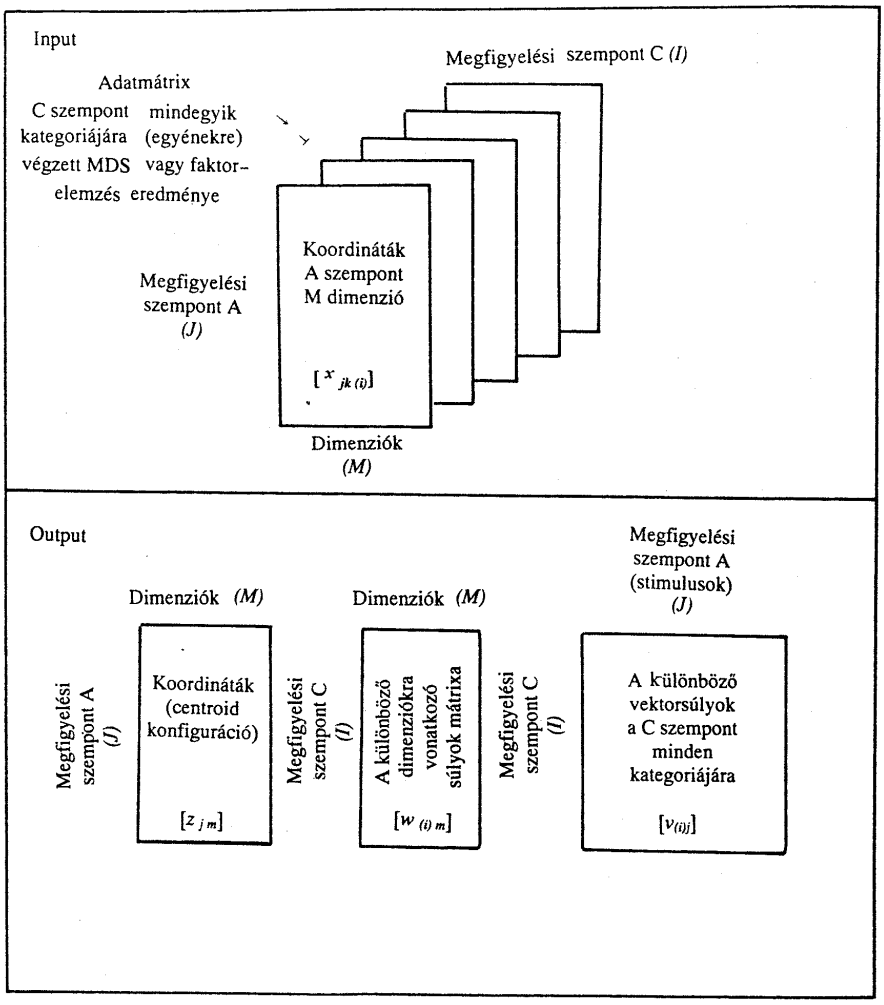


Modell

$$s_{jk(i)} = \sum_{m=1}^{M} c_{im}(a_{jm} - a_{km})^2 + e_{jk(i)}$$

Számítógépes program: INDSCAL

18.8. ábra. Szimmetrikus trilineáris elemzés
(Súlyozott egyedi különbözőségek)



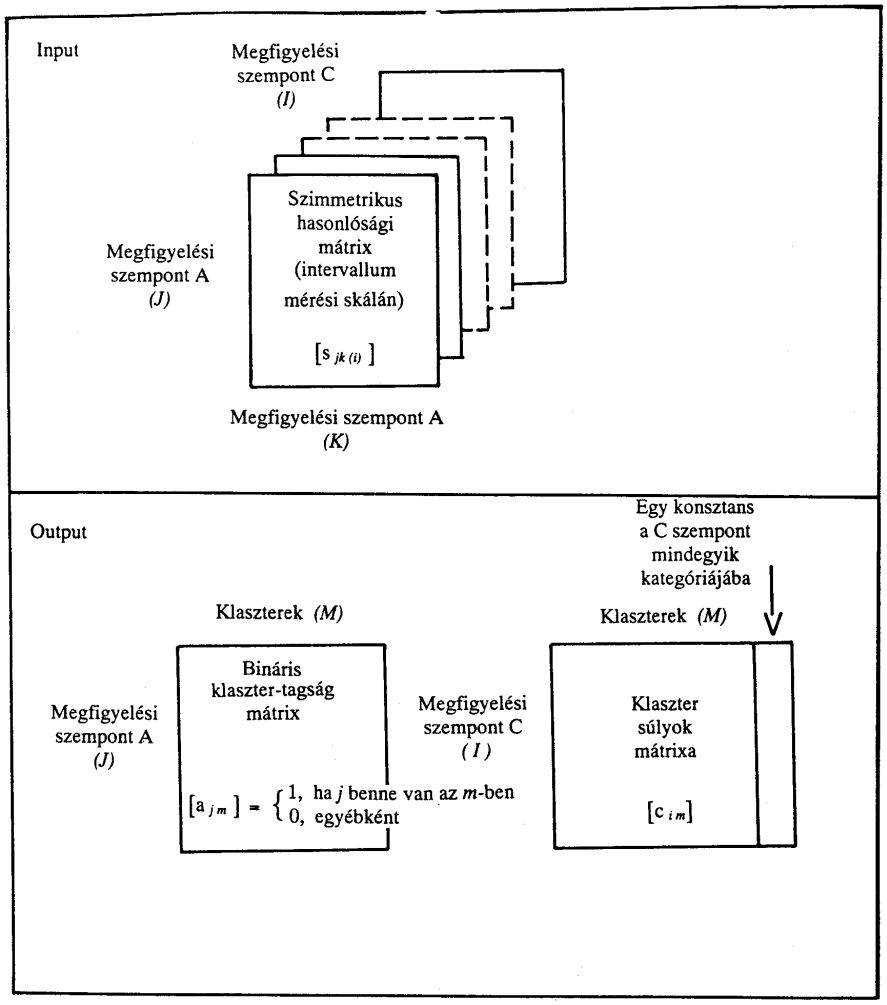
Modell

A transzformációk hierarchikus sorrendje:

1. Hasonlósági transzformáció
2. A dimenziók súlyozása
3. A dimenziók idioszinkratikus súlyozása
4. Vektorsúlyok rögzített középponttal
5. Vektorsúlyok idioszinkratikus középponttal

Számítógépes program: PINDIS

18.9. ábra. Prokrusztészi egyéni különbségek skálázása

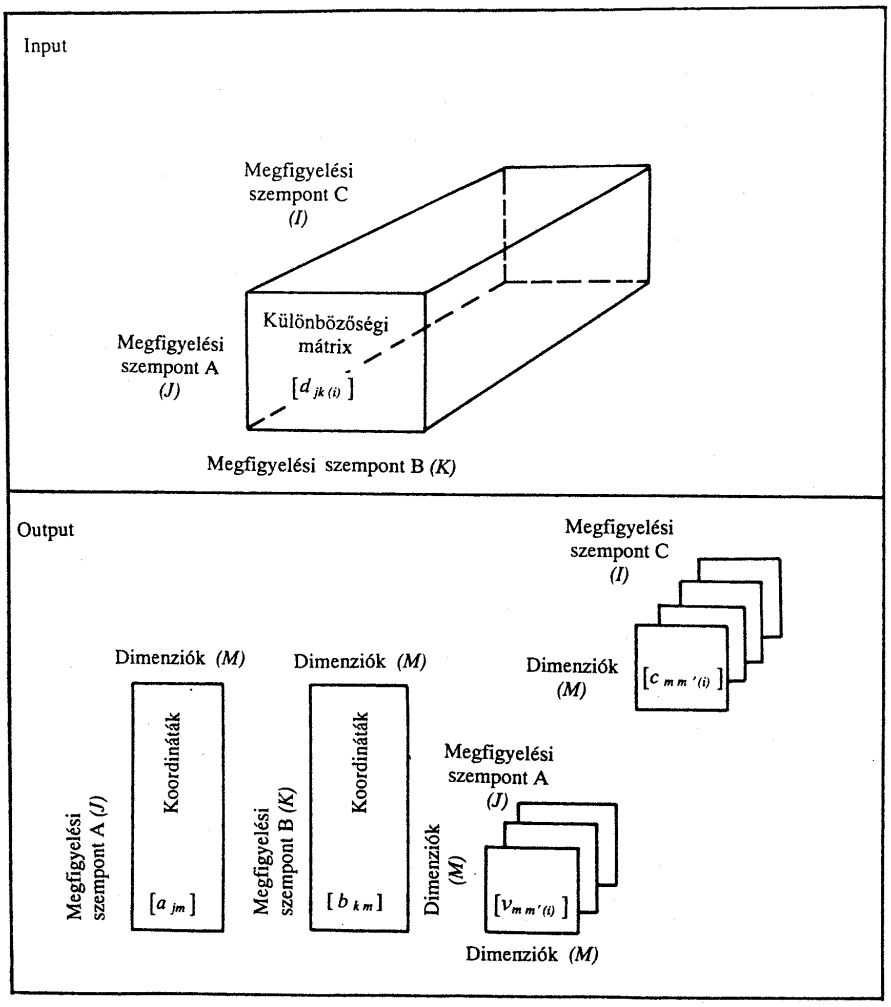


Modell

$$s_{jk(i)} = \sum_{m=1}^{M} c_{im} a_{jm} a_{km} + \text{konstans}_{(i)} + e_{jk(i)}$$

Számítógépes program: INDCLUS

18.10. ábra. Egyéni különbségek klaszterezése

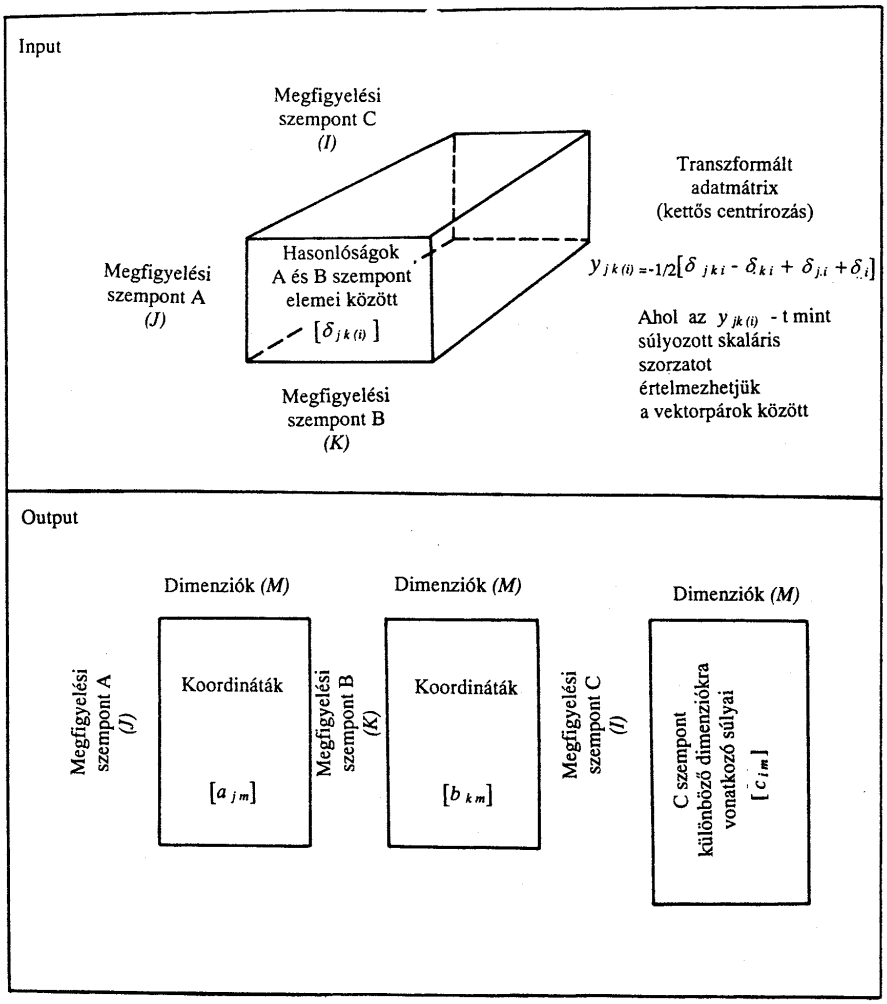


Modell

$$d^2_{jk(i)} = (\mathbf{b}_k - \mathbf{a}_j)' \mathbf{V}_j \mathbf{C}_i (\mathbf{b}_k - \mathbf{a}_j)$$

Számítógépes program: GEMSCAL

18.11. ábra. Általános euklideszi modell skálázása

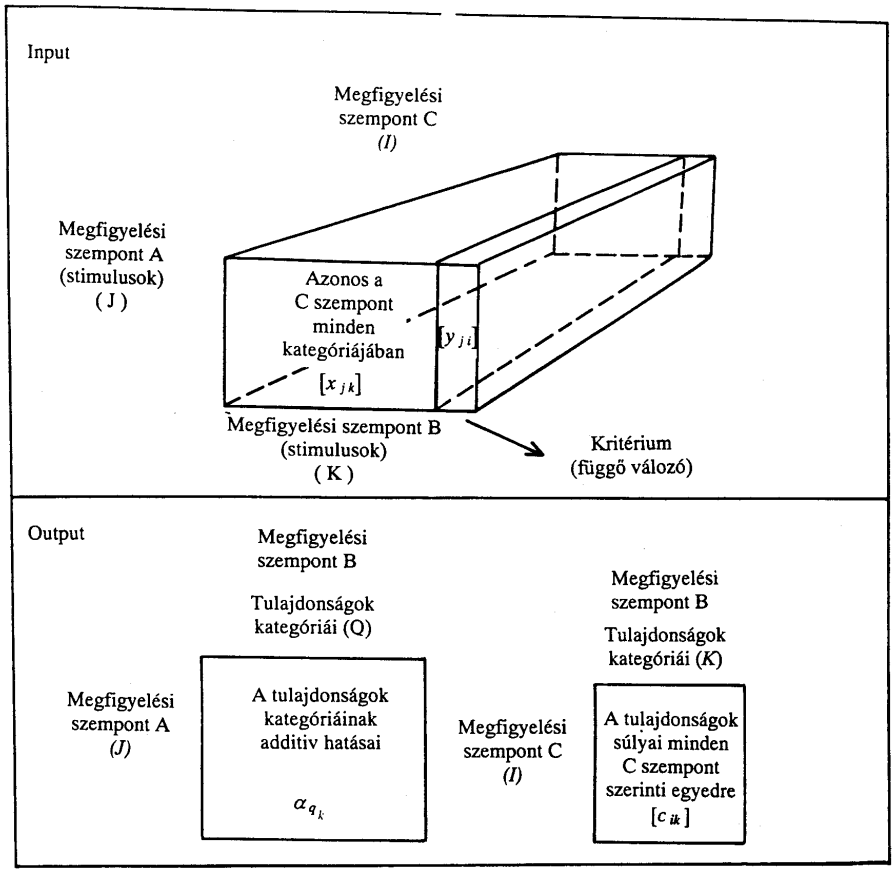


Modell

$$\widehat{y}_{jk(i)} = \sum_{m=1}^{M} a^*_{jm} b^*_{km} c_{im} \quad \text{ahol} \quad \begin{matrix} a^*_{jm} = \alpha(a_{jm} - a_{.m}) \\ b^*_{km} = \beta(b_{km} - b_{.m}) \end{matrix}$$

Számítógépes program: CANDECOMP

18.12. ábra. Nemszimmetrikus trilineáris elemzés (Háromutas unfolding)



Modell

$$y_{j(i)} = \sum_{k}^{K} c_{ik} \left(\sum_{q_k}^{Q_k} h_{jq_k} \alpha_{q_k} \right) + e_{j(i)} \quad \text{ahol}$$

$$h_{jq_k} = \begin{cases} 1, & \text{ha } j \text{ stimulus a tulajdonság} \\ & q_k \text{ kategóriájába tartozik} \\ 0, & \text{egyébként} \end{cases} \qquad \text{és } \widehat{x}_{jk} = \sum_{q_k}^{Q_k} h_{jq_k} \alpha_{q_k}$$

Számítógépes program: WADDALS MAXADD

18.13. ábra. Súlyozott additív modell

az egyének kölönbözőképpen súlyozhatják. Az előző módszerekhez képest a különbség alapvetően ott van, hogy ezek a faktorok nem latens változók, hanem a tulajdonságok (változók) újrakvantifikálását jelentik. A WAM (Weighted Additiv Model) modellt a 18.13. ábra illusztrálja (Tokane, 1982).

18.3. A háromszempontú főkomponens-elemzés

A klasszikus főkomponens-elemzésnél abból a feltételezésből indulunk ki, hogy egy változóhalmaznak a vizsgálati egységekben, megfigyelési egységekben mért, megfigyelt értékeit jól becsülhetjük lényegesen kevesebb számú, nem megfigyelt, latens változókkal, amelyek az eredeti változók lineáris kombinációi. Ezek a latens változók a megfigyelt változók varianciáinak nagy részét megmagyarázzák, és főkomponensnek nevezzük őket. Amennyiben a kovarianciamátrix lehető legjobb reprodukálása a cél, akkor faktorokról – és faktorelemzésről – beszélünk.

Tételezzük fel, hogy a változókat ugyanazon megfigyelési egységeken (ugyanazon a mintán, vizsgálati személyeken) többször, különböző feltételekkel megfigyeltük, így egy újabb szempontot bevezetve az adatokat háromdimenziós mátrixba rendezhetjük.

A főkomponens-elemzés logikáját alkalmazhatjuk mind a három szempontra. A latens változók feltételezése mellett lehetséges, hogy a megfigyelési egységeket reprodukálni tudjuk „ideális" egységek, személyek lineáris kombinációjával. Feltételezhetjük azt, hogy léteznek ilyen „ideális" személyek, amelyek különböző súlyú lineáris kombinációjával becsülhetők a vizsgált személyek.

Hasonlóan lehetséges, hogy a különböző feltételeknek, a harmadik szempontnak is van valamilyen „prototípusa", ideális feltétele. Vagyis mindhárom szempont esetében feltételezhetjük, hogy van mögöttük egy latens, nem megfigyelt dimenzió (vagy dimenziók), amely(ek) az adott szempont elemei közötti kapcsolatokat jól magyarázzák. Hogy ezeket a mögöttes, latens dimenziókat ne keverjük össze, a változók esetében a változók kapcsolatait jól leíró, becslő, mögöttes, nem megfigyelt, a változók számánál lényegesen kevesebb számú, így információt sűrítő, komplex változót „latens változónak" nevezzük, a hasonló tulajdonságú latens megfigyelési egységet „ideális" személynek, míg a harmadik szempont szerinti latens dimenziókat „prototípusoknak". A fentiek mellett még azt is feltételezhetjük, hogy a latens változók, „ideális" személyek és a prototípusok között is lehet kapcsolat.

Tételezzük fel az egyszerűség kedvéért, hogy 2 latens változónk p_1 és p_2, két ideális személyünk m_1 és m_2, és két prototípusunk q_1 és q_2 van. Tételezzük fel, hogy ismerjük az m_1 ideális személy p_1 latens változóra és q_1 prototípusra vonatkozó értékét:

$$g_{p_1q_1(m_1)} \quad \text{vagy} \quad g_{11(1)}\,.$$

Hasonlóan az m_1 ideális személy p_1 latens változóra és q_2 prototípusra vonatkozó értéke

$$g_{p_1q_2(m_1)} \quad \text{vagy} \quad g_{12(1)}.$$

Hasonlóan tételezzük fel, hogy ismerjük a többi értéket is:

$$g_{21(1)},\ g_{22(1)}.$$

Ugyanígy a második ideális személynek a másik két szempont latens dimenziójára vonatkozó értékei:

$$g_{11(2)},\quad g_{12(2)},\quad g_{21(2)},\quad g_{22(2)}.$$

A $g_{pq(m)}$ értékek, a belső struktúramátrix $\mathbf{G}$ értékei a különböző szempontok latens dimenzióinak kapcsolatát írják le. Nézzük most azt meg, hogy egy tényleges megfigyelés (i) j-edik manifeszt változóra és a k-adik feltételre (mérésre) vonatkozó $x_{jk(i)}$ értékét hogyan kaphatjuk meg.

Az i-edik megfigyelt személynek a p_1 latens változóra és q_1 prototípusra vonatkozó értékét megkaphatjuk az m_1 és m_2 ideális személyek lineáris kombinációjával:

$$s_{p_1q_1(i)} = c_{im_1} g_{p_1q_1(m_1)} + c_{im_2} g_{p_1q_1(m_2)},$$

vagy

$$s_{11(i)} = c_{i1} g_{11(1)} + c_{i2} g_{11(2)}.$$

Hasonlóan a p_2 latens változóra és a q_1 prototípusra az i-edik megfigyelt személy értéke:

$$s_{p_2q_1(i)} = c_{im_1} g_{p_2q_1(m_1)} + c_{im_2} g_{p_2q_1(m_2)},$$

vagy

$$s_{21(i)} = c_{i1} g_{21(1)} + c_{i2} g_{21(2)}.$$

A többi változó-kombinációra hasonlóan számíthatjuk az i-edik megfigyelt személy értékeit:

$$s_{12(i)} = c_{i1} g_{12(1)} + c_{i2} g_{12(2)},$$
$$s_{22(i)} = c_{i1} g_{22(1)} + c_{i2} g_{22(2)}.$$

A c_{im} súlyok a megfigyelt személy ideális személyre vonatkozó súlyai. Ezek a súlyok függetlenek attól, hogy melyik latens változóról és melyik prototípusról van szó. Az ideális entitások (szempontok) közötti kapcsolatokat a belső struktúramátrix ($\mathbf{G}$) tartalmazza. A következő lépésben azt mutatjuk meg, hogy az i-edik megfigyelés j-edik manifeszt változóra vonatkozó értékét hogyan számíthatjuk a különböző prototípusokra.

A q_1 prototípus esetén az i-edik személy j-edik megfigyelt változóra vonatkozó értéke:

$$\begin{aligned} v_{j1(i)} &= a_{jp_1} s_{p_1q_1(i)} + a_{jp_2} s_{p_2q_1(i)} \\ &= a_{j1} s_{11(i)} + a_{j2} s_{21(i)}. \end{aligned}$$

Hasonlóan a q_2 prototípusra:

$$v_{j2(i)} = a_{j1} s_{12(i)} + a_{j2} s_{22(i)},$$

ahol az a_{jp} súlyok a latens változók súlyai a j-edik megfigyelt változóra vonatkozóan. Végül nézzük meg, hogy az i-edik személy j-edik manifeszt változóra és k-adik feltételre vonatkozó megfigyelt értéke hogyan határozható meg:

$$\begin{aligned} x_{jk(i)} &= b_{kq_1} v_{jq_1(i)} + b_{kq_2} v_{jq_2(i)} \\ &= b_{k1} v_{j1(i)} + b_{k2} v_{j2(i)}, \end{aligned}$$

ahol a b_{kq} súlyok a latens feltételek, prototípusok súlyai a megfigyelt feltételre vonatkozóan.

Egyesítsük ezek után a három lépést:

$$x_{jk(i)} = \sum_q^2 b_{kq} v_{jq} = \sum_q^2 b_{kq} \left(\sum_p^2 a_{jp} s_{pq} \right),$$

$$x_{jk(i)} = \sum_q^2 b_{kq} \left(\sum_p^2 a_{jp} \left(\sum_m^2 c_{im} g_{pq(m)} \right) \right),$$

ahol $\sum_{m}^{2} c_{im} g_{pq(m)}$ az m_1 és m_2 ideális személy lineáris kombinációja,

$\sum_{p}^{2} a_{jp} \left(\sum_{m}^{2} c_{im} g_{pq(m)} \right)$ a p_1 és p_2 latens változó lineáris kombinációja,

$\sum_{q}^{2} b_{kq} \left(\sum_{p}^{2} a_{jp} \left(\sum_{m}^{2} c_{im} g_{pq(m)} \right) \right)$ a q_1 és q_2 ideális feltétel, prototípus lineáris kombinációja.

Az $x_{jk(i)}$-t a következőképpen írhatjuk:

$$x_{jk(i)} = \sum_{p}^{2} \sum_{q}^{2} \sum_{m}^{2} a_{jp} b_{kq} c_{im} g_{pq(m)}.$$

Általában feltételezzük, hogy a latens változók száma P, a prototípusok száma Q, az ideális személyek száma M. Mivel a gyakorlatban a modell illeszkedése nem tökéletes, feltétezzük, hogy létezik egy hibakomponens a becsült tag mellett:

$$x_{jk(i)} = \widehat{x}_{jk(i)} + e_{jk(i)},$$

$$x_{jk(i)} = \sum_{p}^{2} \sum_{q}^{2} \sum_{m}^{2} a_{jp} b_{kq} c_{im} g_{pq(m)} + e_{jk(i)}.$$

Az általános modellt a fentiek alapján a következőképpen írhatjuk:

$$x_{jk(i)} = \sum_{p}^{P} \sum_{q}^{Q} \sum_{m}^{M} a_{jp} b_{kq} c_{im} g_{pq(m)} + e_{jk(i)}.$$

Ezt az egyenletet a háromszempontú főkomponens-elemzés alapegyenletének nevezzük. A Tucker-féle kompozíciós szabály értelmében az alapegyenletet hat hierarchikus elrendezésben is felírhatjuk:

1. $$x_{jk(i)} = \sum_{p} a_{jp} \sum_{q} b_{kq} \sum_{m} c_{im} g_{pq(m)} + e_{jk(i)}$$

2. $$x_{jk(i)} = \sum_{p} a_{jp} \sum_{m} c_{im} \sum_{q} b_{kq} g_{pq(m)} + e_{jk(i)}$$

3. $$x_{jk(i)} = \sum_{q} b_{kq} \sum_{p} a_{jp} \sum_{m} c_{im} g_{pq(m)} + e_{jk(i)}$$

4. $$x_{jk(i)} = \sum_{q} b_{kq} \sum_{m} c_{im} \sum_{p} a_{jp} g_{pq(m)} + e_{jk(i)}$$

5. $$x_{jk(i)} = \sum_{m} c_{im} \sum_{p} a_{jp} \sum_{q} b_{kq} g_{pq(m)} + e_{jk(i)}$$

6. $$x_{jk(i)} = \sum_{m} c_{im} \sum_{q} b_{kq} \sum_{p} a_{jp} g_{pq(m)} + e_{jk(i)}$$

A modellben a következő paramétermátrixok szerepelnek.

$\mathbf{X}_{(J \times K \times I)}$ háromdimenziós adatmátrix (a megfigyelt, mérési adatokat tartalmazza), amely általános eleme $x_{jk(i)}$ az i-edik megfigyelési egység j-edik változóra és a k-adik feltételre (időpont stb.) vonatkozó megfigyelt értéke,

$\mathbf{A}_{(J\times P)}$ a megfigyelt változók és latens változók kapcsolatait tartalmazó komponens-mátrix (faktorsúlyok mátrixa), általános eleme a_{jp} a j-edik megfigyelt változó p-edig latens változóra (főkomponensre) vonatkozó súlya,

$\mathbf{B}_{(K\times Q)}$ a megfigyelési feltétel és az ideális prototípus kapcsolatait leíró mátrix (komponens-mátrix), általános eleme b_{kq}, a k-adik feltétel q-adik prototípusra vonatkozó súlya,

$\mathbf{C}_{(I\times M)}$ a megfigyelt és az ideális személyek kapcsolatait tartalmazza (komponensértékek, component score matrix), általános eleme c_{im} az i-edik megfigyelt személy m-edik ideális személyre vonatkozó súlya (ez megfelelne a Q-típusú főkomponenselemzés faktormátrixának),

$\mathbf{G}_{(P\times Q\times M)}$ a latens változók, ideális személyek és prototípusok kapcsolatait tartalmazó három-dimenziós belső struktúramátrix, általános eleme $g_{pq(m)}$ a p-edik latens változó q-adik latens feltételre (prototípusra) vonatkozó súlya az m-edik latens személynél (ideális személynél),

$\mathbf{E}_{(J\times K\times I)}$ a hibakomponensek (reziduálisok) háromdimenziós mátrixa, a modell és a megfigyelt adat közötti különbséget tartalmazza, általános eleme $e_{jk(i)}$ az i-edik megfigyelési egység j-edik változóra és k-adik feltételre vonatkozó mérési hibája.

Az **A**, **B** és **C** mátrixokról feltételezzük, hogy ortonormált mátrixok (az oszlopok – amelyek a latens változókat jelölik – merőlegesek egymásra és egységnyi hosszúságúak), és a soroknak a száma mindhárom mátrix esetében nagyobb vagy legfeljebb egyenlő az oszlopok számával.

Az **A**, **B**, **C** és **G** paramétermátrixok elemeinek becsléseit a legkisebb négyzetek módszerével, a

$$\sum_{j}^{J}\sum_{k}^{K}\sum_{i}^{I}\left(x_{jk(i)}-\widehat{x}_{jk(i)}\right)^2$$

függvény minimalizálásával kapjuk.

Ezt az eljárást tartalmazza Kroonenberg TUCKALS3 programja.

A legkisebb négyzetek módszere értelmében a mérési adatok négyzetösszege egyenlő a becsült adatok (a modell által becsült) négyzetösszege plusz a reziduálisok négyzetösszege:

$$\sum_{j}^{J}\sum_{k}^{K}\sum_{i}^{I}x_{jk(i)}^2=\sum_{j}^{J}\sum_{k}^{K}\sum_{i}^{I}\widehat{x}_{jk(i)}^2+\sum_{j}^{J}\sum_{k}^{K}\sum_{i}^{I}\left(x_{jk(i)}-\widehat{x}_{jk(i)}\right)^2,$$

vagy egyszerűbben:

$$SS(\text{adat})=SS(\text{becslés})+SS(\text{reziduális}),$$

ahol SS a négyzetösszeget (sum of squares) jelöli.

Mivel SS(reziduális) nagysága függ a teljes négyzetösszegtől, célszerű a relatív reziduális négyzetösszegek vizsgálata: SS(reziduális)/SS(adat).

A modell relatív illeszkedését az SS(becslés)/SS(adat) hányadossal mérhetjük.

Tucker háromszempontú faktorelemzésének alapegyenletét mátrix alakban is felírhatjuk. Ehhez felhasználjuk a Cartesius-féle szorzatot és a Kronecker-féle szorzatot.

A Tucker háromszempontú faktor modellje mátrix-jelölésekkel:

$$\mathbf{X}_{(I\times\overline{JK})}=\mathbf{C}_{(I\times M)}\cdot\mathbf{G}_{(M\times\overline{PQ})}\cdot\left(\mathbf{A}'_{\overline{(P\times J)}}\otimes\mathbf{B}'_{(Q\times K)}\right)+\mathbf{E}_{(I\times\overline{KJ})},$$

ahol

$\mathbf{X}$ mátrix sorai a megfigyelési egységeket tartalmazzák, az oszlopai pedig a Cartesius-féle szorzat értelmében $J \times K$ elemet tartalmaznak, vagyis a megfigyelt változók és feltételek értékeit mutatják minden megfigyelésre, a $\otimes$ pedig a Kronecker-féle szorzat. Mivel a háromdimenziós mátrixot a Cartesius-féle szorzat szerint háromféleképpen alakíthatjuk át kétdimenziós mátrixszá, az alapmodellt is háromféleképpen értelmezhetjük:

$$\mathbf{X}_{(I\times\overline{JK})} = \mathbf{C}_{(I\times M)} \cdot \mathbf{G}_{(M\times\overline{PQ})} \left(\mathbf{A}'_{(P\times J)} \otimes \mathbf{B}'_{(Q\times K)}\right) + \mathbf{E}_{(I\times\overline{JK})},$$

$$\mathbf{X}_{(J\times\overline{IK})} = \mathbf{A}_{(J\times P)} \cdot \mathbf{G}_{(P\times\overline{MQ})} \cdot \left(\mathbf{C}'_{(M\times I)} \otimes \mathbf{B}'_{(Q\times K)}\right) + \mathbf{E}_{(J\times\overline{IK})},$$

$$\mathbf{X}_{(K\times\overline{JK})} = \mathbf{B}_{(K\times Q)} \cdot \mathbf{G}_{(Q\times\overline{MP})} \cdot \left(\mathbf{C}'_{(M\times I)} \otimes \mathbf{A}'_{(P\times J)}\right) + \mathbf{E}_{(K\times\overline{IJ})}.$$

Ezt könnyen beláthatjuk, ha az alapegyenletet átírjuk a következő formába:

$$x_{jk(i)} = \sum_m c_{im} \sum_p \sum_q g_{pq(m)} \left(a_{jp} b_{kq}\right) + e_{jk(i)},$$

ahol a kettős szumma jelöli a $\overline{pq}$ Cartesius-féle szorzatot, az $(a_{jp}b_{kq})$ szorzat pedig az $(\mathbf{A}' \oplus \mathbf{B}')$ Kronecker-féle szorzat általános eleme az indexek átrendezésével. A p és q szerinti kettős szumma megfelel a $\mathbf{G} \cdot (\mathbf{A}' \oplus \mathbf{B}')$ szorzatnak.

Ha az összes főkomponenst figyelembe vesszük, és így $I = M$, $J = P$ és $K = Q$, ekkor az adatmátrixot pontosan reprodukálni tudjuk a modellel. A gyakorlatban általában csak az első két, három vagy négy latens változó vagy főkomponens érdekes igazán. Ebben az esetben a modellel becsüljük az adatokat és keressük a $\mathbf{C}$, $\mathbf{G}$, $\mathbf{A}$ és $\mathbf{B}$ paramétermátrixok azon becslését, amelyek esetén az $\widehat{\mathbf{X}} = \mathbf{CG}(\mathbf{A}' \otimes \mathbf{B}')$ becslés és a mérési adatok közötti különbség a lehető legkisebb.

A becsléseket a legkisebb négyzetek módszere alapján az

$$f(\mathbf{CG}, \mathbf{A}, \mathbf{B}) = |\mathbf{X} - \widehat{\mathbf{X}}|^2 = \mathbf{X} - \mathbf{CG}(\mathbf{A}' \otimes \mathbf{B}')|^2$$

függvény minimalizálásával kaphatjuk meg, ahol a | | jelölés a távolságfüggvényt jelenti. A minimalizálásnál figyelembe kell vennünk a modell feltételeit, vagyis azt, hogy a $\mathbf{C}$, $\mathbf{A}$ és $\mathbf{B}$ mátrixok oszlopai (a latens dimenziók) ortonormáltak legyenek. Kroonenberg és Jan de Leeuw (1980) bizonyította, hogy a feladatnak mindig létezik megoldása. Először megmutatták, hogy belső struktúramátrix $\mathbf{G}$ kifejezhető a $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ és $\mathbf{X}$ mátrixokkal:

$$\widehat{\mathbf{G}} = \mathbf{C}'\mathbf{X}(\mathbf{A} \otimes \mathbf{B}).$$

A becslési eljárásnál az ún. alternáló legkisebb négyzetek módszerét alkalmazták TUCKALS3-programjukban. Az alternáló legkisebb négyzetek módszere keresi egy paramétermátrix becslését a többi paramétermátrix rögzített értéke esetén. Miután minden paramétermátrixra megkapjuk a becslést, az eljárást folytatjuk egészen addig, amíg az nem konvergál (a konvergencia bizonyítását lásd Kroonenberg és Jan de Leeuw [1980, 77. oldal]).

18.3.1. A Tucker2-modell

Tucker háromszempontú faktormodellje (Tucker3-modell) három komponensmátrixot (faktormátrixot) tartalmaz. Az $\mathbf{A}$ mátrix a megfigyelt változóknak a latens változókra vonatkozó súlyait tartalmazza, a $\mathbf{B}$ mátrix a feltételeknek, azaz a harmadik szempontnak a komponensmátrixa, a $\mathbf{C}$ mátrix pedig (a Q-típusú elemzésnek megfelelően) a megfigyelési egyedeknek a komponens-mátrixa, vagyis az ideális egyedekre vonatkozó

súlya. A gyakorlatban sokszor a harmadik szempont esetében nincs értelme a főkomponensnek – például az idősor esetében –, így azt egységmátrixnak tekintve a következő egyszerűsített modellt írhatjuk fel (ezt nevezik Tucker2-modellnek):

$$x_{jk(i)} = \sum_{p}^{P} \sum_{m}^{M} a_{jp} c_{im} g_{pq(m)} + e_{jk(i)},$$

vagy mátrixjelölésekkel:

$$\mathbf{X}_{(I \times \overline{JK})} = \mathbf{C}_{(I \times M)} \cdot \mathbf{G}_{(M \times \overline{PK})} (\mathbf{A}'_{(\overline{P} \times J)} \otimes \mathbf{I}_{(K \times K)}) + \mathbf{E}_{(I \times \overline{JK})}.$$

A modell a harmadik szempont – a feltételek, idő stb. – minden egyes esetében a másik két szempont, a megfigyelések és a változók komponensei, és a közöttük meglévő kölcsönkapcsolat segítségével becsüli a mérési adatokat.

Ezt a modellt Kroonenberg „subjective metric model"-nek is nevezi (ez a modell azonos Carroll és Chang [1972] IDIOSCAL modelljével).

Ezt a modellt alkalmazhatjuk akkor is, ha nem a harmadik szempont (a különböző feltételekkel, időben történő mérés) az, amelyik nem sűríthető főkomponensekbe, hanem ha akármelyik szempont esetében nincs értelme vagy jelentősége a latens változónak (egyednek, ideális személynek stb). Például ha az „ideális személy" fogalma nem értelmes, vagyis nem léteznek ilyen latens egyedek, amelyek lineáris kombinációival az egyes megfigyelési egységek becsülhetők, kifejezhetők lennének, akkor a következőképpen írhatjuk a Tucker2-modellt:

$$x_{jk(i)} = \sum_{p}^{P} \sum_{q}^{Q} a_{jp} b_{kq} g_{pq(i)} + e_{jk(i)}.$$

Mátrixjelölésekkel:

$$\mathbf{X}_{(I \times \overline{JK})} = \mathbf{G}_{(I \times \overline{PQ})} \cdot (\mathbf{A}'_{(P \times J)} \otimes \mathbf{B}'_{(Q \times K)}) + \mathbf{E}_{I \times \overline{JK}}.$$

A paramétermátrixok becslése megegyezik a Tucker3-modell becslési eljárásával. A különbség csupán az, hogy vagy a **B** vagy a **C** mátrixot egységmátrixnak tekintjük, és így ugyanazt az algoritmust alkalmazhatjuk. Kronnenberg (1980) TUCKALS2 néven elkészítette ennek az egyszerűsített algoritmusnak a számítógépes programját is.

18.3.2. Az input adatok skálázása

Az input adatok skálázásán az adatoknak olyan transzformációját értjük, amely során az eredeti mérési adatokat meghatározott – leggyakrabban az adatoktól függő – mennyiségekkel – általában átlaggal, vagy más középértékkel, szórással, terjedelemmel stb. – transzformáljuk, és így új input adatokhoz jutunk.

A klasszikus főkomponens-elemzésnél a gyakorlatban legtöbbször az adatokat standardizáljuk, és a kovarianciamátrix helyett a korrelációmátrixot elemezzük. A standardizálás két lépésből áll; egyrészt a centrírozásból (minden elemből kivonjuk az adott változó átlagát, így 0 átlagú új elemeket kapunk), másrészt a standardizálásból (minden elemet elosztunk az adott változó szórásával, így 1 szórású új változót kapunk). Ezt a két lépést nevezhetjük általában a skálázás két típusának. Általánosan az átlag és a szórás helyett más konstans értékeket is használhatunk.

Az adatok skálázásánál két alapvető szabályt kell követni: a) azokat az átlagokat kell nullává transzformálni, amelyeknek vagy nincs önmagában értelme, vagy összehasonlíthatatlan a különböző szempontok esetében; b) azokat a szórásokat kell egységnyivé

tenni, amelyeknél a mértékegység önkényes és/vagy nem összehasonlítható a különböző szempontok esetében.

A háromszempontú adatmátrix esetében a centrírozásnak a következő típusait különböztethetjük meg:

1. Teljes centrírozás:

$$z_{jk(i)} = x_{jk(i)} - \overline{x}_{..(.)},$$

ahol az

$\overline{x}_{..(.)}$ a teljes átlagot jelöli (JKI elem átlaga).

2. A j-centrírozás:

$$z_{jk(i)} = x_{jk(i)} - \overline{x}_{j.(.)},$$

ahol

$\overline{x}_{j.(.)}$ egydimenziós marginális átlag (egyik szempont átlaga a másik két szempont értékei figyelembevételével (a j-edik változó átlaga a ki érték alapján).

Mindhárom szempont alapján elvégezhetjük ezt a centrírozást, így beszélhetünk k-centrírozásról és i-centrírozásról is.

3. A jk-centrírozás:

$$z_{jk(i)} = x_{jk(i)} - \overline{x}_{jk(.)},$$

ahol

$\overline{x}_{jk(.)}$ kétdimenziós marginális átlag (két szempont elemeinek minden lehetséges kombinációja a harmadik szempont elemei alapján átlagolva).

A három szempont mindegyikét párosíthatjuk, így a következő lehetőségek vannak:

– jk-centrírozás (ipsatíve)
– ij-centrírozás (abatíve)
– ki-centrírozás (normatív).

4. A jk, ik-centrírozás (performatíve) (dupla centrírozás)

$$z_{jk(i)} = x_{jk(i)} - \overline{x}_{jk(.)} - \overline{x}_{.k(i)} + \overline{x}_{.k(.)}.$$

Hasonlóan számíthatjuk az ij, jk, és a ki, ij-centrírozást is.

5. Hármas-centrírozás.

$$z_{jk(i)} = x_{jk(i)} - \overline{x}_{jk(.)} - \overline{x}_{.k(i)} - x_{j.(i)} + \\ + \overline{x}_{..(i)} + \overline{x}_{j.(.)} + \overline{x}_{.k(.)} - \overline{x}_{..(.)},$$

ahol

$\overline{x}_{j.(i)}$, $\overline{x}_{.k(i)}$, $\overline{x}_{jk(.)}$ a kétdimenziós marginális átlagok,
$\overline{x}_{..(i)}$, $\overline{x}_{j.(.)}$ $\overline{x}_{.k(.)}$ egy-dimenziós marginális átlagok, $\overline{x}_{..(.)}$ pedig a teljes átlag.

Az eredeti adatoktól és a kutatói értékeléstől függ, hogy a gyakorlatban melyik centrírozást választhatjuk, ha egyáltalán kell alkalmazni ezt az adattranszformációt. Mindenesetre a különböző centrírozások különböző eredményre vezetnek, így megfontoltan kell a gyakorlatban eljárni.

A standardizálás bonyolultabb eljárás, mivel ha egy szempont szerint standardizálunk, elronthatjuk a másik szempont szerinti centrírozást.

Harshman a PARAFAC1 nevű háromszempontú faktorelemző programjában alkalmazott egy iteratív standardizálási eljárást (aminek pontos algoritmusát nem közölték). Általában a standardizálást a centrírozással együtt célszerű alkalmazni, amit Kroonenberg (1983) normalizálásnak nevez.

18.3.3. Együttes ábrák

Az egyes szempontok főkomponenseit a komponensmátrixok (faktormátrixok) alapján értékelhetjük ki, míg a különböző szempontok komponensei közötti kapcsolatokat a belső struktúramátrix ($\mathbf{G}$) tartalmazza. Az egyes szempontok ideális egyedei (latens változók, komponensek) kapcsolatait nagyon jól szemlélteti, ha a szempontok komponensmátrixait minden lehetséges párosításban egy ún. közös euklideszi térbe vetítjük. Például, ha az ideális megfigyelési egységeket és a latens változókat („ideális" változók) a harmadik szempont, a megfigyelési feltételek ideális prototípusai minden kategóriája szerint egy közös térbe akarjuk helyezni, ahol az ideális személyek azokhoz a latens változókhoz esnek közel, amelyek jobban jellemzők rájuk (amelyeket jobban preferálnak), akkor a következőképpen járunk el. Legyen a $\mathbf{C}_m$ ($m = 1, \ldots, M$) a $\mathbf{C}$ mátrix oszlopvektora, az $\mathbf{A}_p$ az $\mathbf{A}$ mátrix oszlopvektora, a megfigyelési szempontok latens egyedei közötti távolságot az euklideszi távolság függvény szerint számoljuk, $d^2_{mp}(\mathbf{C}_m, \mathbf{A}_p)$.

A közelséget az $M \times P$ távolság négyzetösszegével mérjük. A $\mathbf{B}$ mátrix minden q ($q = 1, \ldots, Q$) elemére a következő távolságmátrixot számítjuk:

$$\mathbf{D}_q = \mathbf{C}\mathbf{G}_q\mathbf{A}' = \mathbf{C}\left(\mathbf{U}_q\mathbf{\Lambda}_q\mathbf{V}'_q\right)\mathbf{A}',$$

ahol $\mathbf{G}_q$ a belső struktúramátrix harmadik szempont (latens feltétel) szerinti ($M \times P$) típusú metszete.

A $\mathbf{D}_q$ mátrixot a következőképpen írhatjuk:

$$\mathbf{D}_q = \left(\frac{I}{J}\right)^{1/4}\left(\mathbf{C}\mathbf{U}_q\mathbf{\Lambda}_q^{1/2}\right)\left(\frac{J}{I}\right)^{1/4}\left(\mathbf{A}\mathbf{V}_q\mathbf{\Lambda}_q^{1/2}\right)' = \widetilde{\mathbf{C}}_q\widetilde{\mathbf{A}}'_q,$$

ahol $\widetilde{\mathbf{C}}_q = \left(\frac{I}{J}\right)^{1/4}\left(\mathbf{C}\mathbf{U}_q\mathbf{\Lambda}_q^{1/2}\right)$, $\widetilde{\mathbf{A}}_q = \left(\frac{J}{I}\right)^{1/4}\left(\mathbf{A}\mathbf{V}_q\mathbf{\Lambda}_q^{1/2}\right)$, $q = 1, \ldots, Q$.

Ha a $\mathbf{G}_q$ nem kvadratikus mátrix, akkor csak az első $\min(M, P)$ komponenst használhatjuk.

18.4. A Cartesius- és a Kronecker-féle szorzat

18.4.1. Cartesius-féle szorzat

Egy adathalmazt klasszifikáljunk két szempont szerint. Az egyik szempont legyen például a megfigyelési egységek (individumok, egyedek) mintája, i jelölje az i-edik megfigyelt személyt, és N_i a minta nagyságát. Legyen a másik szempont a változóknak (teszteknek) a halmaza, j jelölje a j-edik változót, a változók számát pedig jelölje N_j.

Ezeket nevezzük elemi szempontoknak.

A másodrendű kombinációs szempontot, amit a két elemi szempont Cartesius-féle szorzatának nevezünk és $\overline{ij}$-vel jelölünk, a következőképpen definiáljuk:

$$\overline{ij} = (i - 1)N_j + j.$$

Az i külső húrok minden eleménél a j-belső húrok elemei végigszaladnak.

A kombinációs szempont elemeinek a száma az elemi szempontok elemszámainak a szorzatával egyenlő:

$$N_{\overline{ij}} = N_i N_j .$$

A definíció szerint a kombinációs mód minden eleme az elemi szempontokból választott elem-párokkal függ össze. Illusztráljuk ezt egy egyszerű példával.

Legyen $N_i = 2$ és $N_j = 3$.

Az azonosítási számok (indexek)

Elemi szempontok		Kombinációs szempont
i	j	$\overline{ij}$
$1i$	$1j$	$1\overline{ij}$
$1i$	$2j$	$2\overline{ij}$
$1i$	$3j$	$3\overline{ij}$
$2i$	$1j$	$4\overline{ij}$
$2i$	$2j$	$5\overline{ij}$
$2i$	$3j$	$6\overline{ij}$

ahol $1i$ azt jelöli, hogy az i szempont első eleme, $2i$ jelöli az i szempont második elemét.

18.4.2. A Kronecker-féle szorzat

Legyen **A** egy m-dimenziós mátrix – $(m \times m)$ típusú – és **B** egy n-dimenziós – $(n \times n)$ típusú – mátrix.

Az **A** és **B** mátrixok Kronecker-féle szorzata egy mn-dimenziós mátrix, amit a következőképpen számítunk:

$$\mathbf{C} = \mathbf{A} \otimes \mathbf{B} = [a_{ij}\ \mathbf{B}] .$$

Ha **A** mátrix (2×2) típusú, akkor

$$\mathbf{C} = \begin{pmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} \end{pmatrix} .$$

A Kronecker-féle szorzatmátrix olyan szupermátrix, melynek szubmátrixai (partíciói) a **B** mátrix **A** mátrix megfelelő elemeivel súlyozott mátrixai. Ha **A** mátrix általános eleme $[a_{ij}]$, a **B** mátrix általános eleme $[b_{k\ell}]$ akkor a Cartesius-féle szorzat felhasználásával a **C** mátrix általános eleme:

$$[c_{\overline{ik}\ \overline{j\ell}}] = [a_{ij} b_{k\ell}].$$

A Kronecker-féle szorzatmátrix néhány tulajdonságát bizonyítás nélkül közöljük (lásd Bellman, 1970):

a) A Kronecker-féle szorzatmátrix transzponáltja egyenlő az eredeti mátrixok transzponáltjainak eredeti sorrendű Kronecker-féle szorzatával.

$$\mathbf{C}' = \mathbf{A}' \otimes \mathbf{B}'$$

b) Ha az **A** és **B** mátrixok kvadratikus és szimmetrikus mátrixok, akkor a Kronecker-féle szorzatmátrix is kvadratikus és szimmetrikus lesz.

c) A Kronecker-féle szorzatra érvényesek az alábbi tulajdonságok:

1. $$\mathbf{A} \otimes \mathbf{B} \otimes \mathbf{C} = (\mathbf{A} \otimes \mathbf{B}) \otimes \mathbf{C} = \mathbf{A} \otimes (\mathbf{B} \otimes \mathbf{C})$$

2. $$(\mathbf{A} + \mathbf{B}) \otimes (\mathbf{C} + \mathbf{D}) = \mathbf{A} \otimes \mathbf{C} + \mathbf{A} \otimes \mathbf{D} + \mathbf{B} \otimes \mathbf{C} + \mathbf{B} \otimes \mathbf{D}$$

3. $$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC}) \otimes (\mathbf{BD})$$

d) Értelmezhetjük a Kronecker-féle mátrix hatványát:

$$\mathbf{A}^{[2]} = \mathbf{A} \otimes \mathbf{A},$$

$$\mathbf{A}^{[k+1]} = \mathbf{A} \otimes \mathbf{A}^{[k]}.$$

Ha $\mathbf{A}$ és $\mathbf{B}$ nem kommutatív (kommutatív, ha $\mathbf{AB} = \mathbf{BA}$), akkor

$$(\mathbf{A}\ \mathbf{B})^k \neq \mathbf{A}^k\mathbf{B}^k \quad \text{és soha, ha} \quad k = 2.$$

Azonban bármely $\mathbf{A}$ és $\mathbf{B}$ -re igaz, hogy

$$(\mathbf{AB})^{[k]} = \mathbf{A}^{[k]}\mathbf{B}^{[k]}.$$

e) Ha $\mathbf{A}$ és $\mathbf{B}$ diagonális mátrixok, akkor Kronecker-féle szorzatuk is diagonális mátrix lesz, ahol a diagonális elemek $\mathbf{A}$ és $\mathbf{B}$ diagonális elemeinek páronkénti szorzata.

f) Az $\mathbf{A} \otimes \mathbf{B}$ sajátértékei $\lambda_i \mu_j$, ahol λ_i az $\mathbf{A}$ mátrix, μ_j pedig a $\mathbf{B}$ mátrix sajátértéke. Az $\mathbf{A} \otimes \mathbf{B}$ sajátvektorai

$$\mathbf{Z}_{ij} = \begin{pmatrix} x_1^i \mathbf{y}^j \\ x_2^i \mathbf{y}^j \\ \vdots \\ x_n^i \mathbf{y}^j \end{pmatrix},$$

ahol x_k^i $(k = 1, \ldots n)$ az $\mathbf{A}$ mátrix i-edik sajátvektorának elemei, $\mathbf{y}^j$ a $\mathbf{B}$ mátrix j-edik sajátvektora.

g) Ha az $\mathbf{A}$ és $\mathbf{B}$ mátrixnak létezik az inverze, vagyis $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$, és $\mathbf{B}^{-1}\mathbf{B} = \mathbf{I}$, akkor

$$\left(\mathbf{A}^{-1} \otimes \mathbf{B}^{-1}\right)(\mathbf{A} \otimes \mathbf{B}) = \mathbf{I},$$

ahol $\mathbf{I}$ az egységmátrixot jelöli.

h) Ha $\mathbf{A}$ és $\mathbf{B}$ ortonormált mátrixok, akkor Kronecker-féle szorzatuk is ortonormált mátrix.

i)

$$tr(\mathbf{A} \otimes \mathbf{B}) = (tr\mathbf{A})(tr\mathbf{B}),$$

ahol $tr\mathbf{A}$ az $\mathbf{A}$ mátrix nyoma (trace) (diagonális elemeinek az összege).

18.5. A PARAFAC-modell

A PARAFAC-modell bemutatásához induljunk ki a hagyományos kétdimenziós adatmátrixot elemző faktormodellből. Eszerint a megfigyelt adatokat a mögöttük lévő latens faktorokkal a következőképpen fejezhetjük ki:

$$x_{j(i)} = \sum_m^M a_{jm} f_{(i)m} + e_{j(i)},$$

vagy

$$x_{ji} = \sum_{m}^{M} a_{jm} f_{im} + e_{ji},$$

és mátrixaritmetikai formában

$$\mathbf{X} = \mathbf{A}\mathbf{F}' + \mathbf{E},$$

ahol

X a megfigyelési adatok ($J \times I$) típusú mátrixa, ahol a sorokban a változók vannak ($j = 1, \ldots, J$), az oszlopokban pedig a megfigyelési egységek, a vizsgált objektumok ($i = 1, \ldots, I$) (ez eltér a szokásos jelöléstől, hiszen hagyományosan a változók vannak egy oszlopban és a megfigyelések a sorokban!),

A a faktorsúlyok ($J \times M$) típusú mátrixa, a_{jm} jelöli a j-edik változó m-edik faktorra vonatkozó súlyát (faktorsúly, factor loading),

F a faktorértékek ($I \times M$) típusú mátrix, f_{im} az i-edik megfigyelésnek az m-edik faktorra vonatkozó értéke vagy súlya (factor score),

E a véletlen hibakomponensek mátrixa (specifikus faktorok hatásait fejezi ki).

A fenti faktormodell jelentheti akár a főkomponens-elemzést, akár a faktorelemzést, az **E** mátrixra vonatkozó feltételezésektől függően.

Kruskal (1978, 1981) a faktormodell skaláris formáját ($x_{ji} = \sum a_{jm} f_{im} + e_{ji}$) *bilineáris modellnek* nevezte, mivel a modell strukturális (nem véletlen) része bilineáris kifejezés (bilineáris azért, mert az a_{jm} együtthatók lineáris függvénye, ha az f_{im} együtthatókat konstansnak vagy adottnak tekintjük és vice versa).

A PARAFAC ezt a modellt általánosítja, R. B. Catell (1944), párhuzamos arányos profilok (parallel proportional profiles) alapján. Eszerint ha különböző helyzetekben, állapotokban vagy időpontokban van mérésünk ugyanazon változókról (ugyanazon megfigyelési egységekre), akkor a faktorok relatív pontszámának változását elemezzük. Így ha egy faktor hatása pl. 20%-kal változik, mondjuk növekszik, akkor a faktorsúlyok 20%-kal lesznek nagyobbak ($a_{km} a_{jm} = 1{,}2 a_{jm}$).

Ha a különböző helyzeteket, állapotokat vagy különböző időpontokat a harmadik szempontnak tekintjük, akkor az adatmátrixunk háromdimenziós lesz, általános eleme: $x_{jk(i)}$, ahol a k index jelöli a harmadik szempontot ($k = 1, 2, \ldots, K$).

A PARAFAC általánosítása a kétdimenziós faktor-modellnek, és a következő formát ölti:

$$x_{jk(i)} = \sum_{m}^{M} a_{jm} b_{km} c_{im} + e_{jk(i)},$$

ahol

$x_{jk(i)}$ az i-edik megfigyelési egység j-edik manifeszt változóra vonatkozó megfigyelt értéke a k-adik feltétel (állapot, helyzet) esetében,

a_{jm} a j-edik változó m-edik faktorra vonatkozó súlya,

c_{im} jelöli a kétfaktoros megoldás faktorértékét (f_{im}), azaz az i-edik megfigyelés m-edik faktorra vonatkozó értékét,

b_{km} a k-adik feltétel, állapot faktorsúlya az m-edik faktorra,

$e_{jk(i)}$ a hibakomponens.

Ezt a modellt Kruskal (1981) trilineáris modellnek hívja, mivel most a paramétereknek, faktorsúlyoknak három különböző halmaza van, és a modell lineáris mindegyik paraméterhalmazra, ha a másik kettő értékeit rögzítjük.

Harshman (1984) a c_{im} faktorértékeket (factor score) faktorsúlyoknak is nevezi, mivel az nem más, mint az i-edik megfigyelési egység súlya vagy fontossága az m-edik faktornál.

Ha mátrixjelölést használva bevezetjük az $\mathbf{X}_k$ mátrixot, ami a $(J \times K \times I)$ típusú háromdimenziós mátrix k-adik szelete $(J \times I)$ típusú), akkor az aritmetikai egyenletünket a következőképpen írhatjuk:

$$\mathbf{X}_k = \mathbf{A}\mathbf{D}_k \mathbf{C}' + \mathbf{E}_k,$$

ahol

- $\mathbf{A}$ $(J \times M)$ típusú faktormátrix, a változók és faktorok kapcsolódását írja le, az A szempont esetében,
- $\mathbf{C}$ $(I \times M)$ típusú faktormátrix, a megfigyelési egységek és a faktorok kapcsolatait tartalmazza,
- $\mathbf{D}_k$ $(M \times M)$ típusú diagonális mátrix, diagonális elemei a $\mathbf{B}$ $(K \times M)$ típusú faktormátrix k-adik sorának elemei,
- $\mathbf{E}_k$ $(J \times I)$ típusú, a hibakomponensek-mátrixa a k-adik feltétel (állapot, helyzet, időpont) esetén.

Láthatjuk, hogy a hagyományos faktorelemzéssel az $\mathbf{A}$ és $\mathbf{C}$, a két szempont (A szempont és C szempont) faktorsúlyait különböző névvel láttuk el. Az A szempont súlyait faktorsúlyoknak, míg a C szempont súlyait faktorértékeknek neveztük.

A faktorsúlyokat nevezhetjük standardizált regressziós együtthatóknak – béta súlyoknak vagy korrelációs együtthatóknak – független faktorok esetén, és értelmezhetjük íly módon is. A C szempont súlyait z értékeknek (z-scores) leírva az egyes faktorok hozzájárulását mutatják az egyes megfigyelési egységekhez. Az, hogy melyik értelmezés a helyes, függ a mérési skálától, illetve attól, milyen kontextuális keretben akarjuk értelmezni őket. A PARAFAC esetében, mivel az egyes szempontok felcserélhetők, általánosan úgy értelmezzük őket, mint az egyes faktorok hozzájárulásait, súlyait az adatok becsléséhez. Ez hasonlít a regressziós együtthatók értelmezéséhez, amikor nem tételezzük fel sem a függő, sem a független változókról, hogy standarizáltak. A PARAFAC inputjának megfelelő újraskálázásaival elérhető, hogy az output mátrixokat a kétdimenziós faktorelemzéshez hasonlóan értelmezzük. Például, ha azt akarjuk, hogy az A szempont szerinti faktormátrixot faktorsúlyoknak, a B és C szempont szerintit, komponens- vagy faktorértéknek nevezzük, a következőképpen kell eljárni: centrírozni kell az adatokat a B és C szempont szerint, standardizálni az A szerint (a változók varianciáinak 1-et kell adni). Az eredményeket szintén standardizálni kell A, B és C szempont szerint úgy, hogy a négyzetes átlag 1 legyen.

(Pl. a C szempont szerint az i-edik elem c_{im} a négyzetes átlag négyzetgyökét adja, ami megegyezik a szórással, ha a B szempont szerint az adatok centrírozottak). (A centrírozásról lásd részletesebben a 18.3.2. Az input adatok skálázása című részt.)

18.6. A PARAFAC1-modell alkalmazása kovariancia-adatokra (PARAFAC2)

A paramétermátrixok elemeire, a faktorsúlyokra és a faktorértékekre közvetlen becslést kapunk, ha a faktormodellt az eredeti, megfigyelt adatokhoz illesztjük. Hagyományosan azonban a korrelációk vagy kovarianciák elemzésére koncentrálunk. A korrelációs- ill. kovarianciamátrixot az eredeti adatmátrixból származtatjuk, az ezekre felírt faktormodell paramétereinek becslését Kruskal (1978) „közvetett becslésnek" nevezte. Kétdimenziós modellek esetében a legkisebb négyzetek módszere azonos becslést ad. Háromdimenziós esetben azonban ez nem igaz. A PARAFAC direkt és indirekt becslése különbözik, mégpedig két szinten: statisztikai szinten, a paraméterek és a reziduálisok pontos értékét értve alatta, és sturktúrális szinten, ami az adatok konfigurációját, struktúráját jelzi, amit a megoldás közvetlenül meg tud mutatni.

Az eddig tárgyalt modellt Harshman (1972) PARAFAC1-modellnek nevezi, megkülönböztetve az ebből származtatott, általánosított modellektől.

Ha a PARAFAC1-modellt az eredeti adatokból számított kovarianciákra alkalmazzuk, akkor a PARAFAC2-modellről beszélünk. Induljunk ki az $\mathbf{X}_n$ mátrixra felírt PARAFAC1-modellből, és számítsuk ki a változók egymás közötti keresztszorzatait (cross-products):

$$\mathbf{\Sigma}_k = \mathbf{X}_k \mathbf{X}_k'.$$

Helyettesítsük $\mathbf{X}_k$ helyébe a PARAFAC1-modellt:

$$\mathbf{\Sigma}_k = (\mathbf{A}\mathbf{D}_k\mathbf{C}' + \mathbf{E}_k)(\mathbf{A}\,\mathbf{D}_k\mathbf{C}' + \mathbf{E}_k)',$$

és végezzük el a mátrixszorzást:

$$\mathbf{\Sigma}_k = (\mathbf{A}\,\mathbf{D}_k\,\mathbf{C}')(\mathbf{A}\mathbf{D}_k\,\mathbf{C}')' + \mathbf{E}_k\mathbf{E}_k',$$

figyelembe véve, hogy a hibakomponens független a szisztematikus komponenstől. A mátrix-aritmetikai eljárást tovább folytatva kapjuk, hogy

$$\mathbf{\Sigma}_k = \mathbf{A}\mathbf{D}_k(\mathbf{C}'\mathbf{C})\mathbf{D}_k\,\mathbf{A}' + \mathbf{E}_k\mathbf{E}_k'.$$

Vezessük be a $\mathbf{W} = \mathbf{C}'\,\mathbf{C}$ jelölést, így

$$\mathbf{\Sigma}_k = \mathbf{A}\mathbf{D}_k\,\mathbf{W}\mathbf{D}_k\mathbf{A}' + \mathbf{E}_k\mathbf{E}_k'.$$

Ezt az egyenletet nevezzük PARAFAC2-modellnek (Harshman, 1972).

A modell általánosan mutatja a keresztszorzatokat. Azonban ha az $\mathbf{X}_k$ mátrix a változók átlaguktól való eltéréseit tartalmazza (így minden mátrixban (k) minden változó átlaga nulla), akkor a $\mathbf{\Sigma}_k$ az eltérés-négyzetösszegeket tartalmazza, és ha elosztjuk minden elemét a megfigyelések számával, akkor $\mathbf{\Sigma}_k$ a változók variancia-kovarianciamátrixát jelenti a k-adik feltétel (állapot, idő stb.) esetén. A modellben a $\mathbf{W}$ mátrix új komponens; a faktorok közötti általános kapcsolatokat írja le, ami a faktorok közötti kovarianciamátrixot jelenti. Amennyiben a $\mathbf{C}$ mátrix oszlopvektorainak a négyzetösszege egyenlő eggyel, akkor a $\mathbf{W}$ a faktorok korrelációs mátrixát adja. Ha feltételezzük, hogy a faktorok ortogonálisak (függetlenek) a C szempont szerint, akkor $\mathbf{W} = \mathbf{I}$, és

$$\mathbf{\Sigma}_k = \mathbf{A}\,\mathbf{D}_k^2\mathbf{A}' + \mathbf{E}_k\,\mathbf{E}_k'.$$

Ezt a modellt Harshman (1984. 137 p.) a kovarianciákra (vagy más keresztszorzat adatokra) felírt PARAFAC1-modellnek nevezi. Ebben az esetben az indirekt becslési eljárás nagyon hasonló vagy majdnem megegyező eredményt ad a faktormátrixra ($\mathbf{A}$), mint amit a megfigyelési adatokra közvetlenül illesztett modell ad. Az eltérés onnan

adódik, hogy a két eljárás nem ugyanazt a függvényt optimalizálja; az egyik a legkisebb négyzetek elvét a kovarianciára, a másik a közvetlen megfigyelt adatokra alkalmazza. Az indirekt becslésnél információs veszteségünk is van, mivel a faktorértékek (a C szempont faktorsúlyai) nem szerepelnek a modellben, bár később, a modell becslése után regressziós eljárással ki tudjuk őket számítani. Ha a faktorok nem ortogonálisak, akkor az indirekt PARAFAC1-modell nem megfelelő, a becslések torzítottak lesznek, bár ha az eltérés az ortogonálistól nem túl nagy, akkor az eredmények még elfogadhatók, egyébként pedig nem értelmezhetők.

18.7. A PARAFAC1-modell különböző minták esetében

Tekintsük azt az esetet, amikor nem ugyanazon populációból származó mintán elemezzük az adott változóhalmazt különböző helyzetekben, feltételekkel, hanem ugyanazon változóhalmazt vizsgáljuk több, különböző populációból vett mintán.

Jelölje $\mathbf{X}_k$ a $(J \times I_k)$ típusú adatmátrixot, J változó mérési adatait az I_k elemszámú k-adik mintában. Feltételezve, hogy minden mintában azonos számú és struktúrájú faktor létezik, ezeknek csak a súlya, relatív fontossága változik a különböző populációkban (mintákban), a mért adatokban rejlő struktúrát a következő modell fejezheti ki:

$$\mathbf{X}_k = \mathbf{A}\mathbf{D}_k\,\mathbf{C}_k' + \mathbf{E}_k,$$

ahol

- $\mathbf{A}$ $(J \times M)$ típusú faktorsúlyok mátrixa, azonos minden mintában,
- $\mathbf{D}_k$ $(M \times M)$ típusú diagonális mátrix, M faktor relatív fontosságát adja a k-adik mintában,
- $\mathbf{C}_k$ $(I_k \times M)$ típusú mátrix, a megfigyelési egységek súlyait vagy másképpen a faktorok értékeit tartalmazza a k-adik mintában,
- $\mathbf{E}_k$ a hibakomponensek $(J \times I_k)$ típusú mátrixa a k-adik mintában.

Ezt a modellt nem tudjuk közvetlenül becsülni a PARAFAC-eljárással. A közvetett becslő eljárást viszont alkalmazhatjuk ebben az esetben is. Az $\mathbf{X}_k$ mátrix helyébe a fenti egyenletet helyettesítve az $(\mathbf{X}_k\mathbf{X}_k')$ keresztszorzat mátrixot a korábban már bemutatott eljárással kapjuk:

$$\mathbf{\Sigma}_k = \mathbf{A}\,\mathbf{D}_k(\mathbf{C}_k'\,\mathbf{C}_k)\mathbf{D}_k\mathbf{A}' + \mathbf{E}_k\mathbf{E}_k'.$$

A $\mathbf{W}_k = \mathbf{C}_k'\mathbf{C}_k$ helyettesítést alkalmazva:

$$\mathbf{\Sigma}_k = \mathbf{A}\mathbf{D}_k\,\mathbf{W}_k\mathbf{D}_k\mathbf{A}' + \mathbf{E}_k\mathbf{E}_k'.$$

Ezt nevezzük a faktorelemzés általános egyenletének a k-adik mintában. Amennyiben az adatmátrix soronként centrírozott, akkor $\mathbf{\Sigma}_k$ az eltérés-négyzetösszegeket (keresztszorzat-eltéréseket) jelöli, és ha $\mathbf{\Sigma}_k$ elemeit elosztjuk I_k-val (a minta elemszámával), akkor $\mathbf{\Sigma}_k$ a variancia-kovarianciamátrixot jelöli. A $\mathbf{W}_k$-t standardizálhatjuk, hogy elemei a faktorok közötti korrelációkat (koszinuszokat) adják. Ezt az esetet nevezzük az obligát faktorelemzés általánosításának. Ez a modell általánosabb, mint a PARAFAC2-modell. Ez a modell hasonlít Carroll és Chang IDIOSCAL-modelljéhez (1972), és ezt nevezi Kroonenberg és de Leeuw (1980) Tucker2 modellnek.

Könnyen beláthatjuk, hogy ha $\mathbf{W}_k = \mathbf{W}$, akkor a PARAFAC2-modellt kapjuk, és ha $\mathbf{W}_k = \mathbf{I}$, vagyis feltételezzük a faktorok ortogonalitását, akkor a PARAFAC1-modellt kapjuk indirekt becslés esetére.

Meg kell jegyezni, hogy ha a kovarianciamátrixot ($\mathbf{\Sigma}_k$) átalakítjuk korrelációs mátrixszá (ezt megtehetjük úgy, hogy egy diagonális mátrixszal jobbról, balról beszorozzuk a diagonális mátrix elemei a $\mathbf{\Sigma}_k$ mátrix elemeiből vont négyzetgyök reciprokai, jelöljük $\dot{\mathbf{D}}_k$-val), akkor a változók és faktorok kapcsolatait tartalmazó $\mathbf{A}$ mátrix sorai a különböző feltételeknél (B szempont) transzformálódnak: $\dot{\mathbf{D}}_k\mathbf{A}$, így ez elrontja az $\mathbf{A}$ faktormátrix összehasonlíthatóságát a B szempont különböző kategóriáiban:

$$\dot{\mathbf{\Sigma}}_k = \dot{\mathbf{D}}_k \mathbf{A}\mathbf{D}_k W \mathbf{D}_k \mathbf{A}'\dot{\mathbf{D}}_k + \dot{\mathbf{D}}_k\mathbf{E}_k \mathbf{E}_k'\dot{\mathbf{D}}_k.$$

A PARAFAC ezért nem használható korrelációs mátrixok elemzésére. Ez hasonlóan igaz Tucker modelljeire, de Jöreskog több-populációs faktorelemzésére is. A PARAFAC-program biztosít egy opciót a változók skálakülönbözőségének kiküszöbölésére. Eszerint az egyes változókat nem a B-szempont kategóriáiban kell külön-külön standardizálni, hanem a B szempont mindegyik kategóriájában együtt. A $\dot{\mathbf{D}}_k$ mátrix j-edik diagonális eleme:

$$\dot{d}_{jj} = \left(\frac{1}{K} \sum_{k=1}^{K} \omega_{jj_k} \right)^{-\frac{1}{2}}.$$

Ezt az opciót „equal average diagonal standardization”-nak nevezik (Harshman, 1984).

18.8. Főkomponenselemzés versus faktorelemzés

A közvetett becslő modell egyik előnye, hogy mind a főkomponenst, mind a faktormodellt illeszthetjük, szemben a közvetlen, direkt becsléssel, ahol csak a főkomponensmodell illeszthető. Harris (1975) megállapítása szerint adatelemző szempontból a két különböző faktorbecslés általában nagyon hasonló értékeket ad a faktorsúlyokra.

Elméleti szempontból érdemes megvizsgálni a hibakomponens kovarianciamátrixát ($\mathbf{E}_k\mathbf{E}_k'$). A kovarianciamátrix diagonális feletti elemei kicsik, véletlen ingadozást mutatnak a 0 körül. A diagonális elemek viszont szisztematikusan nagyobbak és pozitív értékűek, mivel a hibakomponensek varianciái. Emiatt célszerűnek látszik a megfigyelt változók kovarianciamátrix diagonális elemeinek elhagyása, mivel a diagonális elemek nagyobb hibakomponenst tartalmaznak, így szisztematikusan felfelé torzítanak a „valódi” értéktől. Általában a különböző faktorelemző eljárások nem hagyják el a diagonális elemet, hanem helyettesítik a változók kommunalitásával (a faktorok által megmagyarázott varianciával). A kommunalitások kezdeti becslésére többféle módszer létezik. Becsülhetjük a többtényezős korreláció négyzetével (minden változónak az összes többi változóra vonatkozó regressziós egyenlete alapján), vagy más eljárással. Az iterációs lépésekben azután behelyettesítjük a faktorok által magyarázott résszel, egészen addig, amíg az eljárás nem konvergál. A PARAFAC számítógépes program Harman és Jones (1966) MINRES-eljárásával megegyező algoritmust alkalmazva a diagonális feletti elemeket becsli.

Empirikus vizsgálatok (Hann, 1981) azt mutatják, hogy nincs nagy különbség a faktormátrix **A** faktorsúlyainak becslésénél, akár elhagyjuk vagy megtartjuk a diagonális elemeket a becslési eljárás során.

18.9. A PARAFAC valódi tengely tulajdonsága

A faktorelemzés, hasonlóan a sokdimenziós skálázáshoz, alapvetően két információval szolgál: a) a pontok konfigurációjával alacsonyabb dimenziószámú térben; b) a tengelyekkel, amelyek kifeszítik a redukált dimenziószámú teret.

A pontok konfigurációja a pontok reprezentálta változók (egyedek stb.) megfigyelt kapcsolatrendszerét fejezi ki egy egyszerűsített, redukált latens térben. Ez alapvetően nem ad új információt, csak esetleg prezentálhatóvá teszi (két- háromdimenziós térben) a kapcsolatrendszert, megkönnyíti a kapcsolatrendszer egészének áttekintését. Ami lényegesen új információt nyújt, az a tengelyek meghatározása. A valódi, helyes irányba mutató tengelyekre vetített pontok vetületei fejezik ki a megfigyelt változók és a mögöttes, latens változók (faktorok) kapcsolatait, és ezek a faktorok jelenthetik az új információt, amelyek felelősek, magyarázzák a megfigyelt folyamatokat. Ha a tengelyek iránya önkényes, vagy valamilyen önkényes – nem belső – kritériumon alapul, akkor a faktorok értelmezése – ahogyan a klasszikus faktorelemzésnél – bizonytalan, kétséges lehet.

A PARAFAC-modellt Harshman éppen azért fejlesztette ki, hogy a faktorok valódi irányát elméleti és empirikus alapon biztosítsa. Vannak, akik a faktorok irányát (és a faktorok rotációját) nem tekintik lényegesnek (pl. Thurstone 1947, 332), mivel szerintük mindegyik irány kifejezheti egy lényeges dimenzióját a bonyolult vizsgált jelenségnek. Ez a felfogás lehet helyes, ha csak a változók sűrítése, redukálása a célunk. Ha a megfigyelt változók konfigurációjának, kapcsolatrendszerének a magyarázata a feladatunk, akkor a különböző rotált megoldások nem lehetnek egyenlően érvényesek.

Jennrich (1970), Harshman (1972), Kruskal (1977) bevezették az „adekvát" adat fogalmát, mely szerint az alternatív megoldás csak azt jelentheti, hogy a megoldás paramétermátrixának oszlopait permutáljuk, átrendezzük, ill. az oszlopvektorokat egy konstans értékkel beszorozzuk. Legyen egy alternatív megoldás a következő:

$$\mathbf{X}_k = \overset{*}{\mathbf{A}}_k \overset{*}{\mathbf{D}}_k \overset{*}{\mathbf{C}}_k' + \mathbf{E}_k,$$

ahol

$\overset{*}{\mathbf{A}}$ a faktormátrix (**A**) alternatív megoldása,

$\overset{*}{\mathbf{D}}_k$ a $\mathbf{D}_k$ mátrix alternatív megoldása,

$\overset{*}{\mathbf{C}}_k$ a $\mathbf{C}_k$ mátrix alternatív megoldása.

A valódi tengely megoldás biztosítja, hogy $\overset{*}{\mathbf{A}}$ annyiban különbözik csak A-tól, hogy oszlopvektorait átrendezzük, permutáljuk (ami azt jelenti, hogy beszorozzuk egy **P** permutáló mátrixszal), és/vagy az oszlopvektorait beszorozzuk egy konstans értékkel (ez azt jelenti, hogy beszorozzuk az **A** mátrixot egy $\overset{\circ}{\mathbf{D}}_k$ diagonális mátrixszal, hasonlóan transzformálhatjuk a **C** és a Σ mátrixokat). Így az **A**, **C** és $\mathbf{D}_k$ mátrixokra a következő

azonosságok igazak:

$$\overset{*}{\mathbf{A}} = \mathbf{A}\ \overset{\circ}{\mathbf{D}}_a\ \mathbf{P},$$

$$\overset{*}{\mathbf{C}} = \mathbf{C}\ \overset{\circ}{\mathbf{D}}_c\ \mathbf{P},$$

$$\overset{*}{\mathbf{D}}_k = \mathbf{P}'\,\mathbf{D}_k\ \overset{\circ}{\mathbf{D}}_\Sigma\ \mathbf{P}.$$

Mivel $\overset{*}{\mathbf{D}}_k$ mátrix diagonális eleme a $\boldsymbol{\Sigma}$ mátrix sorvektora, így írhatjuk, hogy

$$\overset{*}{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}\ \overset{\circ}{\mathbf{D}}_k\ \mathbf{P}.$$

A skálázások hatása semlegesíti egymást, így

$$\overset{\circ}{\mathbf{D}}_a\ \overset{\circ}{\mathbf{D}}_c\ \overset{\circ}{\mathbf{D}}_\Sigma = \mathbf{I}.$$

Ezek a skálázások nem befolyásolják az eredmények értelmezését. Geometriailag a $\mathbf{P}$ permutációs mátrix csupán átrendezi a tengelyek sorrendjét, a $\mathbf{D}_a$ mátrix pedig a tengely fontosságát rendezi át. Ezek a transzformációk nem befolyásolják a tengely fontosságát, a három tér egyikében sem.

A valódi tengely megoldás éppen azt biztosítja, hogy a tengelyeket ne lehessen különbözőképpen értelmezni.

A faktorok rotációjának problémáját nézzük meg először a kétdimenziós faktorelemzés esetében. A faktormodell:

$$\mathbf{X} = \mathbf{A}\mathbf{C}' + \mathbf{E}.$$

Transzformáljuk az $\mathbf{A}$ faktorsúlyok mátrixát tetszőleges nemszinguláris lineáris transzformációval ($\mathbf{T}$) és kompenzációként a faktorértékek mátrixával $\mathbf{T}^{-1}$-vel:

$$\overset{*}{\mathbf{A}} = \mathbf{A}\,\mathbf{T}, \qquad \overset{*}{\mathbf{C}} = \mathbf{C}\mathbf{T}'^{-1}.$$

Ha P számú faktorunk van, akkor $\mathbf{A}$ $(J \times P)$ típusú, oszlopvektorai a változóknak a P számú faktor-tengelyre vonatkozó vetületei. $\mathbf{I}$ nemszinguláris $(P \times P)$ típusú mátrix, oszlopvektorai az $\mathbf{A}$ eredeti tengelyeknek az új $\overset{*}{\mathbf{A}}$ tengelyekre vonatkozó vetületei, így $\overset{*}{\mathbf{A}}$ a változóknak az új faktortengelyekre vonatkozó vetületeit tartalmazza.

Könnyen belátható, hogy

$$\mathbf{X} = \overset{*}{\mathbf{A}}\,\overset{*}{\mathbf{C}}{}' + \mathbf{E}$$

ekvivalens az eredeti modellel:

$$\mathbf{X} = (\mathbf{A}\,\mathbf{T})\left(\mathbf{C}\mathbf{T}'^{-1}\right)' + \mathbf{E} = \mathbf{A}(\mathbf{T}\,\mathbf{T}^{-1})\mathbf{C}' + \mathbf{E} = \mathbf{A}\mathbf{C}' + \mathbf{E}.$$

Nézzük most meg a PARAFAC-modell esetében a fenti transzformáció eredményét:

$$\mathbf{X}_k = (\mathbf{A}\mathbf{T})\ \overset{*}{\mathbf{D}}_k\ \left(\mathbf{C}\,\mathbf{T}'^{-1}\right)' + \mathbf{E}_k,$$

amit egyszerűsíthetünk:

$$\mathbf{X}_k = \mathbf{A}\left(\mathbf{T}\ \overset{*}{\mathbf{D}}_k\ \mathbf{T}^{-1}\right)\mathbf{C}' + \mathbf{E}_k.$$

Vezessük be a $\mathbf{H}_k = \mathbf{T}\ \overset{*}{\mathbf{D}}_k\ \mathbf{T}^{-1}$ jelölést:

$$\mathbf{X}_k = \mathbf{A}\mathbf{H}_k\,\mathbf{C}' + \mathbf{E}_k.$$

A PARAFAC-modell szerint a $\mathbf{H}_k$ mátrix diagonális, ezért a $\mathbf{T}$ transzformációs mátrix csak olyan lehet, hogy $\mathbf{H}_k$ diagonális maradjon és egyenlő legyen $\mathbf{D}_k$ mátrixszal. Be lehet látni, hogy $\mathbf{T}$ mátrixnak diagonális mátrixnak vagy permutációs mátrixnak kell lennie (Harshman, 1972). Ez összefügg azzal, amit korábban már tárgyaltunk, $\mathbf{A}$ és $\overset{*}{\mathbf{A}}$ kapcsolatával, ahol $\mathbf{T} = \overset{\circ}{\mathbf{D}}_a\ \mathbf{P}$. Ha a $\mathbf{T}$ mátrix kielégíti ezt a feltételt, akkor alternatív modellt kaphatunk:

$$\begin{aligned}\overset{*}{\mathbf{A}}\,\overset{*}{\mathbf{D}}_k\,\overset{*}{\mathbf{C}}{}' &= (\mathbf{A}\ \overset{\circ}{\mathbf{D}}_a\ \mathbf{P})(\mathbf{P}'\mathbf{D}_k\ \overset{\circ}{\mathbf{D}}_k\ \mathbf{P})(\mathbf{C}\ \overset{\circ}{\mathbf{D}}_c\ \mathbf{P})' \\ &= \mathbf{A}\ \overset{\circ}{\mathbf{D}}_a\ (\mathbf{P}\,\mathbf{P}')(\mathbf{D}_k\ \overset{\circ}{\mathbf{D}}_\Sigma)(\mathbf{P}\,\mathbf{P}')(\overset{\circ}{\mathbf{D}}_c\ \mathbf{C}').\end{aligned}$$

Mivel $\mathbf{P}\mathbf{P}' = \mathbf{I}$ bármely permutációs mátrix esetén, valamint $\overset{\circ}{\mathbf{D}}_a\ \mathbf{D}_k = \mathbf{D}_k\ \overset{\circ}{\mathbf{D}}_a$, mert diagonális mátrixok szorzata kommutatív:

$$\mathbf{X}_k = \overset{*}{\mathbf{A}}\,\overset{*}{\mathbf{D}}_k\,\overset{*}{\mathbf{C}}{}' + \mathbf{E}_k = \mathbf{A}\mathbf{D}_k(\overset{\circ}{\mathbf{D}}_a\,\overset{\circ}{\mathbf{D}}_k\overset{\circ}{\mathbf{D}}_c)\mathbf{C}' + \mathbf{E}_k.$$

18.10. A PARAFAC- és TUCKER3-modellek összehasonlítása

Tucker (1963, 64, 66) volt az első, aki a faktormodellt kiterjesztette háromszempontú adatmátrixok elemzésére, és modellje mind a mai napig a referenciapontot jelenti, amihez a többi modellt hasonlítani kell.

Kroonenberg és de Leeuw megkülönböztette Tucker modelljének két változatát: a) az eredeti modellt, amelyben mindhárom szempont szerint redukáljuk a dimenzionalitást (ezt nevezzük Tucker3-modellnek); b) változat esetén a harmadik szempont (B szempont, feltételek, állapot, időpontok stb.) szerint nem csökkentjük a dimenziók számát (a $\mathbf{B}$ mátrix egységmátrix lesz), és csak a másik két szempont szerint redukáljuk a dimenzionalitást (ezt nevezzük Tucker2-modellnek).

A Tucker3-modell skaláris formában:

$$x_{jk(i)} = \sum_{p}^{P}\sum_{q}^{Q}\sum_{m}^{M} a_{jp}b_{kq}c_{im}g_{pq(m)} + e_{jk(i)}.$$

A PARAFAC-modell:

$$x_{jk(i)} = \sum_{m}^{M} a_{jm}b_{km}c_{im} + e_{jk(i)}\,.$$

A két modellben az a, b és c paramétereknek azonos a jelentése: faktorsúlyok vagy együtthatók az A, B és C szempont szerint. Tucker a háromindexű paramétert $g_{pq(m)}$ a belső struktúramátrix (core matrix) együtthatójának nevezte.

Tucker modelljében háromszoros szumma szerepel, mivel ebben a modellben a faktorok (latens változók) száma különböző lehet az egyes szempontokban.

A két modell alapvetően megegyezik a következőkben: egyrészt mindkettő közvetlenül illeszti a modellt a megfigyelt adatokhoz, nem közvetetten a kovarianciamátrixhoz.

Másrészt a faktorsúlyok mindhárom szempontban hasonlóan értelmezhetők; a faktorok fontosságát fejezik ki az adott szempontban. Harmadrészt a három szempont egyike sem kitüntetett vagy preferált.

Harshman (1984) a két modell öt fő különbségét sorolja fel. Ezek közül az első elméleti, a többi négy algebrai különbség, és az elsőből következik:

a) A Tucker3 (T3)-modell más „faktor" fogalmon alapul, mint a PARAFAC.
b) A Tucker3-modellben minden faktorkombináció lehetséges, bármelyik szempont bármelyik faktora interakcióban lehet a többi szempont faktoraival; a faktorsúlyoknak nemcsak az $a_{j1}b_{k1}c_{i1}$ kombinációja megengedett, hanem az $a_{j1}b_{k3}\,c_{i2}$ és így tovább.
c) A T3 tartalmaz egy negyedik paraméterhalmazt is: $g_{pq(m)}$, ami a faktoroknak a különböző szempontok közötti interakcióját tartalmazza. A T3 ezért quadrilineáris, nem pedig trilineáris.
d) A T3-modellben az A szempont faktorainak a száma (az **A** oszlopvektorainak a száma) különbözhet a B szempont és a C szempont faktorainak (komponenseinek, latens változóinak) a számától.
e) A T3 nem rendelkezik a valódi tengely megoldással, de nagyobb a faktorok transzformációjának lehetősége, mint a kétszempontos faktorelemzésnél.

A koncepcionális különbség azt jelenti, hogy a PARAFAC-modellben a faktor empirikus entitás, folyamat vagy hatás, amit a megfigyelt változókon keresztül mérünk, nem az adatok, mérések klasszifikációjának eredménye. A PARAFAC-faktor a három partikuláris szempont mindegyikére ható latens változó. A T3-modell faktora viszont az egyes szempontok idealizált eleme, típusa, karaktere. Tucker koncepciójában tulajdonképpen nem a latens faktorok magyarázzák a mérési adatok varianciáját, hanem az egyes szempontok faktorainak interakciói, kölcsönhatásai. Mivel a három szempont szerint a faktorok különbözőek, nem is kell, hogy számuk megegyezzen, szemben a PARAFAC faktoraival, amelyek közötti interakciónak nincs értelme, ezért nem is szerepel a modellben a faktorok kapcsolatait kifejező belső struktúramátrix.

18.11. A faktorsúlyok értelmezése

A faktorsúlyok általános értelmezését korábban már definiáltuk. Ha a faktorok ortogonálisak, akkor a faktorsúlyokat mint a variancia (vagy négyzetes átlag) összetevőit értelmezhetjük egy adott szempont valamely kategóriájában. Vegyük az A szempont j-edik kategóriáját, a j-edik megfigyelt változót, és számítsuk ki a másik két szempont szerinti négyzetes átlagát (mean square):

$$\frac{1}{KI}\sum_{k}^{K}\sum_{i}^{I}\left(x_{jk(i)}\right)^2 = MSQ_j.$$

Behelyettesítve a PARAFAC-modellt a fenti egyenletbe:

$$\frac{1}{KI}\sum_{k}\sum_{i}\left(\sum_{m}(a_{jm}b_{km}c_{im}) + e_{jk(i)}\right)^2 = MSQ_j.$$

Mivel a hibakomponens ortogonális (korrelálatlan) a szisztematikus komponenssel (a faktorokkal), a következőt kapjuk:

$$\frac{1}{KI}\sum_k\sum_i\left(\left(\sum_m(a_{jm}b_{km}c_{im})\right)^2+(e_{jk(i)})^2\right)=MSQ_j.$$

Rendezzük át az összegezést (kommutativitás) és emeljük ki a konstans a_{jm} együtthatót a k és i indexű szummázás alól:

$$\frac{1}{KI}\sum_m\left(a_{jm}^2\left(\sum_k\sum_i\left(b_{km}^2c_{im}^2\right)\right)\right)+\frac{1}{KI}\sum_k\sum_i e_{jk(i)}^2=MSQ_j,$$

vagy

$$\sum_m\left(a_{jm}^2\left(\frac{1}{K}\sum_k b_{km}^2\right)\left(\frac{1}{I}\sum_i c_{im}^2\right)\right)+\frac{1}{KI}\sum_k\sum_i e_{jk(i)}^2=MSQ_j.$$

Ha standardizáljuk a faktorsúlyokat a B és C szempont szerint úgy, hogy a négyzetes átlaguk 1 legyen, akkor a következőket kapjuk:

$$\sum_k(a_{jm}^2)(1)(1)+\frac{1}{KI}\sum_k\sum_i e_{jk(i)}^2.$$

Végeredményként azt kaptuk, hogy ha a B és C szempontban a faktorok ortogonálisak, akkor az A szempontban az **A** faktorsúlymátrix sorvektorai elemeinek a négyzetösszege egyenlő lesz a megfigyelt változók négyzetes átlagának azzal a részével, amit az M számú faktorral magyarázni tudunk (a teljes négyzetes átlag mínusz a hiba négyzetes átlaga). Ebből következik, hogy egy faktorsúly négyzete az adott faktor hozzájárulását adja ehhez a magyarázathoz, valamint ha az **A** mátrix oszlopvektorai elemeinek a négyzetösszegét vesszük, akkor az egyes faktorok hozzájárulását kapjuk a teljes négyzetes átlaghoz, továbbá ha ezeket is összegezzük, akkor a PARAFAC-modell által magyarázott részét kapjuk a teljes négyzetes átlagnak.

Ha a változók (A szempont kategóriái) várható értéke nulla, akkor a négyzetes átlag (MSQ) egyenlő lesz a varianciával.

18.12. Példa: Tucker-modell *(Gyermeknevelési elvek változásai a magyar társadalomban [1978–1998])*

Adatok

Az MTA Szociológiai Intézet Értékszociológiai Műhelye a ’70-es évek végétől végez kérdőíves felméréseket az ország értékrendszerének, az értékek struktúrájának feltérképezésésről. A vizsgálatok során számos értékteszt használatára sor került. Ezek közül a leghatékonyabbnak a Gyermeknevelési elvekről szóló teszt, és a Rokeach-teszt bizonyult, ami nemzetközileg elfogadott, széles körben használt értékteszt. A vizsgálatok adatai lehetőséget nyújtanak arra, hogy az elmúlt három évtized változásait figyelemmel

kísérhessük. A gyermeknevelési elvek ilyen módon való összehasonlítására az eddigiekben még nem került sor.

A kérdőívekben használt Gyermeknevelési teszt a következő volt:

Néhány olyan tulajdonságot soroltunk fel, amire nevelni lehet a gyerekeket.
Melyeket tartja Ön különösen fontosnak? Kérem, válassza ki az öt legfontosabbat!

GY1 — jó magaviselet
GY2 — udvariasság
GY3 — önállóság
GY4 — a kemény munka szeretete
GY5 — őszinteség
GY6 — felelősségérzet
GY7 — türelem
GY8 — képzelőerő, fantázia
GY9 — mások tisztelete, tolerancia
GY10 — vezetőkészség
GU11 — önfegyelem
GY12 — takarkosság
GY13 — határozottság, állhatatosság
GY14 — vallásos hit
GY15 — önzetlenség
GY16 — engedelmesség
GY17 — hűség, lojalitás

Adatbázis

- **Életút-Értékrendszer 1982**. A kérdőív. MTA Szociológiai Intézet.
  Minta: Az ország felnőtt lakosságát reprezentáló minta.
  Esetszám: 1464 fő. Érvényes esetek száma: 1395 fő.

- **Mobilitás kérdőív**, TÁRKI, **1992.** március
  Minta: Az ország felnőtt lakosságát reprezentáló minta.
  Esetszám: 1500 fő. Érvényes esetek száma: 1393 fő.

- **Értékrendszer vizsgálata. 1996.** MTA Szociológiai Intézet.
  Minta: Az ország felnőtt lakosságát reprezentáló minta.
  Esetszám: 1600 fő. Érvényes esetek száma: 1516 fő.

- **Értékrendszer vizsgálata. 1997.** MTA Szociológiai Intézet.
  Minta: Az ország felnőtt lakosságát reprezentáló minta.
  Esetszám: 1500 fő. Érvényes esetek száma: 1442 fő.

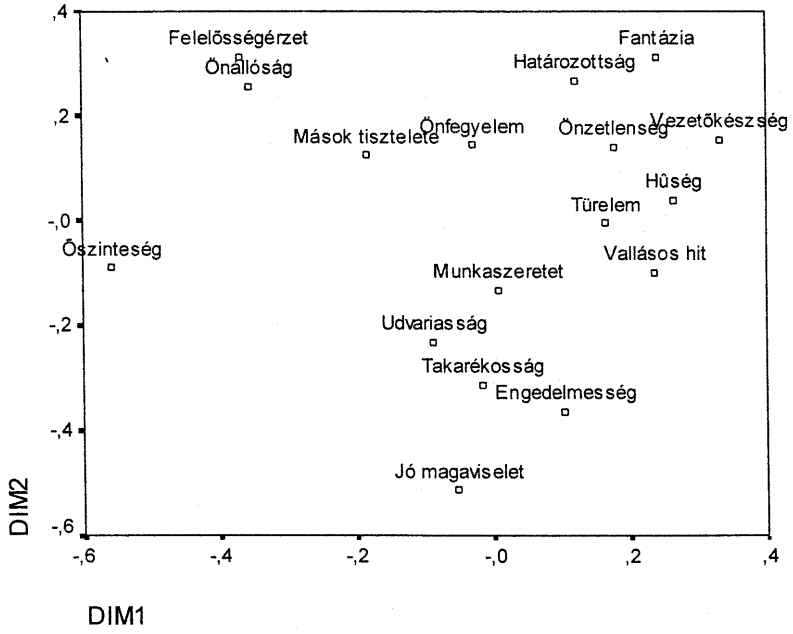
- **Értékrendszer vizsgálata. 1998.** MTA Szociológiai Intézet.
  Minta: Az ország felnőtt lakosságát reprezentáló minta.
  Esetszám: 1521 fő. Érvényes esetek száma: 1419 fő.

Eredmények:

Értékrendszer Tucker3 modellje

Gyermeknevelési elvek (first mode – variables)
Iskolai végzettség (second mode – conditions)
Vizsgálati évek (third mode – subjects)

```
COMPONENT LOADINGS OF MODE   1          Gyermeknevelési elvek
SCALING OF COMPONENTS UNIT LENGTH
```

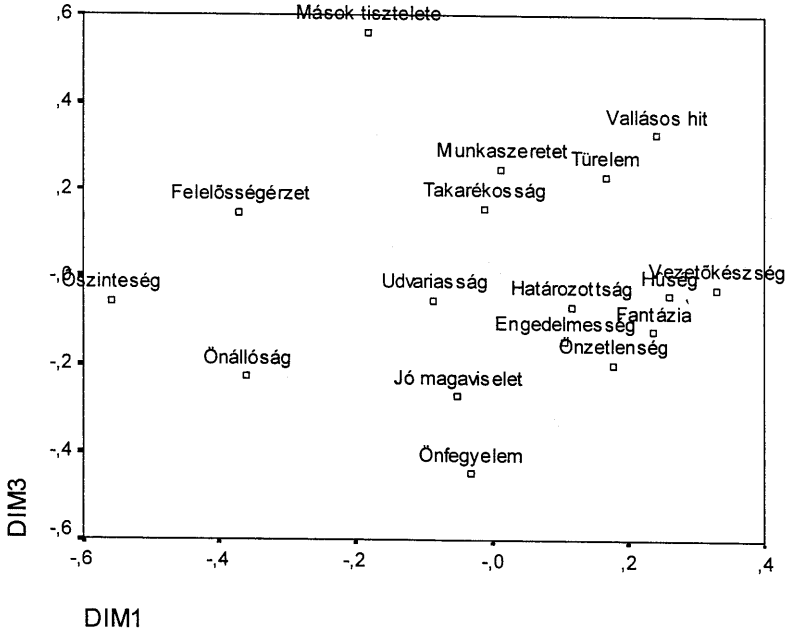



```
STANDARDIZED COMPONENT WEIGHTS OF MODE   1
       :    1.        2.        3.
-------+-----------------------------
     1 :     .8242     .1347     .0119
```

Értékrendszer Tucker3 modellje

Gyermeknevelési elvek (first mode – variables)
Iskolai végzettség (second mode – conditions)
Vizsgálati évek (third mode – subjects)

```
COMPONENT LOADINGS OF MODE   1          Gyermeknevelési elvek
SCALING OF COMPONENTS UNIT LENGTH
```

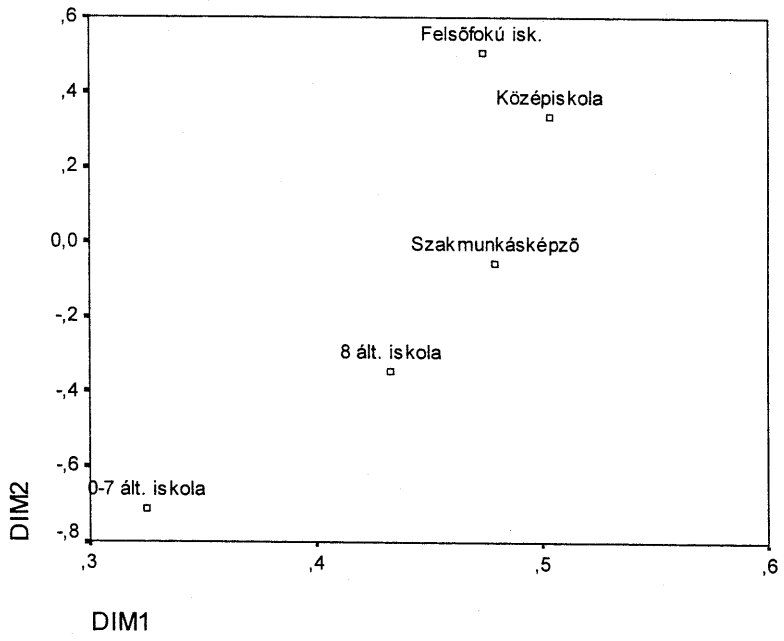



```
STANDARDIZED COMPONENT WEIGHTS OF MODE   1
       :    1.        2.        3.
-------+------------------------------
     1 :     .8242     .1347     .0119
```

Értékrendszer Tucker3 modellje

Gyermeknevelési elvek (first mode – variables)
Iskolai végzettség (second mode – conditions)
Vizsgálati évek (third mode – subjects)

```
COMPONENT LOADINGS OF MODE   2          Iskolai végzettség
SCALING OF COMPONENTS UNIT LENGTH
```

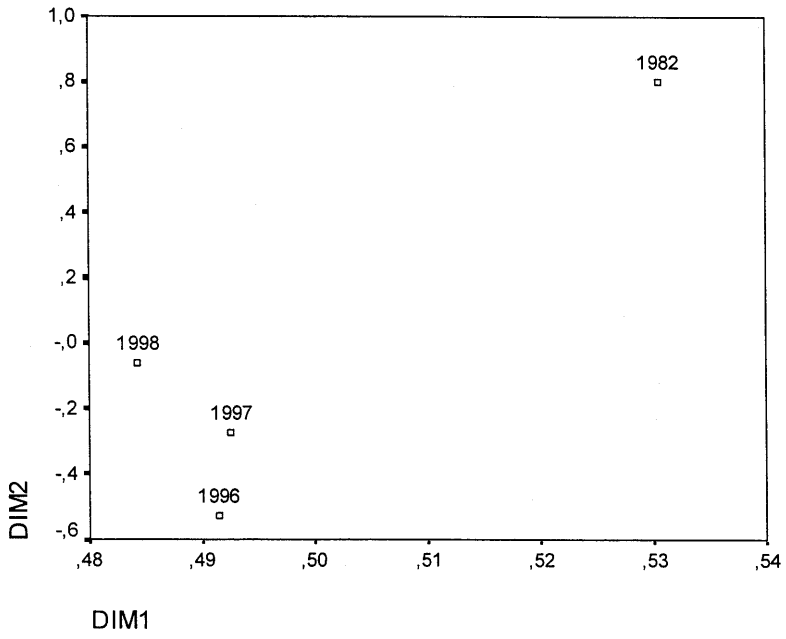



```
STANDARDIZED COMPONENT WEIGHTS OF MODE   2
       :    1.        2.
 ------+--------------------
     1 :     .8392     .1316
```

Értékrendszer Tucker3 modellje

Gyermeknevelési elvek (first mode – variables)
Iskolai végzettség (second mode – conditions)
Vizsgálati évek (third mode – subjects)

```
COMPONENT LOADINGS OF MODE   3          Vizsgálat évei
SCALING OF COMPONENTS UNIT LENGTH
```



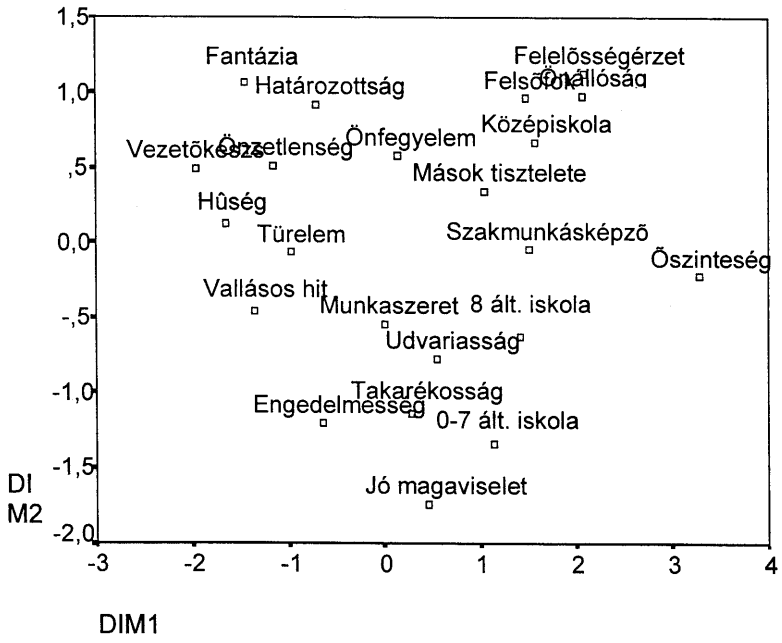

```
STANDARDIZED COMPONENT WEIGHTS OF MODE   3
       :    1.        2.
 ------+--------------------
     1 :     .9551     .0157
```

Értékrendszer Tucker3 modellje

Gyermeknevelési elvek (first mode – variables) és
Iskolai végzettség (second mode – conditions)
Közös tere
Vizsgálati évek (third mode – subjects)

JOINT PLOT OF MODE 1 (Gyermeknevelés) AND MODE 2 (Iskolai végzettség)

FOR Vizsgálati év, COMPONENT 1



COMPONENT WEIGHTS FOR AXES OF JOINT PLOT

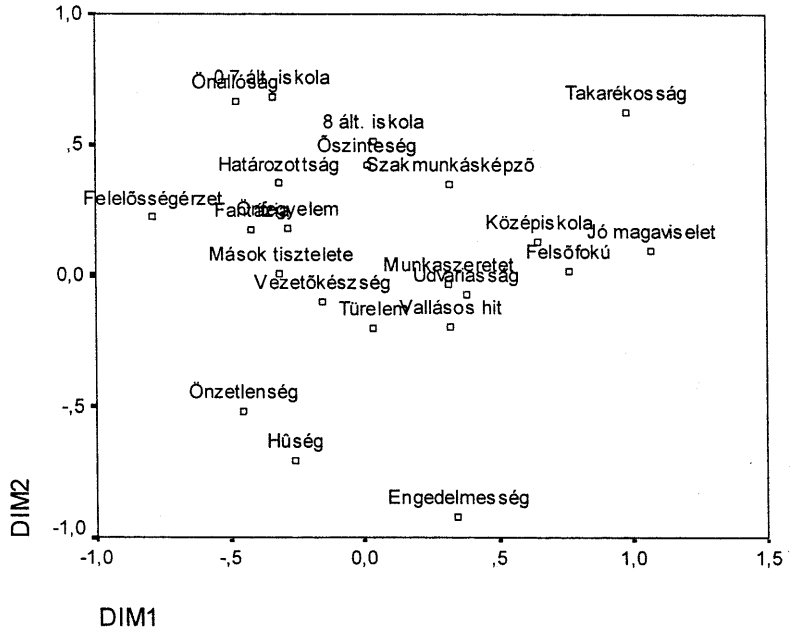
:	1.	2.
1 :	.8239	.1312

Értékrendszer Tucker3 modellje

Gyermeknevelési elvek (first mode – variables) és
Iskolai végzettség (second mode – conditions)
Közös tere
Vizsgálati évek (third mode – subjects)

JOINT PLOT OF MODE 1 (Gyermeknevelés) AND MODE 2 (Iskolai végzettség)

FOR Vizsgálati év, COMPONENT 2



COMPONENT WEIGHTS FOR AXES OF JOINT PLOT

:	1.	2.
1 :	.0153	.0004

IV. rész

Társadalomtudományi alkalmazások

19. fejezet

Kontinuitás és diszkontinuitás az értékpreferenciákban (1977–1998)

Bevezető

Az 1989-es fordulatot követően két olyan kérdés merült fel, amelyekre a szakirodalom azóta is többször visszatért. Ez a két kérdés a következő: 1) Mi az oka annak, hogy a kommunista rendszerek bőséges irodalma nem jósolta meg a változások váratlanságát és radikalitását (Lipset–Bence 1994); 2) Hogyan lehetséges, hogy a kommunista rendszerek összeomlása és a Vörös Hadsereg kivonulása után a kommunista párt öröksége még mindig kísért, holott – legalábbis azt hittük – felszabadultak bűvöletéből (Holmes 1997; Janos 1994; Mokrzycki 1992; Schöpflin 1993; Szakolczai–Horváth 1991; Tismaneanu 1992; Verdery 1996). A két kérdésfeltevés paradox módon szorosan egymáshoz kapcsolódik. Az első kérdés tehát: miként jöhetett létre ez a radikális és előre nem látott változás, a második kérdés pedig arra keresi a választ, miként lehetséges, hogy a változások ellenére sok minden változatlanul tovább él a gondolkodásban.[1]

A kontinuitás és diszkontinuitás ilyen kombinációja részben választ adott, és átfogalmazta a legnyugtalanítóbb kérdések egyikét: a kommunista ideológia és politika hatásának mértékét. Az 1956-os budapesti, az 1968-as prágai és az 1980-as gdanski hősies események hatására az önszerveződő civil társadalom heves ellenállást tanúsított az állam

[1] Készült az OKTK Ny.sz.: A. 1703/VIII. b. 99. *Érékváltozás Magyarországon 1977–1999* és az OTKA Ny. sz.: T016435 *Értékváltozás 1978 és 1994 között Magyarországon* című kutatások keretében, Szakolczai Árpád közreműködésével.

és a párt támadásaival szemben, és szavát hallatta, ha bármikor alkalma volt rá.[2] Ezzel ellentétben az értelmiség már a rendszer létrejöttekor is azt állította, hogy a kommunizmus nem csupán külső diktatúra volt, mivel gyökeresen megváltoztatta a hatalmába kerített országokat.[3]

Első látásra úgy tűnt, a változások ténye és jellege véglegesen cáfolja a „homo sovieticus" képét (Zinoviev 1981). Az örökség jelenléte azonban más forgatókönyvről tanúskodik. A kommunista szellem sikerének vagy összeomlásának paradoxona, akárcsak a múltban, még mindig kísért.

Ha egy problémát nem lehet megoldani, annak oka a kérdés feltevésében rejlik. Esetünkben nem a különböző kérdések szintézise jelenti a megoldást, hanem a kérdés vagy probléma ismételt áttekintése. A kommunizmus hatása esetében ez a rendszer direkt, manifeszt és indirekt, latens hatásának megkülönböztetésében rejlik. A direkt, manifeszt hatás a rendszer ideológiájához való hűség, egy spontán, a létező szocializmus alulról jövő reprodukciója, a *Big Brother* orwelli szeretete. Az elmúlt évtized eseményei kétségtelenül azt bizonyították, hogy ez nem következett be. A még mindig élő örökség bizonyíthatóan egy másik, kevésbé megközelíthető, sokkal inkább latens, de ugyanolyan valóságos szinten van jelen: nem ideológiákban és meghatározott célokban, hanem a mentalitásokban; nem a tartalomban, hanem a viselkedés formájában és modalitásában, nem konkrét, formális intézményekben és szervezetekben, hanem az intézményekhez kötődő attitűdökben és értékekben, valamint az intézményesítés folyamatában.

A manifeszt és a latens között viszonylag könnyű különbséget tenni. Ez épülhet az intellektuális hagyományra, amely különösen a német filozófiai idealizmusban van jelen. Tanulmányunk a fenomenológiai szociológiára és az ezzel kapcsolatos megközelítésekre épít (Schutz 1962; Bateson 1972; Gadamer 1975; Goffman 1986; Searle 1992, 175). E megközelítésekben a latens nem maga a mögöttes fogalom, hanem a magatartás valamilyen formája. E tekintetben az előző rendszer latens és tartós öröksége nem a kommunista ideológia lényegéhez kapcsolódik, hanem a magatartás unalmas, hétköznapi, rutinszerű megjelenéséhez, ami gyakorlói számára bizonyos, adott, nyilvánvaló, gyakorlatias, természetes és általános, és semmiképpen nem a kommunizmus ideológiájának következménye.

Az ilyen perspektíva segíti összekapcsolni az értékeket a mindennapi gyakorlattal, a magatartással.

19.1. Az értékekről

Elfogadott tény, hogy a múlt öröksége – különösen a politikai kultúra, az intellektuális hagyományok, a megrögzött mentalitások, egyszóval az értékek szintjén – rendkívül erős (Schöpflin 1993; Kennedy 1994). Az értékek iránti igényeket közismerten nehéz felbecsülni és igazolni. Ebben a tanulmányban az értékek empirikus tanulmányozásával

[2] Ezt a nézetet vallják Václav Havel, Konrád György és Adam Michnik a „civil társadalom"-ról írott, jól ismert műveikben.

[3] Hasonló nézeteket vallottak az adott országok vezető intellektuális személyiségei, mint például Milan Kundera, Illyés Gyula, Czeslaw Milosz.

próbáltunk bizonyos következtetéseket levonni hat (1977–1997 közötti) országos reprezentatív minta adatainak felhasználásával.

Az ilyen irányultságú megközelítés további ellenvetések sorozatát jelenti. Lehetséges-e a társadalmi értékek vizsgálata egyének válaszai alapján? Hozzáférhetőséget biztosítanak-e ezen preferenciák a mögöttes, latens értékek szintjéhez, ahol a múlttal való folyamatosság feltehetően valótlan? És egy még ezeknél is alapvetőbb ellenvetés: számítanak egyáltalán a felmérés kontextusában az egyének által kifejezett értékek, különös tekintettel a preferenciákra?

Nyilvánvalóan nem lehet ezekre az ellenvetésekre egyetlen tanulmányban kimerítő választ adni. Lehet azonban érvelni azzal, hogy a szociológiában van egy elméletileg következetes és klasszikus perspektíva, ami lehetővé teszi az individuális értékek empirikus tanulmányozásának elméleti alátámasztását. Ezt megtalálhatjuk Max Weber vallásszociológiájában.[4]

Az értékszociológia klasszikus paradigmáiban az értékek vagy átívelő, kötelező normákként, azaz közösségi szinten léteznek és nem alkalmasak mikroszintű individuális elemzésre (ez az elmélet legalább Parsonsig [1968] visszavezethető), vagy feltételezhetően az individuális szükségletekben vannak jelen, és minden ember számára univerzálisak (e nézetet Maslownál találjuk meg [1959]). A weberi érv mindkét állítással szemben áll. Szerinte az értékeket individuális és nem kollektív egységként kell elemezni, mivel azok az egyén szokásaihoz, életvezetésének (Lebensführung) miként jéhez kapcsolódnak. Ez nem a pozitív vagy normatív módszertani individualizmusra jellemző előfeltételnek köszönhető, hanem az empirikus tényhez, szükséghez társuló viselkedésmódhoz kapcsolódik. Az életvezetés akkor kapcsolódik értékekhez, amikor korábban már bizonyosnak vett formái problematikussá válnak (Gadamer 1976; Foucault 1984; Elias 1994, 58, 518).

Ez akkor válik társadalmilag relevánssá, ha több egyénnél ugyanabban az időben jelentkezik. Ez olyan ritkán előforduló esetekben történik, amikor a dolgok addigi rendje felbomlik (Voegelin 1978, 89–115), vagy átmeneti időszakokban (Borkenau 1976; Elias 1983, 1994), illetve „liminal" periódusokban (van Gennep 1960; Turner 1967, 1969). Mivel az ilyen időszakokra jellemző stabil referenciapontok megtörnek, az egyének magukra maradnak, hogy saját életükben stabilitásról és irányításról gondoskodjanak.

Azok a különböző értékrendek, amelyekből a társadalmi értékrendszer összetevődik, ma ilyen elmozdulások „bélyegét" hordozzák (Weber 1948, 268, 80; Elias 1983, 39–40).

A rend felfüggesztése vagy a „küszöb"-feltétel nem csupán az egyént „bélyegzi meg", akinek el kell viselnie az ilyen periódusok nyomását. Ellenkezőleg, arra kényszeríti az egyént, hogy kilépjen mindennapi tevékenységéből, stimulálva a gondolkodást és elmélkedést (Elias 1987; Voegelin 1978, 11–12; Bourdieu 1990). Az átalakulás rendkívül instabil időszakában a gondolkodás nem szorítkozik a létező társadalmi és politikai struktúrák ideáin és megvalósulási formáin való elmélkedésre, mivel az ilyenfajta stabil struktúrák már felbomlottak, hanem inkább reflexív aktivitása annak a gondolkodásmódnak, amely meghatározza a rend alapjait szolgáló mentális és intézményi, társadalmi és individuális szerkezeteket és identitásokat.

Az értékek tehát nem azonosak a (szupraindividuális) normákkal, egy adott társadalom konszenzuális elveivel. Nem gyökereznek az ember (szubindividuális) biológiai vagy fiziológiai szükségleteiben sem. Az értékek többnyire az egyén életvezetéséhez kapcsolódnak, tényleges életelvek, amelyek a viselkedést irányítják, és hozzáférhetővé válnak, amint a gondolkodás által adekvátan megfogalmazódnak.

[4] Lásd Weber (1948 és 1982). Az elméleti megközelítést lásd Szakolczai és Füstös (1994 és 1998), valamint Szakolczai (1998).

Az ilyen értelemben vett értékek individuális szinten és szociológiai felmérésekkel is vizsgálhatók. Ehhez azonban olyan speciális, reflexív megközelítés szükséges, amit általában nem alkalmaznak a közvélemény-kutatásban. A felmérések vagy a tényleges viselkedésre irányulnak (ahol a kérdés az, hogy a válaszoló a valóságos cselekedeteiről számoljon be), vagy attitűdökre (ahol az információ csupán egy gyors impulzusra adott azonnali reakció). A felmérések nem olyan kérdéseket tartalmaznak, ahol a válaszolónak perceken keresztül gondolkodnia kell. Az ilyen megközelítést – anyagi okok miatt – ritkán alkalmazzák.

Létezik azonban egy értékteszt, amelyet speciálisan az egyéni értékek reflexív elemzésére dolgozott ki a lengyel születésű amerikai szociálpszichológus, Milton Rokeach. Többéves klinikai kísérletezés után Rokeach (1973) egy 18 cél- és 18 eszközértékből álló tesztet állított össze, amely véleménye szerint kifejezetten reflexív orientációval a teljes értékteret reprezentálja (lásd a Függeléket). A válaszadóknak mindkét értékcsoporthoz tartozó értékeket 1-től 18-ig kellett rangsorolniuk, elvben összehasonlítva minden értékpárt. A teszt kitöltése 15–20 percet vesz igénybe, ami egy szokásos kutatás számára túlságosan nagy költséget jelent. Azonban ma már elfogadott, hogy a Rokeach-teszt rendkívül érzékeny és megbízható eszköze az egyéni értékek latens dimenzióinak és konfigurációjának feltérképezésére (Feather 1975; Schwarz 1994).

19.2. Adatok és módszerek

Az adatok az MTA Szociológiai Kutatóintézet Értékszociológiai Műhelyének országos reprezentatív mintán végzett vizsgálataiból származnak. A vizsgálatokat Hankiss Elemér, Manchin Róbert és Füstös László vezette 1977–1978-ban, 1982-ben, 1990-ben, 1993-ban, 1996-ban, 1997-ben és 1998-ban.[5] A minta elemszámai sorrendben 807, 2938, 1320, 1538, 1500, 1500, illetve 1521 fő. A tesztet helyesen kitöltők száma 677, 2089, 1063, 1150, 1327, 1347 és 1179. Jelen fejezetben kétoldali *t*-tesztek és varianciaelemzés segítségével a felnőtt népesség átlagos értékpreferenciáit elemezzük.

Háttérinformáció

A 20 évet átölelő, 36 értékből álló tesztben történő szimultán változások értelmezésének megkönnyítése érdekében három további információt adunk meg.

Szemantikus információ. Bár Rokeach célja két 18 – jelentésében egymástól független – értékből álló halmaz kifejlesztése volt, nyilvánvaló, hogy közöttük vannak olyanok, amelyek jelentése jobban hasonlít egymáshoz, mint másokéi, és vannak olyanok, amelyek szemben állnak egymással. Ezeket a hasonlóságokat és különbözőségeket többféle módszerrel lehet vizsgálni, mint például a faktorelemzés vagy a sokdimenziós skálázás. Az ilyen elemzések számos ideológiai érték közötti szoros kapcsolatot tárnak fel, mint például a BÉKE, a HAZA BIZTONSÁGA, a SZABADSÁG és az EGYENLŐSÉG és a velük ellentétes bizonyos materialista vagy hedonista értékek, mint például az ANYAGI

[5] Vannak adatok 1979–1980-ra és 1995-re is. Ezek azonban kisebb elemszámú mintából állnak, és nem reprezentálják az ország felnőtt népességét.

JÓLÉT, az ÉLVEZETES ÉLET közöttieket. A módszer illusztrálására a 19.1. táblázat nyolc magyarországi megfigyelési év első főkomponensének és az amerikai főkomponens súlyait tartalmazza. Az első főkomponens egy adott társadalom értékrendszerének fő választóvonalaként értelmezhető.

Társadalmi információ. Az értékek jelentése és különösen preferenciája erős társadalmi alkotóelemmel rendelkezik. Az értékpreferenciák az életkor, a nem, a foglalkozás és különösen az iskolázottság szerint különböznek. Vizsgálatunk különböző almintákat is tartalmaz, mint például az egyetemi hallgatók, a menedzserek, a kommunista pártiskolákban képzettek vagy a cigányok mintája. Ezek a vizsgálatok bizonyos értékek területén még további lehetőséget nyújtanak az értékek változásainak értelmezésében. Például a BÉKE és a HAZA BIZTONSÁGÁnak erős preferenciáját találtuk azoknál, akik kommunista pártiskolát végeztek. Náluk ezek a legfontosabb hivatalos szocialista értékek (Szakolczai 1987), míg a MUNKA ÖRÖME, a TÁRSADALMI MEGBECSÜLÉS és az EMBERI ÖNÉRZET inkább klasszikus szocialista vagy szociáldemokrata értékek, és ezeket inkább a férfi szakmunkások preferálják.

Rendszerinformáció. Anélkül, hogy az egyes értékeket azonosítanánk a „carrier stratá"-val (Weber 1948), az 1977–1978-as magyar és az 1968-as USA-beli adatok, valamint a megfelelő faktorstruktúrák közötti összehasonlítás körvonalazza a magyar értékpreferenciák sajátosságait (lásd az 19.1. és 19.2. ábrát).

1. Magyarországon jelen volt egy olyan értékrend, ami az amerikai mintában nem található: a „klasszikus" szocialista vagy szociáldemokrata értékrend, amely három érték körül csoportosul: a TÁRSADALMI MEGBECSÜLÉS, a MUNKA ÖRÖME és az EMBERI ÖNÉRZET. E három érték Magyarországon külön faktort alkotott, míg az Egyesült Államokban hasonló faktor nem létezett. Ugyanakkor két érték, a MUNKA ÖRÖME és a TÁRSADALMI MEGBECSÜLÉS fontosabb volt a magyar, mint az USA-beli mintában. Ugyanez azonban nem érvényes az EMBERI ÖNÉRZETre, melynek az Egyesült Államokban más kontextusa volt.

2. Magyarországon nem ez volt az egyedüli, sajátos értékrend, ami kiérdemelte a „szocialista" elnevezést. Egy másik, a HAZA BIZTONSÁGA, BÉKE, EGYENLŐSÉG és SZABADSÁG értékcsoportot az előző klasszikus szocialista értékekkel szemben „hivatalos" vagy „ideológia" szocialista értékeknek nevezhetnénk. Ezek alkották – különösen az első kettő – azokat az értékeket, amelyek a magyar minta értékpreferenciájának erős jellegzetességét adták; Magyarországon 1977–1978-ban a BÉKE volt a legfontosabb érték. A HAZA BIZTONSÁGÁt átlagosan az Egyesült Államokhoz képest Magyarországon közel négy ponttal sorolták előbbre. Az értékpreferenciák struktúrájában a BÉKE és a HAZA BIZTONSÁGÁnak egyedülálló jelentősége volt. Jelen voltak az első két főkomponensben, valamint az öt rotált faktor közül háromban. Mindkét megállapítás nagyon fontos. Egyfelől az első két – a mintán belül a legfőbb érték választóvonalat tükröző, cél- és eszközértékeket egyaránt tartalmazó – főkomponens utal arra, hogy ezek voltak azok az ideológiai, hivatalos szocialista értékek, amelyek a társadalmi tudat szintjén összekapcsolták a célokat és az eszközöket. Másfelől ezeknek az értékeknek jelentős a szerepe a rotált faktorstruktúra előállításában; az ideológiai szempontok mindenkori jelenlétét mutatja az értékrendszer finomabb szerkezetében is.

3. A két különböző szocialista értékrend mellett jelen volt a magyarok által az amerikaiaknál jobban kedvelt értékcsoport, az intellektuális értékeké is (ÉRTELMES,

ALKOTÓ SZELLEMŰ, LOGIKUS GONDOLKODÁSÚ). Az Egyesült Államokban ezek az eszközértékek rangsorában az utolsó helyeken szerepeltek, míg nálunk, Magyarországon, az átlag fölött álltak (2,5–4 ponttal). Ez az eredmény megerősíti az előző, a szocialista rendszer intellektuális jellegét valló nézőpontokat (Bauman 1987; Konrád és Szelényi 1979). Az intellektuális értékek és az ideológia ilyen kapcsolatát az a tény is alátámasztja, hogy az első – a társadalmi értékrendszer legfontosabb választóvonalát definiáló – főkomponens azonos oldalán található ez a két értékrend (lásd 19.1. táblázatot).

4. A hivatalos szocialista értékrenddel ellentétben a második rotált faktor egyik pólusán 1977–1978-ban a vegyes értékcsoport található: CSALÁD, SZERELEM, SZERETETTELJES, BOLDOGSÁG, BELSŐ HARMÓNIA, ANYAGI JÓLÉT. Ezek az értékek általában különböző értékrendekhez tartoznak, a közöttük lévő korreláció nem túl erős. Így a SZERELEM szorosan kapcsolódik a BARÁTSÁGhoz, az ANYAGI JÓLÉT a KELLEMES, az ÉLVEZETES ÉLEThez, a BELSŐ HARMÓNIA a BÖLCSESSÉGhez, a SZERETETTEL TELJES a MEGBOCSÁTÓhoz. A faktor mégis ezen értékpároknak csupán az egyikét tartalmazza. Egyetlenegy negatív jellemvonásuk közös: mindegyik a privát szférához tartozik, és ellentétben áll a nyilvános szféra hivatalos szocialista értékeivel. A merev ideológiai orientációval szemben ezek az értékek az emberi kapcsolatokat és érzelmeket (SZERELEM, SZERETETTELJES, BOLDOGSÁG), a kisközösséghez (CSALÁD), a személyiséghez (BELSŐ HARMÓNIA) vagy az anyagi jóléthez (ANYAGI JÓLÉT) való ragaszkodást hangsúlyozzák. Hangsúlyozottságuknak köszönhetően ezek az értékek Magyarországon fontosabbaknak bizonyultak, mint az Egyesült Államokban. Ez még inkább igaz, ha az „iker"értékekkel hasonlítjuk össze. Eszerint az ANYAGI JÓLÉT és a BOLDOGSÁG 1,5 ponttal vezet Magyarországon, a BELSŐ HARMÓNIA és a SZERELEM 2 ponttal, míg a BARÁTSÁG ugyanazon a szinten áll, a BÖLCSESSÉG pedig 6 ponttal alacsonyabban. A SZERETETTELJES az egyetlen olyan erősen preferált „iker"érték, amely az amerikai mintában 2,5 ponttal, míg párja, a MEGBOCSÁTÓ 4,5 ponttal került alacsonyabbra a magyar mintában. Közülük 1977–1978-ban a párttagok és párton kívül állók közötti preferenciakülönbségeknek megfelelően a SZERELEM, a CSALÁD és a BELSŐ HARMÓNIA a leginkább „ellenzéki" értékek (Szakolczai 1987).

5. Nem meglepő, hogy a magyar és az amerikai minta közötti legnagyobb különbség a vallásos értékekben található. Az Egyesült Államokban az ÜDVÖZÜLÉS egyike a legfontosabb célértékeknek, míg Magyarországon az utolsó helyet foglalja el. A mediánok közötti különbség 9 pont. Hasonló a különbség a két fent említett – SZERETETTELJES és különösen a MEGBOCSÁTÓ – eszközérték esetében is. Fontos megemlítenünk, hogy ezek nem szigorúan vallásos értékek, hanem általában a konkrét személyes kapcsolatokra, a közösségi életre vonatkoznak. Az a tény, hogy Magyarországon nagyon alacsony a preferenciájuk, nem csupán a vallásosság hiányát tükrözi, hanem azt is jelzi, hogy ezen értékszempontnak nem tulajdonítanak jelentőséget. Mind Magyarországon, mind az Egyesült Államokban az intellektuális értékek ellenkező pólusán jelennek meg. Korábbi, a projekthez kapcsolódó kiadványok: Hankiss (1982, 1986), Szakolczai (1982, 1987), Hankiss, Manchin és Füstös (1990).

6. Az utolsó különbség a pragmatikus értékekre vonatkozik. Szemantikailag akár az intellektuális (mint például a HATÉKONY), akár a materialista értékekhez tartoznak (mint például a TÖREKVŐ); az amerikai mintában lényegesen előbbre rangsorolták őket.

Az eredmények azt bizonyítják, hogy Magyarországon egyrészt inkább az ideológiai, mint a vallásos értékek kerültek előtérbe, másrészt kiemelten hangsúlyozott az intellektualizáció – fogalmi ellentétének, a szociabilitásnak és az interperszonális kapcsolatoknak a rovására. A pragmatizmus hiánya jellemzi mind az ideologizációt, mind az intellektualizációt. Kérdés, hogy ezek 1989 óta milyen mértékben változtak.

19.3. Általános hipotézisek

Tanulmányunk azt a központi állítást teszteli, amely szerint a kommunizmus lényeges és tartós, bár többnyire indirekt és latens hatást gyakorolt a társadalomra és az értékpreferenciákra.[6] Az egyik feltevés szerint Magyarországon a hetvenes évek értékpreferenciáinak néhány sajátossága a kilencvenes évek közepén is megmaradt. A másik állítás szerint a kommunista időszakhoz viszonyítva mégis jelentős elmozdulás jött létre az értékpreferenciák szintjén, cáfolva elsősorban a legnyilvánvalóbb kommunista értékek esetében a jelentős folytonosságot.

Ezzel csupán az 1989 előtti és utáni folytonosság és szakadás kombinációja támasztható alá, ami önmagában triviális. Kiegészítjük tehát pozitív, általánosabb hipotézisekkel. Ezek a direkt versus indirekt, manifeszt versus latens ellentétpárok megkülönböztetésére vonatkoznak. Az érvelés elsősorban azt tartalmazza, hogy a célértékekben nagyobb változás történt, mint az eszközértékekben. A célértékek kifejezetten az életcélokra, míg az eszközértékek a cselekvés és viselkedés módjára vonatkoznak. A célértékek nyilvánvalóbbak és közvetlenebbek, a tudatosság felszínéhez állnak közelebb, míg az eszközértékek a megrögzött szokásokhoz, az életvezetés körülményeihez (Mumford 1952; Weber 1995) vagy a habitushoz (Elias 1991; Bourdieu és Wacquant 1992) kapcsolódnak. El kell tehát fogadnunk, hogy 1989 aligha jelent fontos elmozdulást az eszközértékek szintjén, míg az alapvető változások inkább a célértékek átlagos preferenciájában mutatkoznak.

Figyelembe véve a szocialista vagy kommunista értékekhez való kapcsolatot, *első hipotézisként* arra számítunk, hogy a célértékek között elsősorban a látencia és közvetlenség ugyanazon dimenziójában számottevő különbségeket találunk. *Második általános hipotézisként* azt állítjuk, hogy 1989-et megelőzően és azt követően a klasszikus szocialista ideológiához tartozó értékek átlagos preferenciái radikális különbséget mutatnak. *A harmadik általános hipotézis* az 1993–1994-es évek átmeneti nosztalgiájára vonatkozik, amelyet a posztkommunista pártnak a választásokon aratott győzelme jellemez (Tworzecki 1994; Fitzmaurice 1995; Szelényi et al. 1997). Ez a hangulatváltás azon értékek preferenciájának változásában mutatkozik meg, amelyek szorosan kapcsolódnak a kommunista rendszer melletti vagy elleni érzelmekhez. És végül, a *negyedik hipotézisünk* szerint – eltekintve a materializmus és a hedonizmus általános irányzatától – a térségben 1989 után nem jelentkezett új politikai eszme (Vachudová és Snyder 1997, 1). Koherens értékrendszer sem jelenik meg, ami az előző értékpreferencia-mintákhoz viszonyítva visszatérést vagy megújulást jelentene. Feltételezzük, hogy a materializmus növekedése a célértékekben nem jár együtt a pragmatikus eszközértékek hasonló preferencianövekedésével.

[6] Korábbi, a projekthez kapcsolódó kiadványok: Hankiss (1982, 1986), Szakolczai (1982, 1987), Hankiss, Manchin és Füstös (1990).

19.4. Operacionalizált hipotézisek

Az általános elméleti szempontok alapján a különböző értékcsoportok vagy egyedi értékek változásával kapcsolatosan tíz konkrét hipotézist fogalmazunk meg.

1. Hipotézis. A klasszikus szocialista értékeket illetően egyszerű a hipotézis: radikális és visszafordíthatatlan törésnek kell lenni az 1989-es és az azt követő átlagos preferenciákban. Ez mindhárom szociáldemokrata érték, de elsősorban a MUNKA és TÁRSADALMI MEGBECSÜLÉS esetén nyilvánvaló, míg az ideológiai értékek esetén az EGYENLŐSÉGre érvényes ugyanez, ami szemantikailag legközelebb áll ehhez a csoporthoz.

2. Hipotézis. Számítunk arra, hogy a hivatalos szocialista értékek – elsősorban a BÉKE és a HAZA BIZTONSÁGA – szerepe csökken. 1982 és 1990 között a törés a klasszikus szocialista értékekhez hasonlóan nyilvánul meg, esetleg kisebb mértékben. Jelentős visszatérésre számítunk 1990 és 1993 között, ami az 1990 előtti szintet is elérheti. A posztkommunista mozgalom diadalútja valószínűleg – amint a régió többi országaiban is – 1996-ra véget ér Magyarországon. Az 1990-es szintre való visszatérés várható tehát 1996-ra, ami 1997–98-ban is megmarad.

3. Hipotézis. Feltételezzük, hogy kimutatható egy azonos minta (séma), de ellentétes irányú változással, szabad utat engedve eltérő, „ellenzéki" értékeknek. A CSALÁD, a BELSŐ HARMÓNIA és a SZERELEM iránti preferenciák 1990-ben megnőnek, majd 1993-ban csökkennek. 1996-ban ismét megnőnek, és 1997–1998-ban is megtartják ezt a szintet.

4. Hipotézis. Negyedik általános hipotézisünkkel kapcsolatban feltételezzük, hogy a materializmus és a hedonizmus általános irányzatával egybevágóan az ehhez kapcsolódó értékeknek, a KELLEMES, ÉLVEZETES ÉLETnek és az ANYAGI JÓLÉTnek 1989 után növekvő szerepe lesz. Ugyanakkor a szemantikailag rokon pragmatikus érték, a TÖREKVŐ esetében nem számítunk ugyanolyan átlagos preferencianövekedésre. Ami a hedonisztikus értékeket, a BOLDOGSÁG, ÉRDEKES, VÁLTOZATOS ÉLETet és a JÓ KEDÉLYŰt illeti, hasonló változásokra számítunk, mint a materialista értékek esetén. Ez a két csoport szemantikai közelségének is köszönhető.

5. Hipotézis. Az 1977–1978-as magyar és az 1968-as amerikai értékpreferenciák közötti legmegdöbbentőbb különbség az intellektuális értékekben volt. Érvelhetnénk azzal, hogy a változást követően az ilyen nagyméretű különbségek csökkenése várható, lényegesebb változásra azonban mégsem számítunk. Két alhipotézist is megfogalmazunk. Az első alhipotézis szerint a hivatalos szocialista és az intellektuális értékek közötti szoros, az előző rendszer „meghatározó hatásának" („stamping effect") köszönhető kapcsolatra vonatkozik. Eszerint a posztkommunista nosztalgia kapcsán az intellektuális értékek preferenciájának növekedése is várható lenne 1993-ra. A második alhipotézis pedig azt tételezi, hogy nem számítunk hasonló változásra a szemantikailag rokon HATÉKONY pragmatikus értékben.

6. Hipotézis. A magyar és az amerikai értékpreferenciákat illetően a legnagyobb, de kevésbé meglepő különbség a vallásos értékekben mutatkozott. Magyarországon 1989 után – az 1989-es első (poszt-) kormány törekvései ellenére – alig nőtt a vallásosság szerepe. Ez az eredmény a negyedik általános hipotézisünkből is következik, mely szerint az 1989-es értékekben a materializmus és a hedonizmus az egyetlen új változás. Ezért legjobb esetben csak gyenge növekedés várható az ÜDVÖZÜLÉS átlagpreferenciájában; 1993-ra érezhető csökkenéssel.

7. Hipotézis. Teljesen eltérő mintára számítunk az eszközértékek egyik szemantikailag közeli csoportjában, az interperszonális kapcsolatok és a közösségi élet területén. Figyelmen kívül hagyva a MEGBOCSÁTÓ értéket, ide tartoznak a SZERETETTELJES és a SEGÍTŐKÉSZ értékek, amelyek az intellektuális értékekkel szemantikailag ellentétesek. Ezeknek az értékeknek az Egyesült Államokban fontosabb szerepet tulajdonítottak, mint Magyarországon, kivéve a szocialista-kollektivista mellékértelemmel is rendelkező SEGÍTŐKÉSZ értéket, amely lényegesen fontosabb Magyarországon. Általános hipotéziseink tükrében Magyarországon 1989 után nem számítunk ezen értékek átlagos preferenciájának emelkedésére. Ellenkezőleg; a szocialista értékek csökkenésével hazánkban a SEGÍTŐKÉSZ érték is veszít jelentőségéből.

8. Hipotézis. Az eszközértékeknek az intellektuális értékekkel szemantikailag ellentétes másik csoportja az ENGEDELMES, TISZTA és UDVARIAS. Ezek hagyományosfegyelmező értékek, szubsztanciális tartalmuk kisebb, mint az interperszonális és közösségi értékeké. Feltételezzük tehát, hogy nem lesz meghatározott irány a preferenciák változásában, amelyek – ha előfordulnak – ellentétesen tükrözik az intellektuális értékekben mutatkozó oszcillációkat.

9. Hipotézis. Kontrollhipotézis, amellyel eredményeink általános érvényességét akarjuk igazolni. Míg legtöbb hipotézisünk változásokat és diszkontinuitásokat jósolt, addig egyesek a kommunista minta folytonosságát feltételezték. Van azonban egy különálló értékcsoport, amelyben folytonosságra számítunk anélkül, hogy a kommunista örökséget idéznék fel. Ezek az etikus-sztoikus értékek: a BÁTOR, az SZAVAHIHETŐ, a FELELŐSSÉGTELJES, a FEGYELMEZETT és a BARÁTSÁG. A mai társadalmakban annyira általánosan elfogadottak – különösen a SZAVAHIHETŐ és a FELELŐSSÉGTELJES központi morális értékek –, hogy preferenciájukban nem számítunk egyik vizsgálati évben sem szignifikáns változásra. Preferenciájuk relatív stabilitása tehát az alkalmazott módszer megbízhatóságát bizonyítja.

10. Hipotézis. És végül egy jól meghatározott értékcsoport; a kontemplatív személyiségértékek (a BÖLCSESSÉG, a SZÉPSÉG VILÁGA és a BELSŐ HARMÓNIA), amelyek inkább esztétikai, mint etikai elemeket tartalmaznak, és a saját személyiséghez tartoznak. Ezek esetében nem lehet az általános hipotéziseken alapuló koherens hipotézist kidolgozni. Feltételezzük tehát, hogy ha egyáltalán előfordul bennük változás, az olyan lesz, ami a többi csoporttal ellentétben nem egy általános minta alapján történik, hanem követi a szemantikailag közelebb eső értékek mintáját. A BELSŐ HARMÓNIA tehát mint ellenzéki érték viselkedik; a SZÉPSÉG VILÁGA a hedonista értékekre jellemző változások szerint módosul, míg a BÖLCSESSÉG az intellektuális értékekkel együtt mozdul el.

19.5. Eredmények

A hipotézisek kiértékeléséhez használt eredmények többsége a 19.2. táblázatban látható, ami egyrészt a hat vizsgálati év 36 értékének átlagos értékpreferenciáját tartalmazza. Másrészt az átlagok között az egy, két vagy három csillag a kétoldali t-próba szerint jelzi, hogy a két év közötti preferenciaváltozások a 0,05, a 0,01 vagy a 0,001 valószínűségi szinten szignifikánsak. A varianciaelemzés szignifikanciaszintjei a hat vizsgálati

évre szintén megtalálhatók. Végül a táblázat utolsó oszlopa a közel húszéves időszak változásainak linearitását mutatja. A 19.3. és 19.4. ábra a cél- és eszközértékekre vonatkozó alapadatokat tartalmazza.

A következőkben előbb az operacionalizált hipotéziseket értékeljük ki, és az általános hipotéziseket később vizsgáljuk.

1. Hipotézis. Az első hipotézist, amely szerint az 1989-es év minden klasszikus szocialista érték számára egyedülálló és tartós törést jelentett, adataink egyértelműen bizonyítják. A klasszikus szocialista vagy szociáldemokrata értékcsoport mindhárom értékének esetében (MUNKA, TÁRSADALMI MEGBECSÜLÉS és EMBERI ÖNÉRZET) minden posztkommunista évben alacsonyabb az átlagos preferencia, mint a kommunista időszakban, és majdnem minden esetben lényeges a különbség. Ugyanaz érvényes a hivatalos vagy ideológiai szocialista értékcsoport legklasszikusabb szocialista értékének, az EGYENLŐSÉGnek az esetében is. Az eredmények részben még ennél is világosabbak. A legnagyobb törést mutató értékek (TÁRSADALMI MEGBECSÜLÉS és EGYENLŐSÉG) azok, amelyek a két különböző értékcsoportból szemantikailag a legközelebb álltak egymáshoz. Így tehát megállapíthatjuk, hogy a szocialista értékrendszer hanyatlása a klasszikus és hivatalos szocialista értékek metszőpontjában a legerősebb.

2. Hipotézis. A kommunista rendszer két legfontosabb célértékének (a HAZA BIZTONSÁGA és a BÉKE) változása teljes mértékben a várt minta szerint alakult. Preferenciájuk 1977–1978-ban és 1982-ben, a kommunista időszakban, valamint a posztkommunista nosztalgia idején, 1993-ban volt a legmagasabb. 1990-ben jelentőségük lényegesen csökkent, és az 1996–1997–1998-as átlagos preferenciájuk erre a szintre tért vissza. Helytálló tehát az a hipotézis, mely szerint 1993–94-ben csak időszakos visszatérésről volt szó, és hogy az előző rendszerhez kapcsolódó értékek preferenciaszintjének 1989 előtt és után nincs kontinuitása.

3. Hipotézis. A privátszféra korábban „ellenzéki" értékeiről szóló „iker"hipotézist is majdnem mértani pontossággal erősítik meg az adatok. Minden alkalommal, amikor a hivatalos szocialista értékek háttérbe szorulnak, a CSALÁD, a BELSŐ HARMÓNIA és a SZERELEM kerülnek előtérbe, és ha az előzők szerepe nő, akkor az utóbbiaké gyengül (lásd 19.5. ábra). Ha csupán a számok változását nézzük, párhuzamosságot tapasztalunk. A változások magyarázatánál azonban óvatosabban kell fogalmaznunk. A CSALÁD, a BELSŐ HARMÓNIA és a SZERELEM közötti szemantikai kapcsolat meglehetősen gyenge. Az amerikai mintában egyszer sem tartoztak egy faktor ugyanazon pólusához (Rokeach 1973, 47). Csak Magyarországon kerültek egymás mellé a szocialista értékekkel szembeni együttes oppozíciójukkal. Relatív fontosságuk a preferenciákban a hivatalos szocialista értékek elmozdulását követi és másolja. A változások mögötti latens dimenzió – akár jelentőségük növekedéséről, akár csökkenéséről van szó – kizárólag a szocialista értékekkel való kapcsolódásukban keresendő.

4. Hipotézis. Az adatok negyedik operacionalizált hipotézisünket is igazolják. A materializmus és hedonizmus két legreprezentatívabb értékét (ANYAGI JÓLÉT és BOLDOGSÁG) illetően 1989 egyszer és mindenkorra változást jelentett. Ez nem csupán a negyedik hipotézist, hanem az első négy hipotézis alapját alkotó teljes geometriai mintát is igazolja. Amint a 19.5. és 19.6. ábrán látható, a klasszikus szocialista értékek szerepének meghatározó csökkenésével szemben az alapvető materialista és hedonista értékek szerepe erősödik, míg a hivatalos, szocialista értékek zegzugos pályagörbéje az „ellenzéki" privát értékek ellentétes változásaiban tükröződik.

A hipotézis második részét, amely az első általános hipotézissel együtt az anyagi jóléthez legközelebb álló eszközértékek megfelelő változásának hiányát vallotta, szintén igazolják az adatok. A TÖREKVŐ az az eszközérték, amely szemantikailag és az amerikai adatelemzés szerint is legközelebb áll az ANYAGI JÓLÉT célértékhez, nem kapott jelentősebb szerepet a posztkommunista években sem.

Az adatok a változás egy további jellegzetességét is feltárták, amelyet pedig az általános hipotézisek nem jósoltak meg. A kevésbé a jólét és elégedettség (ANYAGI JÓLÉT, BOLDOGSÁG) állapotához, mint inkább egy aktív élet öröméhez (KELLEMES, ÉLVEZETES ÉLET; ÉRDEKES, VÁLTOZATOS ÉLET és SZÉPSÉG) kapcsolódó hedonista értékek maradtak többé-kevésbé stabilak. Megmaradt tehát, mint egy elkövetkezendő boldog állapot mentalitása, a jövő képzete, bár a szocialista értékek helyett ez most a kényelemmel és az anyagi jóléttel társult. Ennek az állapotnak a megvalósításához szükséges eszközértékek most sem voltak meghatározva.

Tovább nő e változás fontossága, ha a két, szemantikailag idetartozó két értékben (TÖREKVŐ, MUNKA) bekövetkező preferenciavesztéssel hasonlítjuk össze. Az 1996–1997-es változás ezeket az értékeket érintette a legdrámaibban. Ez alatt az egy év alatt a MUNKA átlagos rangsorolása 1,5 pontot veszített. Ez önmagában is hatalmas veszteség, és még drasztikusabbnak tűnik, ha figyelembe vesszük, hogy az 1996 és 1997-es változások egyetlenegy cél- és eszközérték esetében sem haladták meg a 0,5 pontot, és hogy az 1989-es szinthez viszonyítva a legjelentősebb csökkenést mutatta. Míg tehát a kommunista időszakban a MUNKA az ötödik legfontosabb célérték volt, 1997-re, 1977–1978-hoz viszonyítva majdnem lineárisan 3,5 átlagos rangpontot veszített, és 1997-re a 18 célérték közül egyike lett a legkevésbé fontos értékeknek (a rangsor 14. értéke a 18. érték között).

Érvelhetnénk azzal, hogy ez csupán válasz a szocialista rezsimmel, valamint a munkának és termelésnek a fogyasztással ellentétes túlzott jelentőségével szemben. Ez csak részben igaz. Valódi relevanciája igazából Max Weber szociológiájának szemszögéből érzékelhető, különösen az „innerwordly asceticism" gondolatai kapcsán (Weber 1978, 542). Véleménye szerint az aszketikus vallások, életvezetések a változás bizonytalan periódusában jelennek meg, amikor belső stabilitást és megerősödést nyújtanak az egyénnek, és társadalmi szinten is elősegítik a konszolidációt. Ez elsősorban a modern kapitalista piac kialakulásának esetében igaz. A MUNKA értékének drasztikus csökkenése, amint azt az adatok is bizonyítják, különösen élesen illusztrálja a posztkommunizmus nehézségeit. A probléma nemcsak a múlt leküzdésére, hanem a jelen sikerességére is kihat. A kommunista rendszer nemcsak a szocialista értékeket kompromittálta, hanem azokat is, amelyek alapvetően fontosak a kommunizmus utáni újjáépítéshez. A MUNKA értékének 1997-es drámai változása 1998-ra visszafordul, és a legnagyobb mértékű pozitív változást produkálja. Ezt a változást erősíti a TÖREKVŐ eszközérték – bár lényegesen csekélyebb – növekedése.

5. Hipotézis. Láttuk a magyar és amerikai társadalom értékpreferenciái közötti különbséget, és az intellektuális és hivatalos szocialista értékek közötti kapcsolatot. Ezekben az értékekben egy esetleges elmozdulás érdekes a kontinuitás és diszkontinuitás általános hipotézisének szemszögéből. Az eredmények egyértelműen bizonyítják a kontinuitás elméletét. A LOGIKUS és ÉRTELMES érték a visszaesés helyett 1989 után fontosabbá vált; a magyar értékrendszer magas fokú intellektualizációja 1989 után változatlanul tovább folytatódott.

Az adatok a két kisebb hipotézist is igazolták. Valóban 1993 volt az az év, amikor az intellektuális értékek rendkívül magas preferenciájáról beszélhetünk. A magyar érték-

rendszer intellektualizációja olyan magas volt ebben az évben, hogy az ÉRTELMES érték az egyetlen legfontosabb eszközértékké vált hazánkban, még az alapvető etikai értékeknél (SZAVAHIHETŐ, FELELŐSSÉGTELJES, BÁTOR) is fontosabbá vált, míg a LOGIKUS a hetedik volt a rangsorban (az amerikai mintában az ÉRTELMES és a LOGIKUS sorrendben a 15. és 17. helyen állt). 1996-ban, a korábbi rendszer iránti nosztalgia megszűnésekor, mindhárom intellektuális érték az előző helyére került, majd 1997-re a LOGIKUS és az ÉRTELMES értéknél is ismét preferencianövekedés volt észlelhető, de 1998-ra az ÉRTELMES az 1996-os, a LOGIKUS az 1990-es szintre esett vissza. E változás a hedonizmus növekedésével járt együtt. Változatlan maradt azonban 1997-ig a HATÉKONY preferenciája, 1998-ban azonban csökkent, és így az utolsó lett az eszközértékek között.

6. Hipotézis. E hipotézis eredményei nyilvánvalóak. Míg más kelet-európai országban végzett vizsgálatok a vallás fontosságának látványos növekedéséről számolnak be, mint például Oroszország esetében (Greely 1994), Magyarországon hasonló eredmény nem mutatható ki. Az ÜDVÖZÜLÉS még mindig a legutolsó helyen álló célérték Magyarországon, és 1989 után sem történt változás az 1982-es szinthez viszonyítva.

7. Hipotézis. A kontinuitás és diszkontinuitás keveredése a négy általános hipotézisből ered, és az adatok most is ezt igazolják. A MEGBOCSÁTÓ és SZERETETTELJES interperszonális értékek alacsony preferenciája 1989 után is változatlan maradt. Tulajdonképpen a MEGBOCSÁTÓ azon kevés érték közé tartozik, amelyben az ANOVA-teszt szerint lényeges elmozdulás nem történt az elmúlt húsz évben. Ez részben a SZERETETTELJES értékre is vonatkozik, ha az 1982-es évet outlier esetnek tekintjük. A csoport harmadik értéke, a SEGÍTŐKÉSZ az egyetlen, amelyet Magyarországon inkább preferálnak, mint az Egyesült Államokban. Mivel azonban nem csupán vallásos és közösségi, hanem szocialista értékjelentéssel is bír, veszített fontosságából. A 18 eszközérték közül az 1977 és 1998 közötti periódusban a legkövetkezetesebben és közel lineárisan került hátrább a preferencia-sorrendben.

8. Hipotézis. Feltételeztük, hogy mivel az ENGEDELMES, a TISZTA és az UDVARIAS értékek preferálása a modern társadalom civilizációs folyamatának természetes eredményei, így jelentőségbeli változásaik a velük pozitív vagy negatív korrelációban lévő értékek változását követi. Az erre vonatkozó jelenlegi adatok nem tartalmaznak semmiféle ellentmondást. Preferenciájukban 1989 nem jelentett törést. Az 1982 és 1990 közötti változások voltak minden más szomszédos vizsgálati évhez viszonyítva a legkisebbek. Másrészt 1989 után évről évre történő elmozdulásuk (mint azt az 19.7. ábra is mutatja) szigorúan követte szemantikai ellentétük, az intellektuális értékek preferenciájának oszcillációit. Elmozdulások, amelyek önmagukban a pozitív és negatív kommunista örökségre egyszerűsödtek: 1993-ban a múlt iránti nosztalgia, valamint a materializmus és a hedonizmus kétlépcsős megerősödése 1990-ben és 1997-ben.

9. Hipotézis. Ebben a kontrollhipotézisben feltételezzük, hogy a változások és a kommunista örökség közepette is számos, minden modern társadalomhoz hozzátartozó érték (BÁTOR, FEGYELMEZETT, BARÁTSÁG, SZAVAHIHETŐ, FELELŐSSÉGTELJES) esetében kontinuitás áll fenn. Ez a hipotézis is beigazolódott (lásd 19.8. ábra). A húsz év alatt az öt stabil érték közül három ehhez a csoporthoz tartozott, beleértve a két legfontosabb etikai értéket, a SZAVAHIHETŐt és a FELELŐSSÉGTELJESt. A FEGYELMEZETT esetében csupán 1982 outlier év, s a BÁTOR preferenciájában sem tapasztalható változás 1977–78-tól 1993-ig. Megállapíthatjuk tehát, hogy ami a morális értékeket illeti, sem a kommunista rendszer, sem összeomlása nem hagyott nyomot

Magyarország értékrendszerében. Elvárásainkat igazolva a kommunista örökség nem a tudatosság felszínén, a világos életcélokban vagy erkölcsi alapokban kereshető, hanem az életvezetés nyilvánvaló struktúráiban vagy a viselkedésben.

10. Hipotézis. A személyiségértékek szemantikailag homogén csoportja – a kontemplatív személyiségértékek (BÖLCSESSÉG, A SZÉPSÉG VILÁGA, BELSŐ HARMÓNIA) – a szocialista, a materialista és – kisebb mértékben – a hedonista értékekkel szemben egy alternatív, független életstílus lényegét képviselhette volna, ellenállva az elmúlt rendszer direkt és indirekt hatásainak. Feltételeztük azonban, hogy ez nem következett be, és egy-egy érték változása csak a vele rokon értékcsoport elmozdulásait tükrözi. Ez a hipotézis egy jelentős kivétellel beigazolódott. A BELSŐ HARMÓNIA módosult leginkább az elvárások szerint, követte a többi „ellenzéki" érték oszcillációit. A SZÉPSÉG preferenciája nem mozdult el együtt a BELSŐ HARMÓNIÁval, mivel 1990-ben szignifikáns, bár nem túl nagy és inkább rejtélyes csökkenése volt észlelhető. Preferenciájában a másik nyilvánvaló változás 1996 és 1997 között történt, amikor a hedonizmus irányába való általános elmozdulás részeként a SZÉPSÉG szerepe is nőtt.

A csoport harmadik értéke, a BÖLCSESSÉG külön minta szerint alakult. Ez fontos eredmény, és talán a legígéretesebb jel is. Ez a három közül Magyarországon leginkább elhanyagolt érték fejezte ki a magyar és amerikai értékpreferenciák közötti legnagyobb különbséget. Számos alapvető lineáris elmozdulás következtében 1997-re a BÖLCSESSÉG két pontot nyert a rangsorban (a legnagyobb változás a 18 célérték változása közül). Ilyen preferencianövekedés különösen az előző rendszer paternalizmusának és az általa létrehozott infantilizmus fényében fontos (Hankiss 1982).

A BÖLCSESSÉG preferálásának rendületlen előmenetelét ábrázolhatjuk a MUNKA hanyatlásának ellentétében. Amint azt az 19.9. ábra mutatja, az értékek egyikének preferenciájában mutatkozó legcsekélyebb elmozdulást tökéletesen letükrözi a másik ellentétes irányú, azonos mértékű elmozdulása. Az ilyenfajta ellentét vagy verseny a két érték között annál is inkább szokatlan, mivel az amerikai minta első főkomponensében együtt jelennek meg ugyanazon a póluson. Így azt mondhatjuk, hogy a BÖLCSESSÉG fontosságának látványos növekedése Magyarországon nagymértékben a „munkaetika" hanyatlásának kompenzálása.

19.6. Általános hipotézis

Mivel az adatok a tíz operacionalizált hipotézis majdnem mindegyikét alátámasztották, nem meglepő, hogy a négy általános hipotézis – amelyekre az előzőek épültek – szintén megfelelt elvárásainknak. Az 19.3. és 19.4. ábrán látható, hogy az 1989 előtti és utáni évek változásai a célértékeknél sokkal számottevőbbek voltak, mint az eszközértékeknél. Az utóbbiakban inkább oszcillációk jelentek meg, míg a célértékek preferenciája mindig meghatározott irányban változott. Ha az 1977 és 1997 közötti változásokat összehasonlítjuk, akkor az átlag azonosságára vonatkozó hipotézist a 18 célérték közül 14 esetben 0,01-es szignifikanciaszinten el kell vetnünk. Az eszközértékeknél 18-ból csak 7 esetben.

A másik három általános hipotézist illetően csak ismétlésbe bocsátkozhatnánk, mivel minden egyes, a szocialista rendszerrel pozitív vagy negatív értelemben asszociált érték

számára az 1989-es év törést jelentett. Ez érvényes minden, a rendszer hivatalos ideológiájához kapcsolódó értékre, bár 1993-ban kisebb mértékű és időleges visszatéréssel. A materializmus és a hedonizmus növekedése bizonyult az egyetlen új fejleménynek, amely a szocialista értékek veszteségét ellensúlyozta. A BÖLCSESSÉG kivételével nem volt olyan érték, amelynek változása lényeges és előre meg nem jósolt lett volna.

19.7. Következtetések

A fejezet húsz év értékpreferencia-változásainak általános és specifikus hipotézisei sorozatát igen bonyolult és összetett értékteszt, a Rokeach-teszt segítségével vizsgálta. Eddig a hipotézisek többsége beigazolódott, és a 36 alapvető humán értékben bekövetkezett majdhogynem minden számottevő változást tartalmazza a hét különböző időpontban. Ilyen magas fokú előreláthatóság már egy kicsit gyanúsnak is tűnhet.

Érvelhetünk azzal, hogy a magyarázat nem az adatoktól, a módszerektől vagy egy kezdetleges post hoc elemzéstől függ. Inkább arról van szó, hogy a változások alapvetően nem vezethetők le sem tíz operacionalizált, sem négy általános hipotézisből, hanem egyetlen, mindent átívelő állításból, amely egységben szemléli a folytonosságot és a szakadást. Az elmúlt húsz év magyarországi értékpreferenciájában egyetlenegy változás történt, amelyből minden egyedi értékpreferencia-változás levezethető, és ez nem más, mint a szocialista rezsim összeomlása és ezzel együtt a kommunisták hatalmának felbomlása.

A szocialista értékrendszernek két különböző, bár egymásba kapcsolódó összetevője volt; a klasszikus és a hivatalos szocialista értékek. 1990-re a mindkét csoport iránti preferencia lényegesen csökkent. Ez egyértelműen a létező szocializmus összeomlásának jele volt. Ez a jelenség meghatározta azonban az ellenzéki mozgalmak irányát és módját is. A klasszikus szocialista értékek ellentétei a materializmus és a hedonizmus értékei voltak, amelyeknek fontossága 1990-ben valóban rögtön megnőtt. Ha azonban e változás iránya a szocializmussal ellentétben állt, akkor jellegét a rendszer összeomlása határozta meg. Ahogyan a szocializmus egy boldog állapotot ígért a fényes kommunista jövőben, a változásokat is hasonló csodaszerekbe vetett naiv hit követte: jólét, kényelmes élet, boldogság és elégedettség. A hivatalos szocialista értékek szembekerültek egy sor „ellenzéki" értékkel, amelyeknek szerepe 1990-ben hasonlóan nőtt.

Az 1990 és 1993 közötti változásokat az előző rendszer iránti ideiglenes nosztalgiának tulajdonították. Ez nem jelentett változást az előbb vizsgált komponensben, a klasszikus szocialista és a hedonista materializmus közötti különbségben, hanem a második komponensben, a hivatalos szocialista és „ellenzéki" értékekben egy közel teljes értékű visszatérést igazol az 1989 előtti preferenciák szintjére. Az előző rendszer egy másik indirekt hatásaként ezt követte az eszközértékek erős intellektualizálódása, amelyet az intellektuális értékek ellentétes pólusán lévő hagyományos értékek ellensúlyoztak.

Ez azt jelenti, hogy 1996-ra a kommunista és antikommunista értékcsoport második komponenséhez tartozó értékek visszatértek az 1990-es szintre. Noha az 1990-es és 1996-os évek átlagainak hasonlósága meglepő, alig van különbség közöttük, ugyanakkor ez azt is jelenti, hogy 1996-ra a kétpólusú szocialista értékcsoportok még mindig befolyásolták az értékpreferenciák változásait.

Még 1997-ben is kevés változás történt e téren. Az egyetlen jelentős különbség, hogy a hedonizmus megerősödése és a munkaetika fontosságának további csökkenése

megerősíti a korábbi, 1990-es trendet anélkül, hogy világosan megmutatkozna egy új értékrend.

A reménysugár – legalábbis adataink szerint – a BÖLCSESSÉG szerepének látványos megerősödése. A BÖLCSESSÉG pedig az az érték, amelyre nagy szükség van a posztkommunista országokban.

19.8. Táblázatok, ábrák

A Rokeach-féle értékteszt

A. Célértékek

1. ANYAGI JÓLÉT (jómód, bőség)
2. BÉKE (háborúktól és konfliktusoktól mentes világ)
3. BOLDOGSÁG (megelégedettség)
4. BÖLCSESSÉG (életbölcsesség)
5. CSALÁDI BIZTONSÁG (szeretteinkről való gondoskodás)
6. BELSŐ HARMÓNIA (belső feszültségektől mentes élet)
7. EGYENLŐSÉG (testvériség, mindenki számára azonos lehetőségek)
8. AZ ELVÉGZETT MUNKA ÖRÖME (teljesítményekben gazdag, aktív élet)
9. ÉRDEKES, VÁLTOZATOS ÉLET (élményekben gazdag, aktív élet)
10. A HAZA BIZTONSÁGA (külső támadásokkal szembeni védettség)
11. IGAZI BARÁTSÁG (szoros emberi kapcsolat)
12. IGAZI SZERELEM (meghitt lelki és testi kapcsolat)
13. KELLEMES, ÉLVEZETES ÉLET (örömök, sok szabadidő)
14. EMBERI ÖNÉRZET (öntudat, önbecsülés)
15. SZABADSÁG (függetlenség, a választás lehetősége)
16. A SZÉPSÉG VILÁGA (a természet és műalkotások szépsége)
17. TÁRSADALMI MEGBECSÜLÉS (elismerés, tisztelet)
18. ÜDVÖZÜLÉS (megváltás, örök élet)

B. Eszközértékek

19. ALKOTÓ SZELLEMŰ (újító, eredeti gondolkodású)
20. BÁTOR, GERINCES (kiáll a nézeteiért)
21. ELŐÍTÉLETEKTŐL MENTES (elfogulatlan, nyílt gondolkodású)
22. ENGEDELMES (kötelességtudó, tisztelettudó)
23. ÉRTELMES (gondolkodó, intelligens)
24. FEGYELMEZETT (önuralommal rendelkező)
25. FELELŐSSÉGTELJES (megbízható, felelősségtudó)
26. HATÉKONY (hozzáértő, szakértő)
27. JÓ KEDÉLYŰ (vidám, könnyű szívű)
28. LOGIKUS GONDOLKODÁSÚ (racionális, ésszerű)
29. MEGBOCSÁTÓ (nem bosszúálló)
30. ÖNÁLLÓ (független, erős egyéniség)
31. SEGÍTŐKÉSZ (mások jólétéért dolgozik)
32. SZAVAHIHETŐ (becsületes, őszinte)
33. SZERETETTEL TELJES (ragaszkodó, gyöngéd)
34. TISZTA (rendes, ápolt)
35. TÖREKVŐ (szorgalmas, vinni akarja valamire)
36. UDVARIAS (jó modorú, jól nevelt)

Rokeach –féle értékteszt

Cél-értékek	**Terminal values**
1 ANYAGI JÓLÉT	A COMFORTABLE LIFE (a prosperous life) *(in Hungary: MATERIAL WELL-BEING)*
2 BÉKE	A WORLD OF PEACE (free of war and conflict)
3 BOLDOGSÁG	HAPPINESS (contentedness)
4 BÖLCSESSÉG	WISDOM (a mature understanding of life)
5 CSALÁDI BIZTONSÁG	FAMILY SECURITY (taking care of the loved ones)
6 BELSŐ HARMÓNIA	INNER HARMONY (freedom from inner conflict)
7 EGYENLŐSÉG	EQUALITY (brotherhood, equal opportunity for all)
8 MUNKA ÖRÖME	A SENSE OF ACCOMPLISHMENT (lasting contribution) *(in Hungary: THE SATISFACTION OF WELL-DONE WORK)*
9 VÁLTOZATOS ÉLET	AN EXCITING LIFE (a stimulating, active life)
10 A HAZA BIZTONSÁGA	NATIONAL SECURITY (protection from attack)
11 IGAZI BARÁTSÁG	TRUE FRIENDSHIP (close companionship)
12 IGAZI SZERELEM	MATURE LOVE (sexual and spiritual intimacy)
13 ÉLVEZETES ÉLET	PLEASURE (an enjoyable, leisurely life)
14 EMBERI ÖNÉRZET	SELF-RESPECT (self-esteem)
15 SZABADSÁG	FREEDOM (independence, free choice)
16 A SZÉPSÉG VILÁGA	A WORLD OF BEAUTY (beauty of nature and the art)
17 TÁRS MEGBECSÜLÉS	SOCIAL RECOGNITION (recpect, admiration)
18 ÜDVÖZÜLÉS	SALVATION (saved, eternal life)

Eszköz-értékek	**Instrumental values**
19 ALKOTÓ SZELLEMŰ	IMAGINATIV (daring, creative)
20 BÁTOR GERINCES	COURAGEOUS (standing up for your beliefs)
21 ELŐITÉLETEKTŐL MENT.	BROADMINDED (open-minded)
22 ENGEDELMES	OBEDIENT (dutiful, respectful)
23 ÉRTELMES	INTELLECTUAL (intelligent, reflexive)
24 FEGYELMEZETT	SELF-CONTROLLED (restrained, self-disciplined) *(in Hungary: DISCIPLINED)*
25 FELELŐSSÉGTELJES	RESPONSIBLE (dependable, reliable)
26 HATÉKONY	CAPABLE (competent, effective)
27 JÓKEDELYŰ	CHEERFUL (lighthearted, joyful)
28 LOGIKUS GONDOLKODÁSÚ	LOGICAL (consistent, rational)
29 MEGBOCSÁTÓ	FORGIVING (willing to pardon others)
30 ÖNÁLLÓ	INDEPENDENT (self-reliant, self-sufficient)
31 SEGÍTŐKÉSZ	HELPFUL (working for the welfare of others)
32 SZAVAHIHETŐ	HONEST (sencere, truthful)
33 SZERETETTEL TELJES	LOVING (affectionate, tender)
34 TISZTA	CLEAN (neat, tidy)
35 TÖREKVŐ	AMBITIOUS (hard-working, aspiring)
36 UDVARIAS	POLITE (sourteous, well-mannered)

- Megjegyzés: in three cases, the re-translation of the word used in the Hungarian version into English is also given.

19.1. táblázat. Az első főkomponens struktúrája

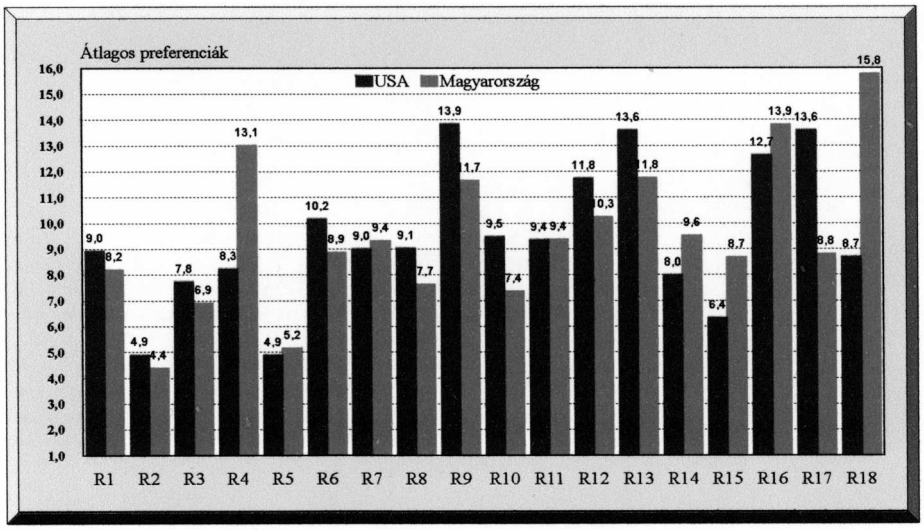
	U.S.A. 1968	HUNGARY 1978	HUNGARY 1982	HUNGARY 1990	HUNGARY 1993	HUNGARY 1996	HUNGARY 1997	HUNGARY 1998
ALKOTÓ SZELLEMŰ	.49	.58	.57	.63	.57	.54	.52	.61
LOGIKUS	.57	.47	.55	.60	.58	.51	.46	.51
HATÉKONY	.27	.35	.48	.56	.51	.46	.37	.33
ÉRTELMES	.52	.32	.31	.43	.35	.38	.34	.32
ELŐÍTÉLETEKTŐL MENTES	.20	.34	.47	.19	.29	.13	.21	.24
ÖNÁLLÓ	.37	.42	.21	.28	.30	.34	.27	.27
BÁTOR	.19	.33	.26	.21	.15	.20	.24	.20
FELELŐSSÉGTELJES	.23	.36	.37	.18	.34	.15	.06	.07
BÖLCSESSÉG	.31	.10	.OS	-.04	.11	.10	.01	.17
BELSŐ HARMÓNIA	.23	.01	.04	-.02	-.05	.23	.02	.11
MUNKA ÖRÖME	.46	.03	.14	.07	.07	.10	.O1	.05
SZABADSÁG	.29	.42	.24	.11	.17	-.16	-.01	-.14
EGYENLŐSÉG	.04	.23	.24	-.16	.04	-.31	-.19	-.30
HAZA BIZTONSÁGA	.06	.38	.26	-.13	.14	-.32	-.32	-.39
BÉKE	-.15	.33	.22	-.20	.07	-.36	-.33	-.40
TÁRSADALMI MEGB.	.05	.24	.13	.02	.19	-.17	-.21	-.20
ANYAGI JÓLÉT	-.29	-.30	-.24	.09	-.06	.11	.26	.13
ÉLVEZETES ÉLET	-.27	-.39	-.27	.06	-.14	.17	.41	.21
BOLDOGSÁG	-.34	-.40	-.31	-.07	-.24	.13	.01	.18
JÓKEDÉLYŰ	-.44	-.40	-.43	-.19	-.34	.00	-.03	.01
ÜDVÖZÜLÉS	-.24	-.26	-.33	-.41	-.31	-.41	-.41	-.32
SEGÍTŐKÉSZ	-.25	-.21	-.19	-.45	-.34	-.33	-.34	-.35
ENGEDELMES	-.39	-.45	-.42	-.45	-.34	-.42	-.39	-.44
TISZTA	-.55	-.48	-.48	-.41	-.44	-.38	-.24	-.33
UDVARIAS	-.49	-.46	-.49	-.43	-.54	-.47	-.36	-.45
SZERETETTEL TELJES	-.39	-.53	-.52	-.49	-.52	-.36	-.37	-.25
MEGBOCSÁTÓ	-.38	-.53	-.53	-.57	-.54	-.50	-.52	-.46
(sajátértékek	3.39	3.91	3.62	3.33	3.14	3.07	2.80	3.01)

19.2. táblázat. Az átlagos rangszámok változása

	1978	1982	1990	1993	1996	1997	1998	ANOVA	linearity
1. ANYAGI JÓLÉT	8.2	8.7***	6.8	6.5	6.5	6.8	6.8	.001	.001
2. BÉKE	4.4***	3.9***	4.9***	3.7***	4.8	4.9	5.1	.001	.001
3. BOLDOGSÁG	6.9***	7.6***	6.0	6.2	5.9	5.9	5.9	.001	.001
4. BÖLCSESSÉG	13.1*	12.7***	12.2***	11.3	11.4**	11.0	11.0	.001	.001
5. CSALÁDI BIZTONSÁG	5.2	5.3***	3.9***	4.6***	3.8	3.7	3.9	.001	.001
6. BELSŐ HARMÓNIA	8.9	8.7***	7.5***	8.4***	7.6	7.4	7.6	.001	.001
7. EGYENLŐSÉG	9.4	9.1***	11.2***	10.7	11.0	11.1	11.2	.001	.001
8. MUNKA ÖRÖME	7.7***	8.2***	8.7***	8.8	9.7***	11.2	9.4	.001	.001
9. VÁLTOZATOS ÉLET	11.7	11.9	11.7	11.9	12.0***	11.5	12.1	Note 1	n.s.
10. HAZA BIZTONSÁGA	7.4**	6.8***	8.1***	6.6***	8.0	7.9	8.2	.001	.001
11. IGAZI BARÁTSÁG	9.4**	8.9	9.2	9.1	9.0	9.1	9.0	n.s.	n.s.
12 IGAZI SZERELEM	10.3*	10.7***	9.7***	10.6**	10.0	10.3	10.8	.001	n.s.
13. ÉLVEZETES ÉLET	11.8	11.9	11.6*	12.0	11.9***	11.3	11.9	Note 1	Note 1
14. EMBERI ÖNÉRZET	9.6	9.6*	9.9***	10.5*	10.1	10.0	9.9	.001	.001
15. SZABADSÁG	8.7	8.8	9.1	9.0**	9.4	9.6	9.3	.001	.001
16. SZÉPSÉG VILÁGA	13.9	13.6**	14.2	14.0	14.1***	13.6	14.1	.001	n.s.
17. TÁRS. MEGBECSÜLÉS	8.8*	9.2***	11.1***	10.5	10.6	10.5	9.9	.001	.001
18. ÜDVÖZÜLÉS	15.8	15.5	15.4	15.7***	15.2	15.4	15.0	.001	.001
19. ALKOTÓ SZELLEMŰ	10.9	10.9	10.7**	10.1***	10.9	10.7	10.9	Note 2	n.s.
20. BÁTOR	7.2	7.4	7.3	7.4*	7.8	7.8	7.2	Note 3	Note 3
21. ELŐÍTÉLETMENTES	10.4*	9.9***	10.7	10.4	10.4	10.7	10.4	.001	.001
22. ENGEDELMES	10.9	10.5***	11.1*	11.6***	11.0	11.2	10.9	.001	.001
23. ÉRTELMES	8.3***	7.6	7.4***	6.4***	7.4	7.0	8.3	.001	.001
24. FEGYELMEZETT	9.2***	8.2***	9.1	9.0	9.3	9.0	9.2	Note 2	Note 2
25. FELELŐSSÉGTELJES	6.9*	7.3	7.2	7.2	7.1	6.8	6.9	n.s.	n.s.
26. HATÉKONY	11.4	11.6	11.4	11.2	11.3	11.3	11.4	n.s.	n.s.
27. JÓKEDÉLYŰ	9.9	10.3*	9.8*	10.2**	9.6	9.7	9.9	.001	.001
28. LOGIKUS	10.9***	10.3	10.2***	9.3***	10.9***	10.0	10.9	.001	n.s.
29. MEGBOCSÁTÓ	11.0	11.2	11.2	11.4	11.1	10.9	11.0	n.s.	n.s.
30. ÖNÁLLÓ	9.1***	10.1*	9.7	10.0**	9.4	9.1	9.1	.001	.001
31. SEGÍTŐKÉSZ	7.9***	8.7	9.0	9.2	9.1	9.3	7.9	.001	.001
32. SZAVAHIHETŐ	6.7	6.4	6.5	6.7	6.4	6.6	6.7	n.s.	n.s.
33. SZERETETTEL TELJES	10.2***	11.0***	10.0	10.2	10.3	10.1	10.2	Note 4	Note 4
34. TISZTA	10.0***	9.2	9.4	9.7***	8.8**	9.4	10.0	.001	n.s.
35. TÖREKVŐ	9.5***	10.8	10.5	10.8	10.8	11.2	9.5	.001	.001
36. UDVARIAS	10.8***	9.8	9.7***	10.4***	9.6**	10.2	10.8	.001	n.s.

- Alacsonyabb átlagos rangérték, magasabb a preferencia.
- A csillag az oszlopok között szignifikáns változást jelent a két év között.
  - * significant at the 0.05 level.
  - ** significant at the 0.01 level.
  - *** significant at the 0.001 level.
- ANOVA test indicates significance of overall change for the six observation years
  - n.s. means not significant.
  - Note 1: n.s. if 1997 is omitted.
  - Note 2: n.s. if 1993 is omitted.
  - Note 3: n.s. if 1996 and 1997 are omitted.
  - Note 4: n.s. if 1982 is omitted.
- linearity test indicates where the changes in between observations have a directionality, according to the ANOVA programme.

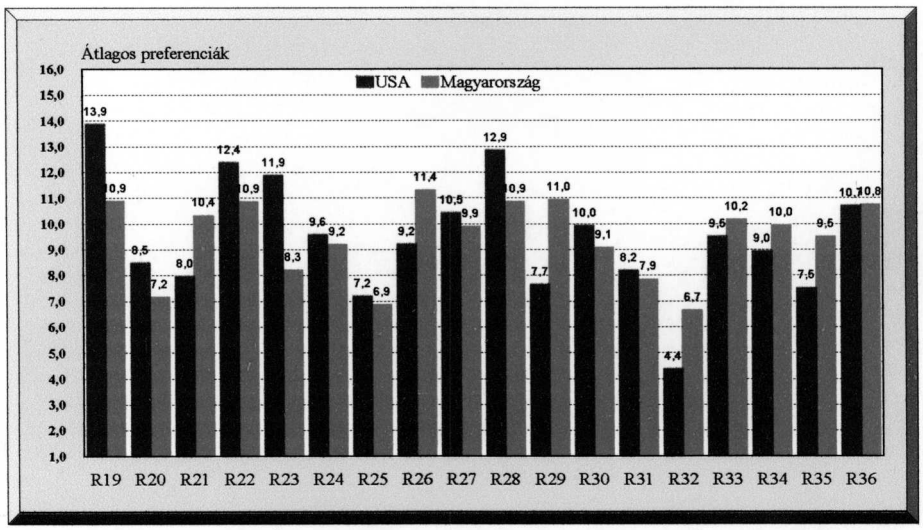
19.1. ábra. Rokeach-féle értékek
USA 1968, Magyarország 1977-1978
Célértékek



Alacsonyabb érték – magasabb preferencia

MTA Szociológiai Kutatóintézet **Értékszociológiai Műhely, 1999**

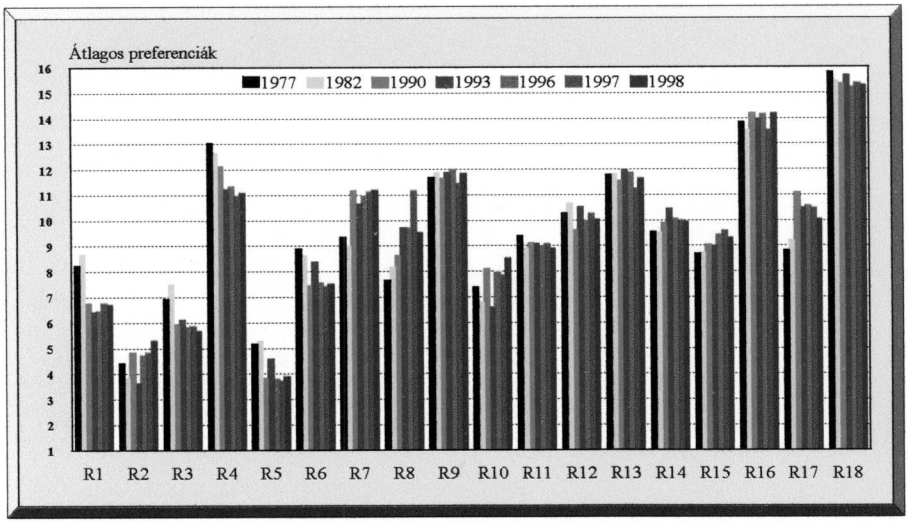
19.2. ábra. Rokeach-féle értékek
USA 1968, Magyarország 1977-1978
Eszközértékek



Alacsonyabb érték – magasabb preferencia

MTA Szociológiai Kutatóintézet **Értékszociológiai Műhely, 1999**

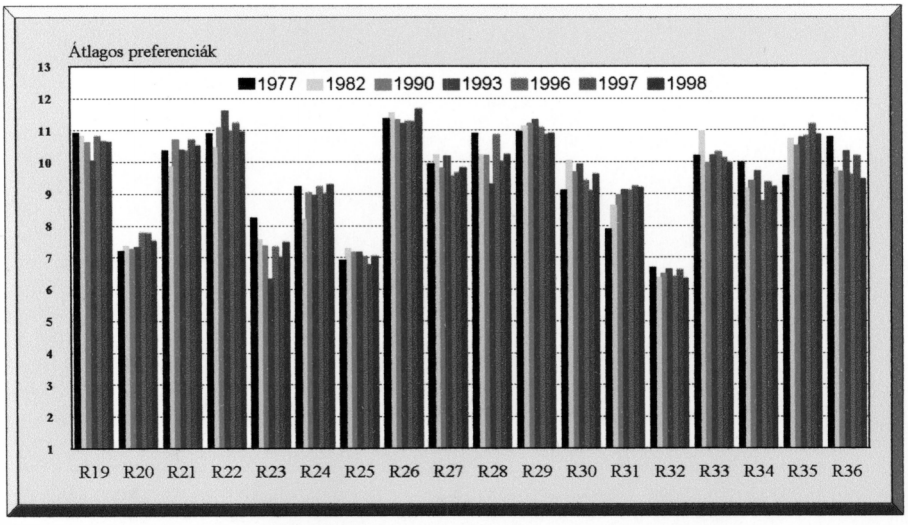
19.3. ábra. Rokeach-féle értékek, 1977–1998
Célértékek



Alacsonyabb érték – magasabb preferencia

MTA Szociológiai Kutatóintézet **Értékszociológiai Műhely, 1999**

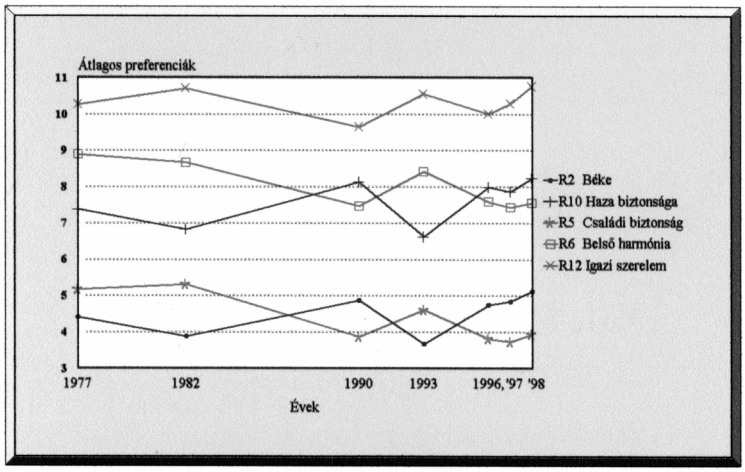
19.4. ábra. Rokeach-féle értékek, 1977–1998
Eszközértékek



Alacsonyabb érték – magasabb preferencia

MTA Szociológiai Kutatóintézet **Értékszociológiai Műhely, 1999**

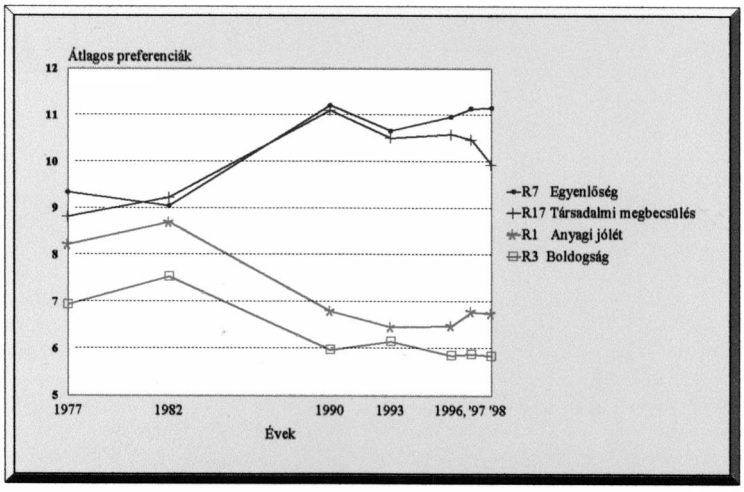
19.5. ábra. BÉKE, A HAZA BIZTONSÁGA
versus
CSALÁDI BIZTONSÁG, BELSŐ HARMÓNIA, IGAZI SZERELEM



Alacsonyabb érték – magasabb preferencia

MTA Szociológiai Kutatóintézet Értékszociológiai Műhely, 1999

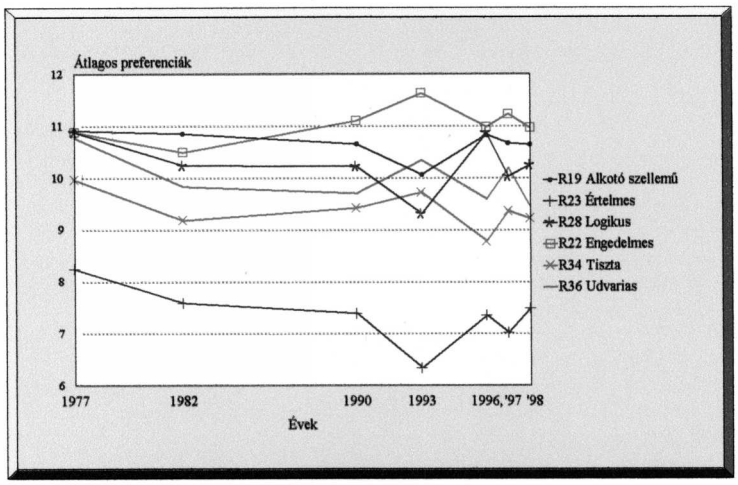
19.6. ábra. EGYENLŐSÉG, TÁRSADALMI MEGBECSÜLÉS
versus
ANYAGI JÓLÉT, BOLDOGSÁG



Alacsonyabb érték – magasabb preferencia

MTA Szociológiai Kutatóintézet Értékszociológiai Műhely, 1999

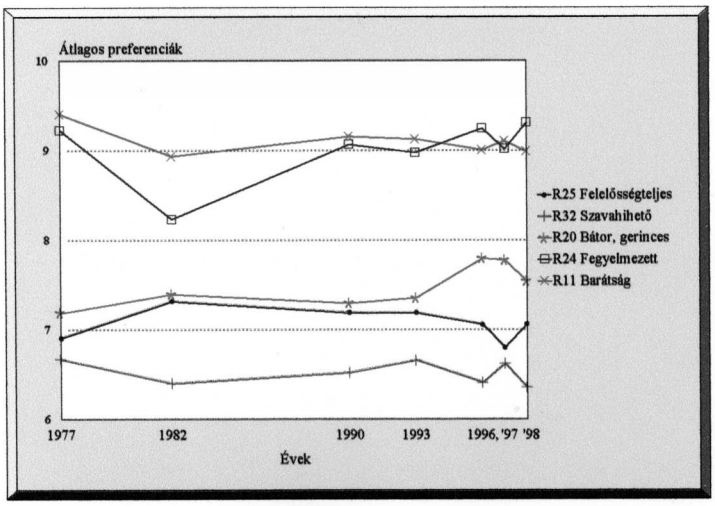
19.7 ábra. ALKOTÓ SZELLEMŰ, ÉRTELMES, LOGIKUS
versus
ENGEDELMES, TISZTA, UDVARIAS



Alacsonyabb érték – magasabb preferencia

MTA Szociológiai Kutatóintézet Értékszociológiai Műhely, 1999

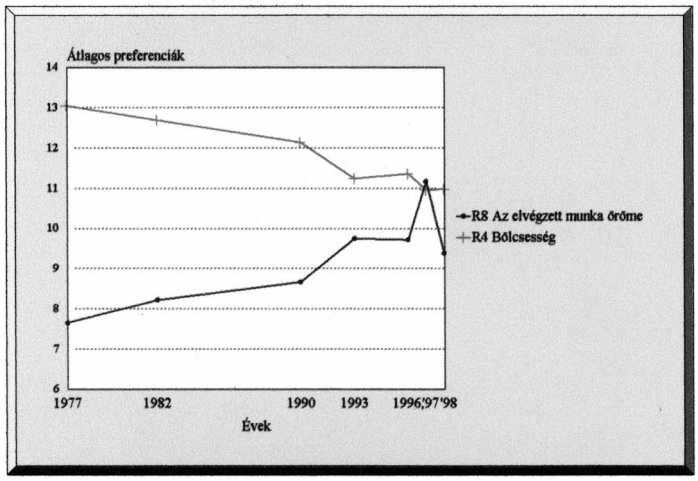
19.8. ábra. FELELŐSSÉGTELJES, SZAVAHIHETŐ,
BÁTOR, FEGYELMEZETT, BARÁTSÁG



Alacsonyabb érték – magasabb preferencia

MTA Szociológiai Kutatóintézet Értékszociológiai Műhely, 1999

19.9. **ábra. AZ ELVÉGZETT MUNKA ÖRÖME KONTRA BÖLCSESSÉG**



Alacsonyabb érték – magasabb preferencia

MTA Szociológiai Kutatóintézet Értékszociológiai Műhely, 1999

20. fejezet

A változó értékrendszer

A változó értékrendszer maga mögött hagyja az exkommunistákat)[1]

Bevezető

A Magyarországon végzett szociológiai felmérések azt mutatják, hogy a magyarországi értékrendszerben a kommunizmus összeomlása után bekövetkezett változások 1993-ra nagyrészt megfordultak, előre jelezvén az exkommunisták egy év múlva bekövetkezett választási győzelmét. A legújabb felmérésekből azonban kiderül, hogy a fordulat csupán átmeneti visszatérés volt a leépülő szocialista értékrendszerhez.

Miután a kommunista utódpártok 1993–1994-ben meglepő győzelmet arattak Közép- és Kelet-Európa három, legkevésbé szovjetbarát országában – Litvániában, Magyarországon és Lengyelországban –, újra és újra felvetődik a kérdés: vajon ezek a győzelmek a korábbi rezsimek gyakorlatához, sőt értékeihez való tartós visszatérést jelzik-e, avagy az egész csupán a kommunizmus összeomlásával kapcsolatos felfokozott várakozásokat kísérő csalódásokból fakadó átmeneti fordulat?

Ezekben az országokban a közvélemény-kutatások havonta mérik a politikai pártok népszerűségét, azonban ezen kutatások nem rendelkeznek 1989 előttre vonatkozó viszonyítási ponttal, és nem nyújtanak segítséget a változó politikai preferenciák mögötti értékváltozások megértéséhez. Ugyanakkor, ha mélyebben elemezzük őket, a Magyar Tudományos Akadémia Szociológiai Intézetének Hankiss Elemér vezette munkacsoportja által elvégzett, szám szerint öt csaknem húszéves időszakot átfogó országos reprezentatív felmérés alapot teremt annak a következtetésnek a megfogalmazásához, amely szerint az exkommunista pártok támogatottságának 1993–1994-es megugrása tulajdonképpen csak egy elhajlás volt. S bár az 1993-as magyarországi felmérés határozottan visszafordulást mutat az 1978-as és 1982-es vizsgálatokban feltárt értékpreferenciákhoz, 1996-ra egy másik döntő elmozdulás következett be, amely megismételte és megerősítette az 1990-ben bekövetkezett változásokat.

A felmérések és elemzések a Rokeach-féle értéktesztre támaszkodtak, amelyet névadója, Milton Rokeach szociálpszichológus használt először 1968-ban az Egyesült Államokban. Milton Rokeach értéktesztje az értékpreferenciákat mérő igényes, megbízható eszköz. A megkérdezetteknek 18 alapvető fontosságú emberi „célt” (ANYAGI JÓLÉT, BÉKE, BOLDOGSÁG, CSALÁDI BIZTONSÁG, IGAZI SZERELEM, SZABADSÁG, A SZÉPSÉG VILÁGA, TÁRSADALMI MEGBECSÜLÉS stb.) és az ezek elérését elősegítő 18 „eszközt” (ENGEDELMES, LOGIKUS GONDOLKODÁSÚ, SZAVAHIHETŐ,

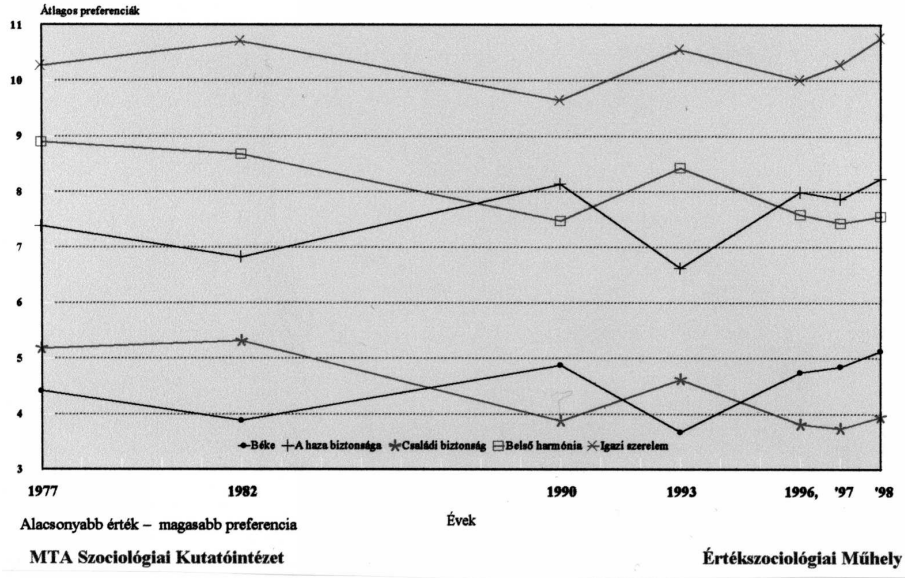
[1] Készült az OTKA T 016435 sz. *Értékváltozás 1978 és 1994 között Magyarországon* című kutatás keretében Szakolczai Árpád közreműködésével.

SZERETETTEL TELJES, ÖNÁLLÓ, TISZTA stb.) kell rangsorolniuk annak megfelelően, hogy ezen értékek saját életükben milyen fontos szerepet játszanak. Itt csak a legfontosabb megállapítások összegzését adjuk meg, a rendelkezésünkre álló egyetlen másik, az 1968-as eredeti amerikai vizsgálatból származó országos reprezentatív adattal való összehasonlítással együtt. S bár ilyen vizsgálatokat csak Magyarországon végeztek, az eredményeknek nyilvánvaló relevanciájuk van a térség más országai számára is.

20.1. „Hivatalos” versus „ellenzéki” értékek

A legfontosabb megállapítások öt érték nyomon követéséből származnak; ezek az értékek a BÉKE, A HAZA BIZTONSÁGA, A CSALÁDI BIZTONSÁG, A BELSŐ HARMÓNIA ÉS AZ IGAZI SZERELEM. A korábbi magyarországi felmérések azt mutatták, hogy a BÉKE és A HAZA BIZTONSÁGA értékei a hivatalos kommunista értékrendszert takarják. Ezek az értékek sokkal fontosabbak voltak 1978-ban és 1982-ben Magyarországon, mint 1968-ban az Egyesült Államokban. Kiderült továbbá az is, hogy ennek a különbségnek a kommunista pártapparátus volt a forrása, minthogy a két érték fontos volt a kommunista párttagok, s különösképpen a pártiskolát végzettek számára. A CSALÁDI BIZTONSÁG, a BELSŐ HARMÓNIA és az IGAZI SZERELEM főként a hivatalos értékrendhez nem tartozók számára voltak fontosak – akik ezeket az értékeket preferálták, a BÉKÉT ÉS A HAZA BIZTONSÁGÁT általában hátrább helyezték a rangsorban (lásd az alábbi ábrát).

20.1 ábra. BÉKE, A HAZA BIZTONSÁGA
versus
CSALÁDI BIZTONSÁG, BELSŐ HARMÓNIA, IGAZI SZERELEM



Az ábra jól mutatja, hogy mind az öt érték átlagos preferálása határozott minta szerint fluktuál 1978 és 1996 között. A kommunista időszakkal összehasonlítva 1990-ben csökkent a BÉKE és A HAZA BIZTONSÁGA relatív fontossága, míg magasabb preferenciaszintre került a CSALÁDI BIZTONSÁG, a BELSŐ HARMÓNIA és az IGAZI SZERELEM. 1993-ra ez a trend mindenestől megfordult, s mind az öt érték majdnem pontosan az 1989 előtti szintre tért vissza, előre jelezve az exkommunista szocialisták választási győzelmét. A múlthoz való ilyetén visszatérés azonban tiszavirág életű volt. 1996-ban következett az újabb gyökeres fordulat kivétel nélkül mind az öt érték esetében; ezúttal az 1990-es szint felé.

Az értékváltozásnak ez a mintája nem tulajdonítható a vizsgált populáció összetételében meglévő különbségeknek. A dolog megismétlődött két, iskolázottság szerint elkülönített csoportban: azoknál, akik csak általános iskolát végeztek és azoknál, akik ennél magasabb képesítést szereztek. Továbbá más értékeknél nem találtunk ilyen változásmintát. Ez persze nem azt jelenti, hogy a többi érték nem változott – bár mind az öt vizsgálatban a perszisztencia figyelemre méltó volt az országos átlagokban. Akadt két olyan értékcsoport, amelyeknél 1990 csupán törést hozott, későbbi visszafordulás nélkül. A határozott és nagy preferenciavesztést elszenvedő értékek tartalmukban a szocializmus lényegéhez kötődtek: az EGYENLŐSÉG, a TÁRSADALMI MEGBECSÜLÉS, a MUNKA ÖRÖME, és bizonyos mértékig az EMBERI ÖNÉRZET. Az első három mindegyike a kommunista időszakhoz képest majdnem két átlagponttal került hátrébb a rangsorban. Az ellenkező pólust a BOLDOGSÁG és főként az ANYAGI JÓLÉT értékek reprezentálják, amelyek sokkal fontosabbá váltak az 1989 utáni mindhárom vizsgálati évben.

Ezek a változások a magyarországi értékrendszert az amerikaihoz (USA) közelítették. Más tekintetben viszont a magyarországi értékrendszer továbbra is élesen eltért tőle. Nem meglepő módon a vallási értékek – az ÜDVÖZÜLÉS és a MEGBOCSÁTÓ – sokkal kevésbé voltak fontosak Magyarországon, míg a racionális-intellektuális értékek – a LOGIKUS GONDOLKODÁSÚ, az ÉRTELMES és az ALKOTÓ SZELLEMŰ – sokkal fontosabbak voltak. S ez összhangban áll Konrád György és Szelényi Iván jól ismert megállapításával: a kommunista országokban az értelmiség és a szellemi értékek különösen erős befolyásáról. 1990 azonban nem hozott eltérést ezen az értékekben a kommunista időszakra jellemző értékrendhez képest. Semmi változás nem mutatkozott a vallási értékek átlagos rangsorolásában, míg a racionális-intellektuális értékek nemhogy nem veszítettek fontosságukból, de súlyuk 1993-ban még nőtt is – bár 1996-ra nagyjából visszatértek 1989 előtti szintjükre. E tekintetben a magyar és az amerikai értékrendszer közötti igazán meglepő különbséget az a tény szemlélteti, hogy az ÉRTELMES értéket 1968-ban az Egyesült Államokban a 18 érték között az utolsó helyre rangsorolták, 1993-ban Magyarországon viszont az elsőre.

20.2. Változó választóvonalak

A változás azonos mintája 1990-ben, a megfordulásé 1993-ban, majd az 1990-es változásokhoz való visszafordulás és annak megerősítése 1996-ban egy olyan változáselemzésben tárható fel, amelyben jelentésükben hasonlónak tekintett és egymással ellentétes értékek is szerepelnek. Az előbbit inkább csoportokban preferálják, míg az utóbbi

ellentétpárokban jelenik meg – vannak olyan válaszadók, akik mindegyik értéket preferálják, de kevés olyan válaszadó akad, aki mindkettőt egyszerre preferálja. Ezeket a változásokat a felmérések adataira alkalmazott faktorelemzés módszerével lehet kimutatni.

Az értékek tekintetében Magyarország egyik 1989 előtti jellemzője volt például a hivatalos kommunista értékek (BÉKE és A HAZA BIZTONSÁGA) és az „ellenzéki" értékek (CSALÁDI BIZTONSÁG, BELSŐ HARMÓNIA, IGAZI SZERELEM) közötti ellentét. Ilyen választóvonal, illetve társadalmi hasadás nem volt jelen az amerikai vizsgálatban, és 1990-re eltűnt a magyar értékrendszerből, s nem tért vissza 1993-ban sem.

Az értékrendszer tekintetében nem ez volt az egyetlen nagy átrendeződés 1989 után. A kommunizmus idején a népességet megosztó fő választóvonal egyfelől a kommunista rendszerre leginkább jellemző két értékcsoport, a hivatalos-kommunista és a racionális-intellektuális értékek, másfelől pedig a hagyományos és közösségi értékek (SEGÍTŐKÉSZ, ENGEDELMES, UDVARIAS, MEGBOCSÁTÓ stb.) és a hedonisztikus értékek (VÁLTOZATOS ÉLET, ÉLVEZETES ÉLET, BOLDOGSÁG stb.) között húzódott. Ennek a társadalmi hasadásnak a koherenciája már 1982-re meggyengült, 1990-re pedig el is tűnt. 1993-ban visszatérés mutatkozott az 1989 előtti társadalmi hasadáshoz, ami majdnem reprodukálta az 1982-es helyzetet. 1996-ban azonban már teljes körű lett az átrendeződés. A hivatalos-kommunista és a racionális-intellektuális értékek mint a legfontosabb társadalmi választóvonal immár nem az egyik oldalon csoportosultak. Most a két csoport egymással ellentétes pólusra került, jelezve a magyar társadalomban az egyéni értékek mentén kialakult új és fontos választóvonalat.

Megerősítik és kiegészítik ezeket a megállapításokat a „hivatalos" és „ellenzéki" értékekben bekövetkező egyéb fontos változások. Az 1989 előtti magyarországi társadalmi választóvonalat a hivatalos-kommunista értékek uralták, minthogy a megkérdezettek ezeket mint értékcsoportot általában vagy elismerték, vagy elutasították. Az 1978-ban kimutatott öt választóvonalból háromnál – és az 1982-ben kimutatott hatból háromnál – ezek az értékek állnak szemben bizonyos más értékekkel. Az 1989 utáni változások három nagy trenddel jellemezhetők. Először is: megnövekedett a választóvonalak száma – 1990-ben és 1993-ban hétre, 1996-ban nyolcra –, ami az értékrendszerek társadalmon belüli fokozatos differenciálódásáról árulkodik. Másodszor: a hivatalos-kommunista értékek elveszítették megosztó szerepüket. Míg 1978-ban három választóvonalból még hármat határoztak meg, 1990-ben és 1993-ban hétből már csak egyet (1996-ban a társadalmi csoportok közötti további differenciálódásnak köszönhetően, nyolcból kettőt). S végül, a magyar értékrendszer struktúrája egyre jobban közelített az amerikaihoz. A Rokeach által 1968-ban az Egyesült Államokban megállapított választóvonalak többsége nem vagy csak alig volt jelen 1978-ban Magyarországon. A magyarországi értékrendszerben 1989 után mutatkozó új választóvonalak azonban könnyen felismerhető módon a korábbi amerikai választóvonalakhoz hasonlítanak, s még a meglévő hasadások változásai is nagyrészt ugyanabba az irányba mutatnak. Magyarországon 1989 előtt erős választóvonal mutatkozott például egyfelől a modern – racionális-intellektuális vagy pragmatista – értékek, másfelől pedig a tradicionális – vallási, közösségi vagy a fegyelemmel kapcsolatos – értékek között. 1989 után ez az egy választóvonal megkettőződött: létrejött egy a racionális-intellektuális (LOGIKUS GONDOLKODÁSÚ, ÉRTELMES, ALKOTÓ SZELLEMŰ ÉS HATÉKONY) és a vallási-közösségi értékek (MEGBOCSÁTÓ, SEGÍTŐKÉSZ, ÜDVÖZÜLÉS, SZERETETTEL TELJES, ENGEDELMES), és egy másik a szabad szellemű aktivitás (ALKOTÓ SZELLEMŰ, ELŐÍTÉLETEKTŐL MENTES) és a fegyelmezett passzivitás értékei (UDVARIAS, TISZTA)

között. Ez utóbbi választóvonal majdnem ugyanilyen formában már 1968-ban jelen volt az amerikai értékrendszerben.

Az adatok is megerősítik, ami várható volt, hogy a hivatalos-kommunista értékek egy sajátos, a régi rezsimben elit státusszal rendelkező csoporthoz kötődnek. Az azonos értékpreferencia-készlettel rendelkező legnagyobb és leginkább koherens válaszolói csoportot 1978-ban és 1982-ben is azok alkották, akik kifejezetten fontosnak tartották a hivatalos-kommunista és a racionális-intellektuális értékeket, és nem preferálták fontos értékekként a fő választóvonal másik oldalán elhelyezkedőket. A válaszolói értékpreferenciák elemzése alapján kialakított tíz klaszterből[2] 1978-ban az összes válaszolók 14,8 százaléka tartozott a hivatalos-kommunista-racionális-intellektuális csoporthoz, 1982-ben pedig 18,6 százalékot ölelt fel ez a csoport. Ráadásul ez a csoport valóban egy elit csoport volt; mindkét vizsgálatban a csoport tagjai kétszer nagyobb valószínűséggel rendelkeztek felsőfokú képzettséggel a népesség egészéhez képest.

Az 1989-es változások eredményeképpen ennek a csoportnak a dominanciája 1990-re eltűnt, ekkor ugyanis a válaszolóknak már csak 9,1 százaléka tartozott ide. 1993-ban a válaszolók 14,3 százalékával megint ez a csoport lett az egyik legnagyobb. A csoport belső összetétele azonban jelentősen megváltozott – iskolázottsági színvonala most nemigen haladta meg az átlagot. Ami azt jelenti, hogy a csoport új tagjai szinte kizárólag az alacsonyabb iskolai végzettségűek közül kerültek ki. Így tehát a kommunizmus domináns, hivatalos értékrendszerének kívánatossága nosztalgikus vagy egyéb okokból 1993-ban fokozódott ugyan, ám ez határozottan egy nemelit jelenség volt. S mint ilyen, nem is tarthatott sokáig.

1996-ban a mintának már csak 6,1 százaléka tartozott ehhez a csoporthoz, amely 1978-ban még az egymástól jól megkülönböztethető tíz csoport közül a legnagyobb volt, és 1996-ban tíz közül a legkisebb lett. Ugyanakkor a csoporton belüli iskolázottsági szint ismét jóval meghaladta az átlagot. Ez valószínűleg azért alakult így, mert a csoport még mindig magában foglalta a korábbi rezsim híveinek „kemény magját", miközben elveszítette nemelit tagjainak többségét, akik számára 1993-ban hirtelen oly vonzónak tűnt a korábbi rendszer.

Mindent egybevetve a vizsgálatokból nemigen lehet arra következtetni, hogy a magyarországi értékrendszer egyszerűen ugyanúgy ingadozik, ahogy a politikai preferenciák ingadoznak a demokratikus országokban. Ezek a vizsgálatok azt mutatják, hogy az 1993–1994-es exkommunista politikai győzelmek nem az 1989–1990-es ambiciózus változások tartós korrekcióját jelentik, hanem csak egy egyszeri és ideiglenes visszatérést a szétesőszocialista értékrendszerhez. A különös év nem 1990 volt, hanem 1993. Magyarországon az exkommunisták magukkal hozott értékrendszerének fényessége végképp megkopott.

[2] A nem-hierarhikus klaszterelemzés McQueen-féle eljárását alkalmaztuk és az MDS MINISSA-eljárását használtuk a klaszterek számának pontos meghatározásához.

21. fejezet

Értékrendszerek az axiális momentumokban

(24 európai ország összehasonlító elemzése)

A fejezet célja [1] egy új elméleti keret kimunkálása és letesztelése az értékszociológia számára. Elméleti, Max Weber műveire támaszkodó része kifejti, hogy az értékrendszer több rétegből áll, s hogy e rétegek mindegyike magán viseli a különböző „axiális momentumok" „bélyegét"; ilyen „axiális momentumok" azok a periódusok, amelyekben a politikai rend és a mindennapi élet rendezőelveinek felbomlása olyan nézetrendszerek megjelenéséhez vezetett, amelyek az egyének oldalára tolták át az értéksorrend alakítását. Az empirikus részben – a Magyarországon 1978 és 1993 között elvégzett négy országos reprezentatív felmérés adatai alapján hidakat építve a Rokeach-értékteszt és a World Values Survey között – írásunk klaszterelemzés, diszkriminancia-elemzés és LVPLS-modellezés segítségével azt mutatja be, hogy azok a 24 kelet- és nyugat-európai országban mutatkozó különbségek, ahogyan a társadalmi háttértényezők individuális szinten befolyásolják az értékpreferenciákat, nem annyira az országos szinten érvényesülő modernizációs vagy gazdasági változóknak, vagy a kommunizmus alatti viszonylagos liberalizációnak köszönhetők, hanem az olyan axiális momentumok bélyegének, mint például a protestantizmus, a felvilágosodás vagy a szocializmus különböző változatai. Az eredmények elég egyértelműen utalnak arra is, hogy a jelenlegi helyzet valóban egy másik axiális momentum jegyeit viseli magán.

A fejezetben 24 európai ország értékrendszerének vizsgálatát, valamint az értékrendszerek összehasonlító elemzésének néhány eredményét ismertetjük. A felhasznált adatok két forrásból származnak. Az első az 1990-es World Values Survey (WVS). Ez a vizsgálat 30, legalább részben európai országra terjedt ki. Hatot végül kénytelenek voltunk kihagyni: Svájcot és Szlovéniát, mivel az adatok nem álltak rendelkezésünkre; Törökországot és Izlandot, a távoli északnyugati, illetve délkeleti peremvidéket, mivel speciális esetnek vagy kivételnek ítéltük őket; és a két Írországot, mivel értékrendjüket oly erősen uralja a vallásosság, hogy esetükben az általunk választott elemzéstípust nem tudtuk alkalmazni. Az elemzésbe így a következő országok kerültek bele: Franciaország, Anglia, Olaszország, Nyugat-Németország, Hollandia, Dánia, Belgium, Spanyolország, Magyarország, Norvégia, Finnország, Svédország, Lengyelország, Belorusszia, Csehszlovákia, Kelet-Németország, Bulgária, Portugália, Ausztria, Moszkva, Litvánia, Lettország, Észtország és Oroszország.[1] A másik adatforrást azok az országos reprezentatív felmérések szolgáltatták, amelyeket a Magyar Tudományos Akadémia Értékszociológiai Kutatóközpontja végzett Hankiss Elemér, Manchin Róbert és Füstös László vezetésével 1977–1978-ban, 1990-ben és 1993-ban.

[1] Készült Szakolczai Árpád közreműködésével az OTKA T 016435 sz. *Értékváltozás 1978 és 1994 között Magyarországon* című kutatás keretében.

21.1. Elméleti keretek

Az értékszociológia még ma is szenved attól a múltbeli megosztottságtól, amely két különböző, sőt ellenséges táborra szakította a szakmát. A derékhadat a strukturalista-funkcionalista szociológia alkotta, amely Durkheimet és Parsonst követve az integratív társadalmi normákként értelmezett értékekre koncentrált. Ez az általános elméleti megközelítés alkalmatlan volt az empirikus elemzésre, s a helyzetet csak tovább nehezítette Durkheim viszolygása az „individualizmustól". Ráadásul „idealistának" minősítve támadták a szociológia legbefolyásosabb marxista áramlatai is. A másik megközelítés pszichológiai jellegű volt; ennek reprezentánsa a szükségletek hierarchiájáról Maslow által megfogalmazott elmélet, amely aztán Inglehart nagy hatású műveiben és G. W. Allport írásaiban köszön vissza. Ez a megkülönböztetés már sokkal jobban illeszkedik az empirikus elemzés módszereihez, de ellenkező előjelű hiányosságokkal rendelkezett: túl individualisztikus volt, nem kapcsolódott megfelelően az általános szociológiai törekvésekhez, és hiányoztak belőle a történelmi dimenziók.

Mindez pedig a következő dilemmához vezet. Egyrészt a két szemléletmód kibékíthetetlennek tűnik, másrészt viszont bármelyiket válasszuk is, bele kell törődnünk, hogy alapvető törés következik be az elméleti és az empirikus kutatás, a kortárs és a történelmi vizsgálódás, valamint az egyén és a társadalom között. Parsons explicit módon érvelt a „kaotikus egyéni akaratok" és a „normatív szabály" közötti „fundamentálisan mély" dichotómia létezése mellett (1968; 378). Ha elfogadnánk, hogy a társadalmi értékeket csak absztrakt, általános szinten lehet tanulmányozni, akkor ebből következően nem volna lehetőségünk vizsgálni, hogy az emberek miképpen alkalmazzák eredményesen az életükben az értékeket.

Furcsa azonban az az elképzelés, hogy feltétlenül választanunk kell e két megközelítésmód között, minthogy a normatív értékek parsonsi elméletének másik fontos forrása, Max Weber művei magukba foglalták mind a kettőt. Weber dolgozta ki a szociológiai elméletben jelenleg is használt elméleti és fogalmi apparátust, s ugyanakkor foglalkozott az empirikus kutatás egy eljárásával is, sőt, úttörője volt Németországban a survey-módszerek (kérdőíves felmérések) alkalmazásának is (Lazarsfeld és Oberschall 1965). „Magyarázó szociológiájának" a korabeli társadalmi valóság megértése volt a fő célja, műveinek nagy részében ugyanakkor távolra tekintő komparatív történeti vizsgálódásokat találunk. Írásunkban kifejtjük, hogy Weber művei valóban fontos kulcsokat kínálnak a dichotómiák leküzdéséhez, valamint elméleti keretek kidolgozásához az olyan értékkutatás számára, amely érzékeny az elméleti és történelmi vonatkozások iránt, s ugyanakkor empirikusan tanulmányozza az egyéni preferenciákat, esetenként kérdőíves technikák alkalmazásával is.

A kiindulási pontot két jól ismert kifejezés kínálja. Az első a különbségtétel a *rendes* és a *rendkívüli* szituációk között; ezt Weber a „Vallási csoportokról" szóló fejezet elején vezeti be *Economy and Society* (1978; 400) című művében. A második pedig a híres váltókezelő-hasonlat. Ennek értelmében „közvetlenül nem az eszmék, hanem az anyagi és a képzeletbeli érdekek befolyásolják az emberek viselkedését. Ám gyakran az 'eszmék' által kreált 'világképek', határozzák meg, váltókezelők módjára, azokat a pályákat, amelyeken az érdek dinamikája a cselekvést mozgatja." (Gerth és Mills 1947; 280). Ezekkel a fogalmakkal lehetőség nyílik felcserélni azt a szemléletmódot, amelyen az értékszociológia két eltérő megközelítése nyugszik, elkerülve így a dichotómiákat.

Az értékszociológia normatív és individualisztikus megközelítésének közös az orientációja: mindkettő azokat a végső tényezőket keresi, amelyek mindig és azonos módon determinálják a viselkedést. A különbség köztük mindössze annyi, hogy az egyik az egyé-

neken kívül létező „társadalmi tényekben” véli fölfedezni ezeket a tényezőket, a másik pedig az emberi természetben. Mi azonban másféle szemléletet választunk, amikor nem kívánjuk fölfedezni az örökké fennálló vagy egyetemes érvényű normákat, sem pedig az emberi természet hasonlóan fundamentális jelentőségű struktúráit, hanem inkább azokra a történelmi pillanatokra koncentrálunk, amelyekben egy társadalmi rend összeomlik. Az ilyen eseményt a (társadalmi) „rend felbomlásának” nevezhetjük. Ez a fogalom közel áll a Parsons-féle „normatív szerephez”, ahogy a fotó negatívja is emlékeztet a tényleges képre. A rend felbomlása nemcsak láthatóvá teszi a normákat, hanem a legfontosabb feladattá teszi a viselkedési szabályok megalkotását. Így aztán első lépésként a hangsúly az általános értékelmélet felépítésének elvont tevékenységéről áthelyeződik a világcivilizációk komparatív történelmére. Tulajdonképpen ezt tette Max Weber is 1910 végén, amikor elkezdett dolgozni két, szorosan egymáshoz kapcsolódó munkáján, a *Gazdaság és társadalom* című művén és *A világvallások gazdasági etikáj*án. S bár ezek közül a második jóval kevesebb érdeklődést váltott ki, mint az első, jelentőségére nemrég fölhívta a figyelmet Tenbruck (1980) és Schluchter (1989); Schluchter külön hangsúlyozva még a két munka párhuzamosságait.

Nem Weber volt azonban az egyetlen olyan befolyásos társadalomtudós, akinek munkája némiképp váratlanul az antikvitás felé fordult. Hasonló fordulat figyelhető meg Norbert Elias, Michel Foucault, Lewis Mumford, Friedrich Nietzsche és Eric Voegelin munkásságában. Karl Jaspers (1953) volt az, aki az „axiális kor” fogalmának megalkotásával koherens történeti kereteket adott ezeknek a megközelítésmódoknak. Az időszámításunk előtti VI. század környékén szimultán módon nagy spirituális megújulás játszódott le egy viszonylag rövid időszakon belül Indiában, Kínában, Izraelben, Perzsiában és Görögországban, amely az átfogó politikai, társadalmi és vallási-szellemi (spiritual) válságra adott válaszként megteremtette a legelterjedtebb világvallások és filozófiák alapjait.[2] Ezeknek a válaszoknak az az elképzelés volt a központi elemük, hogy a fennálló rendszer összeomlásának következtében nemcsak az egyéneket kell megvédeni és vigasztalni a nagy nyomorúságban, hanem a sorrendet is meg kell változtatni, s az egész közösséggel, annak szokásaival és törvényeivel szemben az egyén belső világát kell előtérbe helyezni. Az egyént és annak tudatát, a lelkét és az értékeit kell megtenni a világ új, szilárd tengelyévé (axis). Ebből származik az „axiális kor” kifejezés, valamint az értékfilozófia „axiológia” elnevezése is.

Az axiális kor egy konkrét történelmi időszakot jelent. Ha azonban erre az időszakra korlátozzuk figyelmünket, akkor fennáll a veszélye, hogy a szociológiai elméletet egy radikális historizmusra redukáljuk. A tér- és időbeli korlátozásokat azonban, ha az adott periódus konceptuálisan kiterjesztett, el lehet törölni, méghozzá az „axiális momentum” terminus bevezetésével. Axiális momentumról akkor beszélhettünk, ha a dolgok fennálló rendjében, s vele együtt a politikai rendszerben és a mindennapi élet társadalmi rendjében is *kollapszus* – meglehetősen ritka esemény – következik be, s egy olyan nagy vallási-szellemi megújuláshoz vezet, amely a kollapszusra adott válaszreakcióként a rend forrását az egyénen belülre helyezi. Ilyen események következtek be az i.e. és i. sz. első évszázadban (a Római Köztársaság bukása és a kereszténység megjelenése), az V–VII. században (a Római Birodalom összeomlása és az iszlám megjelenése [lásd Pirenne 1939]), a XV.–XVI. században (a középkor alkonya, a reneszánsz és a protestantizmus [lásd Huizinga 1990]), és végül a politikai abszolutizmus és a hagyományos európai társadalmi rend felbomlásának két nagy szakaszában, a felvilágosodás, illetve a szocializmus előretörésekor.

A Parsosns-féle szemléletmódban az értékszociológia fő kérdése az, hogy miképpen lehet azonosítani azokat a fundamentális, mögöttes rendezőelveket (ordering codes), ame-

lyek normatív módon megszabják az egyének viselkedését. A Weber-féle megközelítésben azonban az elemzés fő kérdése az, hogy mi történik az egyének életvezetésében, ha ezek a normák megszűnnek működni. Weber ezenkívül kulcsfontosságú konceptuális eszközt is kínál a „lenyomatképzés" („stamping") terminus bevezetésével.[3] A rend összeomlásakor erőszakkal tarkított és bizonytalan idők jönnek el, amelyek „nyomot" hagynak vagy „sebet" ejtenek az akkor élőkön. A természeti csapások vagy háborúk által okozott pusztítástól eltérően azonban a rend felbomlásának időszakai után a nyomok tartósan megmaradnak, minthogy belsőleg esnek szét a társadalmi rend lényegi struktúrái és kódjai – még a személyi és a kollektív identitás is. Ha ezek a rendezőstruktúrák érintetlenek maradnak, még a pusztító eseményeket is át lehet vészelni. Ha azonban ezek a háttérkódok maguk is felbomlanak, akkor a végül megjelenő új rendezőelvek magukon fogják viselni a régi események bélyegét.

Ezek az új szabályok azonban nem egyszerűen az erőszakos káosz lenyomatai, és nem is anonim, személytelen folyamatok eredményei. A stabil struktúrák összeomlása a „mi a teendő" kérdését veti fel, de nem a normák elméleti definiálásának szükségleteként, hanem abban az értelemben, hogy konkrét útmutatást kell adni a mindennapi életvezetés számára. Ezt igyekezett megvalósítani minden egyes axiális filozófia és vallás – és tartós hatásuknak is épp ez a magyarázata.[4] A „lenyomatképzés" fogalma nemcsak az események passzív jegyeit foglalja magában, hanem a változás által érintett emberek aktív reflexiós folyamatát is.[5]

A rend felbomlásának axiális momentumként való konceptualizálása két feladatot vet fel. Az egyik történeti, vagyis le kell írni a konkrét eseményeket, és el kell végezni összehasonlító elemzésüket. A másik azonban már szociológiai, és a tényleges következményekre koncentrál. Ez elvégezhető a „civilizációs folyamat" pszicho- és szociogenetikai rekonstrukciója (Elias 1994), vagy a társadalmi fegyelmezés genealógiája (Foucault 1976; Oestreich 1982) segítségével; valamint ezeknek a jegyeknek mint a társadalom szövetében meglévő, egymásra rakódó rétegek empirikus tanulmányozásával is. Ez a tanulmányozás nemcsak az egyéni szintű adatokra, hanem sajátlagos céltárgyára, az egyénre is támaszkodhat. A kérdés mármost az, hogy a távoli történelmi események lenyomatai milyen mértékben láthatók még ma is abban, ahogy az egyének a saját életüket mint reflektáló szubjektumok vezetik. Az axiális momentum fogalmából két következmény adódik a társadalmi értékek empirikus elemzése számára. Először is az axiális hiedelem- vagy nézetrendszerek (belief systems) a tradicionális társadalmi normáktól eltérően az egyéni preferenciákon keresztül akár szociológiai felmérésekkel is tanulmányozhatók. Az ilyen értékek feltérképezése azonban a válaszolóktól nagyfokú együttműködési készséget, egyfajta „reflexív felmérési módszert" igényel. Másodszor: az egyéni értékek vagy akár a csoportértékek tanulmányozása helyett az érték- vagy szubjektivitás-típusok tanulmányozására van szükség. Ez azt jelenti, hogy nem a változók, hanem az esetek, az egyéni válaszolók lesznek az elemzés alapegységei. A cél annak kiderítése, hogy léteznek-e jól körülhatárolható embercsoportok, amelyek összes értékeik kiválasztásában hasonlóságokat mutatnak.[6]

21.2. Adatok és módszerek

A Rokeach-teszt. A komparatív kutatásnak szerencsére rendelkezésére áll egy kitűnő értékteszt, amely megfelel a „reflexív felmérési módszer" kritériumainak, minthogy a társadalom elemzésének lehetőségével ösztönzi az igen kifinomult, mélyre tekintő pszichológiai látásmódot. Ezt a tesztet Milton Rokeach amerikai szociálpszichológus dolgozta ki. Rokeach 18 cél- és 18 eszközértéket választott ki a cél–eszköz felosztásnak megfelelően, arra kérve a válaszolót, hogy „rendezze el őket, mint az ÖN életének vezérlő elveit, aszerint, hogy mennyire fontosak az ÖN számára" (1973; 358), két különálló, 1-től 18-ig terjedő halmazban. Az értékek pontos felsorolása a Függelékben található. A Rokeach-teszt eredményeit széles körben elismerik ugyan, ám a múltban sajnálatos módon inkább a hátrányait hangsúlyozták. Gyakran éri az a vád, hogy túl általános, történelmietlen és elvont, mivel az egyéni értékekről úgy tartják, hogy konkrétabb dolgokhoz kapcsolódnak, nem pedig az átfogó értékekhez. A teszt fő problémája azonban az, hogy igen időigényes, minthogy az életükben alkalmazott értékpreferenciákról kérdezi a válaszolókat. Így aztán ennek az egyetlen tesztnek a kitöltése 15–20 percet vesz igénybe, amit egy országos minta mai költségei miatt általában elfogadhatatlannak tartanak. A tesztet ezért csak diákmintákra vagy kisebb csoportokra alkalmazzák. A teszt eddig publikált, egyetlen országos alkalmazása a Rokeach által elvégzett 1968-as NORC-vizsgálat.

A költségtényezőktől valóban nem lehet eltekinteni, ám a fenti érvelés figyelmen kívül hagyja azt a tényt, hogy a Rokeach-teszt arra készteti a válaszolókat, hogy reflektáljanak saját életvezetésükre, s így sokkal differenciáltabb információkat szolgáltat, mint pusztán a vélemények vagy attitűdök regisztrálása. Elméleti szempontból a reflexív jelleg, a teszt előnyei megérnék a költségeket. Magyarországon az Értékszociológiai Műhelyben négyszer sikerült országos reprezentatív mintán alkalmazni a Rokeach-tesztet: 1978-ban, 1982-ben, 1990-ben és 1993-ban (1993 óta még három alkalommal, 1996-ban, 1997-ben és 1998-ban volt olyan országos reprezentatív felmérés, amelyben szerepelt a Rokeach-teszt). A megfelelő mintanagyság 807, 2938, 1320 és 1538 volt, a hiányzó értékek aránya pedig 25 százalék. Ez az adathalmaz lehetővé tette a felvázolt elméleti keretek relevanciájának tesztelését.

Klaszterelemzés. A vázolt elméleti megközelítésből fakadó másik módszertani következményként a hangsúly a változókként felfogott értékek tanulmányozásáról áttolódik a válaszolók tipológiájára. A mi esetünkben a q-analízis (Brown 1980) technikái nem alkalmazhatók, minthogy nagyszámú változót feltételez viszonylag kisszámú válaszoló esetében. Ezért választottuk inkább a klaszterelemzést (Aldenderfer és Blashfield 1984; Everitt 1984). Igaz, hogy ma már nem nagyon használják sem a klaszterelemzést, sem pedig a válaszolók tipológiáját, érdekes megjegyezni, hogy az adatelemzés egyik alapítója, a Weber műveiből sokat merítő Lazarsfeld a technikáit pontosan az olyan egyéni válaszolók tipologizálására fejlesztette ki, mint a vásárlók, a rádióhallgatók, a mozilátogatók stb. Metodológiai-teoretikai értekezéseiben Gerth (1982) azt írja, hogy Lazarsfeld megpróbálta a maga nyelvére és módszereire „lefordítani" Weber tipológiai megközelítésmódját. S bár akkoriban még nem volt lehetséges a nagyszámú válaszolók osztályozása, ma ez a feladat már nem okoz nehézséget.[7]

21.3. Az elemzés áttekintése

A Rokeach-teszt és a klaszterelemzés megfelelő eszközök ugyan, de a magyar adatok klaszterelemzéséből nem nyerhetnénk szignifikáns megállapításokat a bennünket érdeklő elméleti kérdések megválaszolásához. Az eredmények csak a magyarországi helyzetre volnának érvényesek, s nem tehetnénk semmiféle összehasonlítást. Ezenkívül a klaszterelemzés exploratív módszer. Ennélfogva az eredmények elemzése erősen függ az értelmezéstől, s gyengén van alátámasztva egyéb bizonyítékokkal.

E két okból döntöttünk úgy, hogy hidakat építünk a Rokeach-teszt és a World Values Survey (WVS) között, kihasználva azt a helyzetet, hogy 1990-ben Magyarországon mindkét tesztet végigkérdezték ugyanazon a mintán. A vizsgálat pontos menete a következő volt.

1. Első lépésként klaszterelemzéssel érték- (vagy szubjektivitás-) típusokat deriváltunk a magyar adatok számára. A módszer exploratív jellege miatt ezt a lépést egymást követő szakaszokban végeztük el. Először egy sor különböző klasztermegoldást próbáltunk ki mind a négy mintaévre. Másodszor a 25 esetnél többet tartalmazó klaszstercentrumokból a sokdimenziós skálázás felhasználásával minden év számára kiindulási konfigurációként kiválasztottunk kilenc reprezentatív centrumot. Harmadszor pedig értelmeztük a létrejött kilenc klasztercentrumot minden év esetében, összehasonlítottuk a négy évet, azután pedig az egész adathalmaz számára kiválasztottunk tíz kiindulási klasztercentrumot. Az így kapott klasztercentrumok felhasználásával meghatároztuk a tíz értékvagy szubjektivitástípust. A fő kérdés pedig itt az, hogy az így kapott különböző értéktípusok milyen mértékben azonosíthatóak olyan axiális ideológiákkal, mint a katolicizmus, a protestantizmus, a felvilágosodás, a szociáldemokrácia vagy a szocializmus.

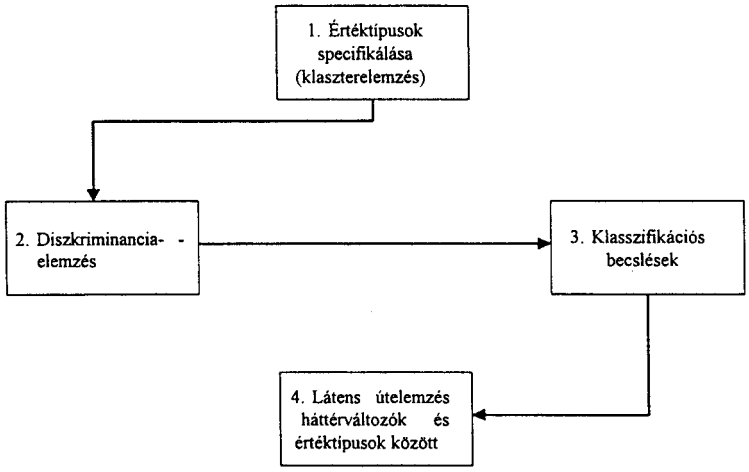
2. A második lépésben, még mindig az 1990-es magyar adatoknál maradva diszkriminanciaelemzéssel és a World Values Survey változóinak alkalmazásával megkülönböztettük egymástól a tíz értéktípust. Az első szakaszba próbaképpen sok változót vettünk fel, majd a másodikban elhagytuk azokat, amelyek nem bizonyultak statisztikailag szignifikánsnak.[8]

3. A harmadik lépésben a magyarországi adatoknál kapott diszkriminanciafüggvény segítségével meghatároztuk a tíz Rokeach-típust 26 európai ország számára. Annak érdekében, hogy ne veszítsünk el túl sok esetet, a hiányzó értékeket a megfelelő országos átlagokkal pótoltuk.

4. Ebben a stádiumban már közvetlenül összehasonlítható volna mind a 26 ország tíz értéktípusa. Az eljárás azonban két nagy nehézségbe ütközik: először is ez így még csak egy interpretációs gyakorlatozás volna, sőt becslések értelmezgetése. E korlátok leküzdése érdekében elvégeztük a becsült Rokeach-értéktípusok LISREL-féle LVPLS strukturális modellezését.[9] Méghozzá a következő megfontolásokból: először is, annak ellenére, hogy az adatok bizonyosan sok becslési hibát tartalmaznak, a strukturális jellemzőket ezek a tévedések csak kisebb mértékben befolyásolhatják, mint a puszta megoszlások; másodszor pedig, egy ilyen elemzés esetén lehetőségünk nyílik hipotézisek tesztelésére, s így a pusztán leíró és interpretáló megközelítés korlátain való túllépésre.

A modell egzakt specifikációja azonban hasonlóságokat mutat a korábbi eredményekkel. Ezért aztán közelebbről meg kell vizsgálnunk ezeket az eredményeket. Az olvasó eligazodásának megkönnyítésére a vizsgálat négy lépését külön is bemutatja a 21.1. ábra.

	Adatok/Minták	Eljárások és módszerek
1.	Rokeach-teszt Magyarország (4 minta: 1978, 1982, 1990, 1993)	Rokeach-féle értéktípusok specifikálása (Klaszterelemzés)
2.	Rokeach-teszt és WVS-adatok Magyarország (1990)	Az értéktípusok megkülönböztetése az 1990-es WVS-ből átvett változók alkalmazásával (Diszkriminancia-elemzés)
3.	WVS-adatok 26 ország (1990)	Az „értéktípusok” becslése (Klasszifikációs becslés)
4.	WVS-adatok 24 ország (1990)	A háttérváltozók értéktípusokra gyakorolt hatásának tesztelése az LVPLS módszerével



21.1. ábra. A kutatási eljárás áttekintése

21.4. A Rokeach-féle értéktípusok specifikálása Magyarország számára

Az ismételt klaszterelemzés révén kapott tíz típus leírását a 21.1. táblázat adja meg. Értelmezésüket segítette a 36 érték már korábban elvégzett faktorelemzése is (Hankiss et al. 1983). Ezen az alapon mind az eszköz-, mind pedig a célértékek két széles láncolatra oszthatók, melyek mindegyike, némi átfedéssel, három értékcsoportot tartalmaz. A négy láncolatot a 21.2–21.5. táblázat illusztrálja. A célértékek esetében az első lánc az elvontabb, ideológiai, vallási és etikai értékekből álló célokat tartalmazza (lásd 21.2. táblázat), a második pedig a konkrétabb, világiasabb, tradicionális, pragmatikus-anyagi és

hedonista értékeket (21.3. táblázat). Az eszközértékek egy elvontabb, aktívabb csoportot, az intellektuális, a pragmatikus és az etikai értékeket (21.4. táblázat), és egy világiasabb, passzívabb csoportot, a tradicionális-fegyelmező, a vallási-közösségi és a hedonista értékeket foglalják magukba (21.5. táblázat).

Típus	Jóval az átlag fölött preferált értékek	Jóval az átlag alatt preferált értékek
CL1	Béke, egyenlőség, *a haza biztonsága*, szabadság, *alkotó szellemű, logikus*, felelősségteljes, bátor, hatékony, előítéletektől mentes	Anyagi jólét, udvarias, boldogság, *jókedélyű*, engedelmes, *tiszta*, *szeretettel teljes*, megbocsátó
CL2	Béke, egyenlőség, a haza biztonsága, szabadság, *engedelmes megbocsátó*, udvarias, fegyelmezett	Szerelem, családi biztonság, belső harmónia, előítéletektől mentes, alkotó szellemű, *értelmes*, logikus, felelősségteljes
CL3	*Munka öröme, emberi önérzet*, társadalmi megbecsülés, barátság, hatékony, segítőkész, fegyelmezett, felelősségteljes	
	Béke, a haza biztonsága, szerelem, szeretettel teljes, *tiszta*, udvarias, boldogság, élvezetes élet, jókedélyű	
CL4	*Belső harmónia*, bölcsesség, szerelem, emberi önérzet, a szépség világa, *logikus*, *alkotó szellemű, értelmes*, előítéletektől mentes	*A haza biztonsága*, béke, egyenlőség, anyagi jólét, *engedelmes*, megbocsátó, tiszta, udvarias, törekvő
CL5	*Üdvözülés, megbocsátó, szeretettel teljes*, engedelmes, segítőkész, udvarias, tiszta, jókedélyű, szavahihető, béke	Változatos élet, szabadság, *alkotó szellemű, önálló, logikus*, bátor, hatékony, előítéletektől mentes
CL6	*Változatos élet, élvezetes élet, szerelem*, boldogság, anyagi jólét, jókedélyű, szeretettel teljes, tiszta, udvarias	*A haza biztonsága, béke*, egyenlőség, szabadság, fegyelmezett, felelősségteljes
CL7	*Anyagi jólét*, boldogság, a haza biztonsága, *önálló*, hatékony, alkotó szellemű, logikus, értelmes, béke, szerelem	*Egyenlőség, megbocsátó, segítőkész*, engedelmes, udvarias, szeretettel teljes, szabadság, emberi önérzet, társadalmi megbecsülés
CL8	*Boldogság*, családi biztonság, béke, szerelem, a haza biztonsága, engedelmes, tiszta törekvő, segítőkész, szeretettel teljes	*Társadalmi megbecsülés, hatékony*, *alkotó szellemű*, értelmes, a szépség világa, barátság, emberi önérzet
CL9	*Társadalmi megbecsülés*, egyenlőség, béke, a haza biztonsága, szavahihető, segítőkész, előítéletektől mentes, felelősségteljes	Anyagi jólét, élvezetes élet, boldogság, változatos élet, barátság, *jókedélyű*, megbocsátó, szeretettel teljes, alkotó szellemű
CL10	*Béke*, a haza biztonsága, szabadság, anyagi jólét, *jókedélyű, tiszta*, udvarias, szeretettel teljes	Társadalmi megbecsülés, belső harmónia, munka öröme, egyenlőség, *felelősségteljes, fegyelmezett*, engedelmes

Megjegyzés: az átlagtól való különösen nagy eltérést dőlt betű jelzi!

21.1. táblázat. A tíz értéktípus

A két csoport közvetlen kombinációjával, a cél- és az eszköz-láncolatok leginkább hasonló csoportjainak párosításával hat típust kapunk. Így az intellektuális és az ideologikus értékek összekapcsolásával jön létre a felvilágosult racionalista típus (CL1), és könnyen megkonstruálhatunk három további típust, a klasszikus szociáldemokratát (CL3), a hedonistát (CL6) és a tradicionális-fegyelmezőt (CL8). Probléma csak a maradék kettővel van, a vallásos és a pragmatikus-materialista típussal, mert ezek az értékek láncot váltanak a cél- és az eszközértékekben. Az okot azonban nem nehéz megtalálni. Annak a protestáns etikának a forradalmi hagyatéka ez, amely háttérbe szorította a tradicionális közösségi értékeket, és egy axiális nézetrendszeren belül a pragmatizmust és az anyagi jólétet hangsúlyozta. Minthogy Magyarország túlnyomórészt katolikus ország, nem meglepő, hogy a megmaradó két alaptípust vallásosnak (CL5) és materialistának (CL7) minősítettük.

A másik négy típus a hat fő típus változata volt, annak következtében, hogy a magyar mintában különösen jelentős szerepet játszottak az ideológiai értékek. Korábbi kutatásainkban már megállapíthattuk, hogy a Rokeach-tesztben ezek a felvilágosult racionalizmus olyan értékei, amelyek a legjobban individualizálják a létező szocializmus hivatalos értékeit. Ezek voltak azok az értékek, amelyeknél a populációs átlagok azt mutatták, hogy a preferencia sokkal magasabb volt Magyarországon, mint az USA-ban (Hankiss et al. 1983). A magyarországi adatok belső elemzése azt is mutatta, hogy ezeket az értékeket különösen fontosnak tekintette a párttagság és a magasabb iskolai végzettségűek egy része, s hogy feltűnően ideologikusnak és racionálisnak mutatkoztak azok, akik speciális pártiskolába jártak (Szakolczai 1987). Az ideológiai tényező túlzott szereplése jellemezte a maradék négy típust. Először is, a felvilágosult racionalista és a tradicionális-fegyelmező értékek közötti szembenállástól elkülönülten létezett egy csoport, amely nemcsak erősen intellektuális volt, hanem ideológiaellenes is. Ezt „nonkonformista értelmiségi" (vagy „poszt-felvilágosult") stratégiának nevezhetjük (CL4). A másik, ezzel homlokegyenest ellentétes típus nemintellektuális és tradicionális-közösségi irányultságú volt a maga eszközértékeiben, s ugyanakkor hangsúlyozottan ideologikus. Ezt „poszt-vallásos" típusnak nevezzük (CL2).

Ideologikus	**Keresztény**	**Szociáldemokrata**
egyenlőség	üdvözülés	munka öröme
		emberi önérzet
béke		társadalmi megbecsülés
szabadság		
a haza biztonsága		

21.2. táblázat. A célértékek elvont láncolata

Tradicionális	**Anyagi**	**Hedonisztikus**
családi biztonság	anyagi jólét	változatos élet
	élvezetes élet	
boldogság		barátság
		szerelem

21.3. táblázat. A célértékek világias láncolata

Racionális-szellemi	**Pragmatikus**	**Etikai**
értelmes		szavahihető
	hatékony bátor	
	előítéletektől mentes	
	alkotó szellemű felelősségteljes	
	önálló fegyelmezett	
logikus		

21.4. táblázat. Az eszközértékek elvont láncolata

Tradicionális-fegyelmező	**Keresztény-közösségi**	**Hedonisztikus**
udvarias	segítőkész	jókedélyű
	engedelmes szeretettel teljes	
tiszta	megbocsátó	

21.5. táblázat. Az eszközértékek világias láncolata

Az ideológia-értékek szintén párhuzamot mutattak két másik, szintén homlokegyenest ellentétes stratégiával – a szociáldemokratával és a hedonistával. Míg a klasszikus szociáldemokrata típus – a régi kommunista–szociáldemokrata elkülönülés egyik jeleként és jellemzőjeként – határozottan ideológiaellenes volt és közel állt a protestáns pragmatizmushoz, itt találtunk egy olyan csoportot is, amely az alapvető szociáldemokrata értékekhez való ragaszkodás mellett igen erős ideológiai orientációt árult el. És végül, míg a „standard" hedonisztikus csoport ideológiaellenességet mutatott, létezett még egy csoport, amely ideologikus is volt. Ezeket a típusokat „ideologikus szociáldemokratának" (CL9) és „ideologikus hedonistának" (CL10) nevezzük. A 21.6. táblázat a további elemzésben használt tíz típus elnevezését és kódjait sorolja fel.

CL1	Felvilágosult racionalista (hivatalos szocialista)
CL2	Posztvallásos
CL3	Klasszikus szociáldemokrata
CL4	Posztfelvilágosult
CL5	Vallásos
CL6	Hedonista
CL7	Materialista (posztkommunista)
CL8	Tradicionális-fegyelmező
CL9	Ideologikus szociáldemokrata
CL10	Ideologikus hedonista

21.6. táblázat. Az értéktípusok összegző leírása

21.5. Az értéktípusok becslése Magyarország esetében

Az értéktípusok becslése Magyarország esetében az 1990-es WVS változói felhasználásával

A 21.7. táblázat az eredeti és a becsült gyakoriság-megoszlásokat veti egybe, és megadja a helyesen klasszifikált eseteket is két különböző sorozat számára. Az elsőben a legjobb becsléseket tüntettük föl mind a tíz típusnál. Ez azt mutatja, hogy az elemzés igen jól működött a típusok többségénél, helyesen klasszifikálva az esetek több mint 33 százalékát, de a dolog már nem így alakult a CL2, a CL6, a CL7 és főként a CL10 esetében. A CL10 azonban elméleti szempontból eléggé érdektelen volt, s mivel az LVPLS-modell specifikációjához egy típust mindenképpen el kellett hagyni, úgy döntöttünk, hogy a CL10 értékelése után újra értékeljük a többi kilenc típust. Így a klasszifikálás a CL6, a CL7 és néhány más típus esetében jelentősen javult, s a CL2 maradt az egyetlen olyan típus, amelynél elég gyenge eredményt hozott a klasszifikációs becslés. Mindent egybevetve: eseteink 35 százalékát tudtuk helyesen megbecsülni. Ez elfogadható volt, mivel azt jelentette, hogy az értéktípusok átlagában a válaszolók mintegy 8 százaléka került a nyolc másik „rossz" típusba, míg 35 százalékuk a helyes típusba.

	CL1	CL2	CL3	CL4	CL5	CL6	CL7	CL8	CL9	CL10
Eredeti	9,1	5,8	4,8	10,5	14,4	15,0	16,1	8,0	6,3	9,7
Becsült	10,3	8,5	8,3	9,1	14,4	8,2	10,0	13,3	8,9	9,1
1%	37,3	22,2	41,0	37,8	44,4	29,0	30,5	33,3	37,0	13,5
2%	42,7	22,2	41,0	37,8	44,4	29,0	30,5	33,3	35,2	n. a.

21.7. táblázat. Az értéktípusok eredeti és becsült megoszlása Magyarországon (%-ban)

Megjegyzés: az 1% a helyesen klasszifikált eseteket adja mind a tíz klaszterben, a 2% a CL10 elhagyása után pedig a helyesen klasszifikált eseteket.

21.6. Az értéktípusok becslése a többi ország esetében

Az értéktípusok becslése az 1990-es WVS-ben szereplő többi 25 ország esetében, a Magyarország számára kifejlesztett diszkriminancia-függvény felhasználásával

Így aztán hozzáláthattunk az értéktípusok becsüléséhez a többi 25 ország esetében. Egy ilyen eljárásnak komoly veszélyei is vannak. A szokásos becslési kockázatot itt még az is növelte, hogy a hiányzó értékeket országos átlagokkal kellett pótolni, s hogy a típusokat a keleti világban ritkaságnak számító esetre, egy olyan egykori szocialista országra dolgoztuk ki, amely túlnyomórészt katolikus volt. Így aztán kétség merült föl, hogy a tipológia egyáltalán alkalmazható lesz-e a többi országra. A kapott becsléseket a 21.8. táblázat tartalmazza.

A táblázat első látásra is jól mutatja a becslés jóságát. Ezek szerint a vallásos típus még mindig a vallásos katolikus országokban volt a leginkább uralkodó; Lengyelországban, Portugáliában, Olaszországban, Spanyolországban és a két Írországban,[10] s

ugyanakkor minimális jelentőségű volt a protestáns (főként északi) és az egykori kommunista országokban. Hasonló módon a szociáldemokrata típus Csehszlovákiában és Svédországban volt a legerősebb, míg az ideologikus szociáldemokraták és az ideologikus hedonisták az egykori Szovjetunió országaiban mutatkoztak a legnagyobb arányban. A táblázat ugyanakkor egyéb érdekességekkel is szolgál. A szociáldemokraták aránya például Magyarországon volt messze a legalacsonyabb, ami némi magyarázattal szolgálhat a szociáldemokrata párt itteni, furcsa bukására, míg a posztvallásos értéktípus (CL2), a hivatalos szocialista értékrendszer nemelit verziója különösen a katolikus országokban volt erős.

	CL1	CL2	CL3	CL4	CL5	CL6	CL7	CL8	CL9	CL10
Oroszország	5,2	11,1	21,4	7,8	6,4	6,7	7,5	0,9	19,3	13,7
Moszkva	5,1	5,9	24,7	12,3	4,8	9,1	11,7	1,7	13,6	11,1
Belorusszia	6,1	5,2	23,8	5,8	6,1	7,2	11,5	1,3	20,4	12,5
Észtország	10,2	2,9	26,8	9,8	3,4	7,0	14,3	1,9	13,1	10,6
Lettország	13,5	4,2	27,9	10,9	4,4	6,2	16,4	1,9	5,5	9,1
Litvánia	15,3	8,2	21,2	8,0	12,1	7,3	7,5	4,7	3,3	12,4
Bulgária	10,7	13,2	15,8	12,2	6,0	8,6	11,3	2,0	9,4	10,9
Magyarország	10,3	8,5	8,3	9,1	14,4	8,2	10,0	13,3	8,9	9,1
Csehszlovákia	3,6	16,4	33,7	1,5	12,0	11,7	3,7	0,8	3,1	13,6
Lengyelország	3,4	11,8	14,7	7,1	30,8	34,3	2,9	1,5	1,3	2,1
Kelet-Németország	24,1	6,4	21,9	11,9	7,6	6,5	9,5	2,5	5,7	4,0
Nyugat-Németország	12,9	13,6	20,7	13,7	9,6	11,6	8,9	1,4	2,1	5,6
Ausztria	17,0	8,2	23,8	10,8	15,1	7,5	5,2	2,9	3,3	6,2
Finnország	7,0	13,8	12,9	13,1	6,1	19,4	8,8	1,4	4,4	13,1
Svédország	11,6	1,5	28,4	17,8	3,4	12,6	3,6	0,7	14,6	5,8
Norvégia	16,1	4,0	16,3	23,3	10,6	6,7	7,7	0,8	6,7	7,8
Dánia	12,8	2,5	21,3	17,9	7,5	15,4	10,2	1,7	5,3	5,4
Hollandia	8,3	22,0	17,3	14,4	9,2	11,7	7,2	0,4	3,0	6,5
Anglia	8,2	4,3	18,4	10,7	17,2	11,9	6,3	3,7	8,7	10,6
Észak-Írország	7,6	4,3	10,9	6,9	50,0	4,3	1,6	3,0	7,9	3,6
Írország	6,4	3,1	12,5	6,3	55,5	3,9	1,8	2,9	2,3	5,3
Belgium	2,9	23,6	17,1	9,0	16,3	12,0	5,3	2,1	3,8	7,9
Spanyolország	3,8	10,3	13,0	17,1	24,7	12,1	3,6	4,6	5,4	5,4
Portugália	1,8	20,7	14,7	7,2	30,6	5,6	2,0	5,7	9,1	2,7
Olaszország	3,6	14,6	26,6	9,1	25,0	5,9	4,6	2,8	3,9	4,0
Franciaország	3,5	19,3	26,1	7,4	9,5	14,5	6,0	2,6	4,8	6,4

21.8. táblázat. Az értéktípusok becsült megoszlása 26 országban

A táblázat már olyan adatokat is tartalmaz, amelyek nem voltak elérhetők a WVS-adatok közvetlen elemzésével. Ott különálló kérdésekbe foglalták, s így nem vetették össze egymással a valláshoz, a társadalmi demokráciához, az erkölcshöz és a jóléthez kapcsolódó dolgokat. Megoldást itt az sem jelentene, ha egyetlen elemzés alá vonnánk mindegyiküket, az így előálló módszertani nehézségekről nem is beszélve. A táblázat ezenkívül a vázolt elméleti keretek alkalmazhatóságát is bizonyítja. Értéktípusok kikövetkeztetésével, a válaszolói értékpreferenciák egész mintájának összevetésével és a reflexív Rokeach-teszt alkalmazásával lehetővé vált, méghozzá egyéni szintű adatokból, a legfontosabb axiális nézetrendszerek reprodukálása, amelyek ténylegesen még ma is irányítják az egyének életvezetését. Már részletezett okokból azonban úgy döntöttünk, hogy még tovább megyünk, s ezeket az eredményeket a latens változós modellezésben is felhasználjuk.

21.7. A háttérváltozók értéktípusokra gyakorolt hatásának tesztelése

A háttérváltozók értéktípusokra gyakorolt hatásának tesztelése 24 ország esetében (LVPLS-modellezés)

A központi elképzelés

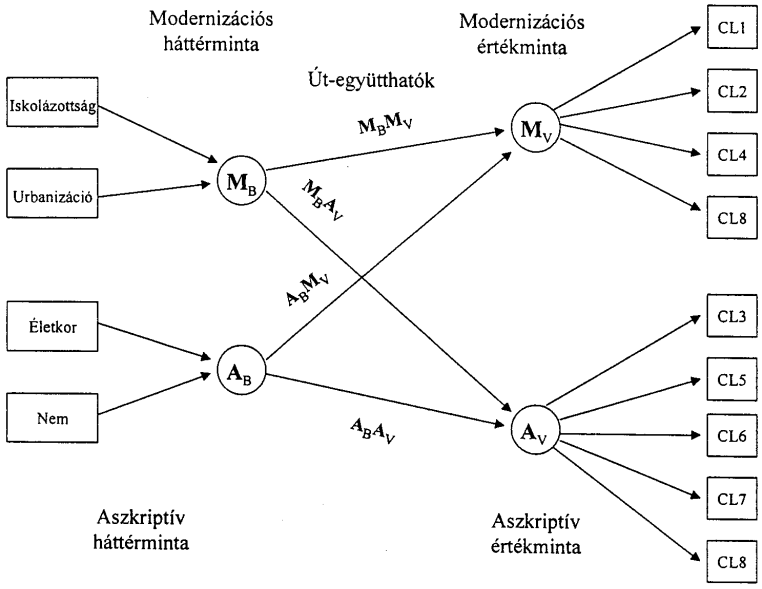
Az írásunk elején kifejtett elméleti alapvetés szerint az axiális nézetrendszerekként értelmezett értékek, amelyek a rend felbomlásának periódusaiban (az „axiális momentumokban") bukkantak fel és nyomódtak bele a társadalom szövetébe, igenis sokat számítanak, és független szerepet játszanak az egyének életvezetésének tényleges alakításában. Ha ez a felvetés helyes, akkor az egyes országok közötti különbségeket erősen befolyásolják az olyan axiális megfontolások, mint például a katolicizmus öröksége vagy a protestantizmus hatása, a felvilágosodás és a szocializmus. Ezzel az elképzeléssel szemben két ellenérv fogalmazható meg. Az első szerint az egyének (és az országok) közötti értékeltérések kizárólag a modernizáció eltérő szintjeiből fakadnak. Egyéni szinten ezt az iskolázottsággal, a foglalkozással vagy a lakóhellyel lehet mérni; országos szinten pedig jól alkalmazhatók a gazdasági jólét vagy a társadalmi és politikai fejlettség standard mércéi. A második ellenérv szerint az egyéni értékeltérések másik fő forrását a demográfiai vagy a születéssel adott (ascriptive) tényezők jelentik, például az életkor vagy az iskolai végzettség.

A három felvetés azonban nem zárja ki egymást. Ostobaság volna kijelenteni, hogy ma az egyéni szintű értékpreferenciák eltérései az iskolázottságtól, a lakóhelytől, a kortól vagy a nemtől függetlenek, és kizárólag vallási vagy ideológiai megfontolásokból fakadnak. A struktuális modellezés mögött húzódó központi elképzelés épp az ellenkező pólusról indul ki, és azt állítja, hogy az axiális nézetrendszerekként értelmezett értékeknek megvan a maguk autónómiája – még akkor is, ha a modernizációs és a születéssel adott tényezők természetesen befolyásolják őket. Ezt a hipotézist úgy ellenőrizhetjük, hogy kiválasztjuk a viselkedés bizonyos (politikai, családi, házasodási stb.) elemeit, és a modernizációs és a születéssel adott háttértényezők ellenőrzése után megnézzük, hogy a kapott értéktípusok rendelkeznek-e önálló magyarázóerővel. Tanulmányunkban azonban mi egy másfajta stratégiát követtünk. A háttérváltozók hatásának ellenőrzése helyett inkább arra kerestünk magyarázatot, hogy maguk a háttérváltozók miként idézik elő az értéktípusok szerinti megoszlást. Ilyenképpen az elemzés egyszerre két feladatot végez el. Közvetlen módon felméri, 24 európai ország esetében, hogy a társadalmi háttérváltozók milyen mértékben befolyásolják az egyéni szintű értékpreferenciákat. Közvetve pedig lehetővé teszi az axiális momentumra vonatkozó hipotézisek tesztelését, éspedig a 24 európai ország között abban mutatkozó különbségek alapján, ahogyan a háttérváltozók előre jelzik az egyéni szintű értéktípusba sorolódást.

A modell. A modellnek két blokkja van; az első a háttérváltozók mintáit tartalmazza, a második pedig az értéktípusokéit, a háttérváltozóktól az értéktípusokhoz vezető útegyütthatókkal együtt (lásd 21.2. ábra). A koefficiensek a két mintában a faktorsúlynak felelnek meg.

A modellt elméleti megfontolások és a magyarországi adatok átfogó elemzése alapján szerkesztettük meg. A modell bal oldala, a háttérváltozók mintája azt tételezi fel, hogy modernizációs háttértényezőt (M_B) alkotva kapcsolat lesz a magasabb iskolai végzettség és a városi lakóhely között, továbbá születési háttértényezőt (A_B) alkotva a

fiatalabb és férfi válaszolók között. A jobb oldalon hasonló elkülönülés látható az értéktípusok mintáiban. Az egyik tényezőt főként a modernizációs értékdifferenciák (felvilágosult és posztfelvilágosult racionalizmus) jellemzik, míg a másikat inkább a születési értéktípusok alkotják, amelyeket az (egyéni) életciklus- vagy (társadalmi- történeti) korszakhatások befolyásolják. S végül, középen két „közvetlen" útegyüttható ($M_B M_V$ és $A_B A_V$) látható a megfelelő háttér- és értéktényezők között, valamint két „kereszt" koefficiens ($M_B A_V$ és $A_B M_V$) is.



21.2. ábra. A modell

Általános hipotézisek

Elméleti megközelítésünkkel összhangban első alapvető hipotézisünk az, hogy az egyéni viselkedést irányító fundamentális axiális kategóriaként értelmezett értékeknek igenis van jelentőségük, és bizonyos módokon fontosabbak, mint a gazdasági jólét vagy a modernizáció szintje. Az értékrendszer jelenlegi szerkezete nyilvánvalóan ezeknek az axiális kategóriáknak valamilyen hierarchikus alakzata, amely a hagyományos értékek maradékaival kezdődik, a kereszténység, majd a protestantizmus, azután a felvilágosodás hatásával folytatódik, s a szocializmus különböző változataival fejeződik be. Konkrétan arra számítunk, hogy a kommunizmus lenyomata olyan határozott osztóvonalat képez az országok között, amely a katolikus-protestáns megosztottsághoz mérhető, s ugyanakkor nem érzékeny a modernizációs szintek csoporton belüli eltérései iránt. Itt Lengyelország lesz a különösen fontos példa, minthogy elméleti alapvetésünk értelmében a katolicizmushoz való ragaszkodás ezt az országot nyilván ellenállóbbá tette a kommunizmusból fakadó, maradandó hatásokkal szemben. Azt várjuk tehát, hogy a háttértényezőknek az értékekre gyakorolt befolyása miatt Lengyelország ebben az értelemben nem a többi kelet-európai országhoz fog közel állni, hanem sokkal inkább az olyan katolikus országokhoz, mint Spanyolország vagy Portugália.

Második hipotézisünk szerint az axiális hiedelemrendszerek nem egyszerűen csak fontosak, de sok függ kialakulásuk körülményeitől is. Az axiális momentum lényege az, hogy belsőleg omlik össze a korábbi rendbe vetett hit. Ha ilyen válság nem lép fel, akkor a rendezőelvek eltörlése és új axiális nézetrendszerek előírása felszínes és hatástalan marad, s az emberek az egészet kényszerként, diktatúraként fogják megélni. Az ilyen országokban valószínűleg erősebb az új ideológiával szembeni ellenállás. Pontosan ez történt Írországban az abszolutista korszakban, s nagyrészt ez a magyarázata az ország még ma is jól érzékelhető speciális értékrendszerbeli státusának. Így aztán azokban az országokban, amelyekben a szocializmust a világháború kezdetén s nem az azt követő általános összeomlás időszakában vezették be, és amelyekben a nemzeti kultúra erős maradt, a szocializmus hatásának jól láthatóan eltérőnek kell lennie. Azt várjuk tehát, hogy a többi szocialista országhoz képest a leginkább specifikus eltéréseket a balti országok fogják mutatni, főképpen Litvánia, s nem az olyan, gazdaságilag vagy társadalmilag modernizáltabb vagy politikailag liberálisabb országok, mint Magyarország, Kelet-Németország vagy az egykori Csehszlovákia.[11]

Harmadszor: sok szó esik mostanában, s különösen 1989 óta, korunk válságáról. Könnyű ezt a dolgot azzal elintézni, hogy a válság eszméje mindig is jelen volt a modernitás történetében. Ugyanakkor rövidlátás volna figyelmen kívül hagyni a bennünket körülvevő válságjeleket. Elméleti alapvetésünkben a történelmet axiális momentumok történeteként fogtuk föl. Utolsó általános kérdésünk pedig az, hogy adataink alapján érzékelni tudjuk-e egy axiális krízis jeleit. Ez minden axiális nézetrendszertől való gyors elfordulást feltételez. Ezért aztán külön figyelmet fordítottunk azokra az országokra, amelyek az axiális nézetrendszerek első vonalába kerültek; ide tartozik Anglia (a protestantizmus, de a munkásmozgalom miatt is), Franciaország (a felvilágosodás miatt), Svédország (a szociáldemokrácia miatt), valamint Oroszország (a kommunizmus miatt), és az is feltételezhető, hogy egy axiális nézetrendszer összeomlásának a rendszer legbelső országaiban lesz a legnagyobb a hatása.

Operacionalizált hipotézisek

I. A háttérminta

1. Hipotézis: Azt feltételezzük, hogy az iskolázottság relatív jelentősége nagyobb lesz, mint az urbanizációé, minthogy a modern társadalmakban az iskola a szocializáció legfontosabb terepe. A német típusú településmintával rendelkező, vagyis a történelmileg meggyökerezett, erős falu–város különbségekkel terhelt országokban ugyanakkor az urbanizációnak is igen nagy a jelentősége, szemben például a latin országokkal, ahol az ilyen különbségek jóval kisebbek (Brunner 1992; Tribe 1988).
2. Hipotézis: Azt feltételezzük, hogy az életkor relatív jelentősége nagyobb, mint a nemé. Ugyanakkor a nemek közötti eltérések feltétlenül jelen vannak azokban az országokban, amelyekben fennmaradtak a tradicionális minták. Így arra lehet számítani, hogy a nem jelentősége nagyobb a kelet-európai és a katolikus országokban, s különösen ott, ahol e két tényező átfedi egymást, például Lengyelországban.

II. Az értékminta

3. Hipotézis: A magyarországi adatok elemzése azt mutatta, hogy a szocializmus hivatalos elit ideológiája a felvilágosult racionalista értéktípus (CL1) volt, a hivatalos nemelit ideológiája pedig a posztvallásos értéktípus (CL2). Így azt várjuk, hogy Keleten a CL1 sokkal pozitívabban fog korrelálni a modernizációs értékfaktorral,

mint Nyugaton, különösen a posztfelvilágosult értéktípussal (CL4) szemben, míg a CL2-nek erős negatív korrelációt kell mutatnia.

4. Hipotézis: Azt várjuk, hogy a vallásos értéktípus (CL5) koefficiensei, a szekularizáció általános trendjét tükrözve, pozitívan korrelálnak az életkorral, s különösen magasak azokban a katolikus országokban, amelyek még ma is viszonylag erősen vallásosak. Ez önmagában még nem egy nagy fölfedezés, ám mégiscsak megerősíti a modellt és az adatokat.
5. Hipotézis: A szociáldemokrata értékek koefficienseinek pozitívan kell korrelálniuk az életkorral mindenütt, ahol ennek az ideológiának a háttérbe szorulása feltételezhető, és negatívan azokban az országokban, amelyekben ez az ideológia még 1990-ben is erős volt. Az első csoportba nagy valószínűséggel besorolhatjuk Svédországot, Nyugat-Németországot és az egykori Szovjetunió országait, ahol elég elmosódott a szociáldemokrácia és a szocializmus közötti különbség; a második csoportba pedig az erős, de nem domináns szociáldemokrata hagyománnyal rendelkező dél- és kelet-európai országokat, például Olaszországot, Spanyolországot, Portugáliát, Csehszlovákiát és Kelet-Németországot. S végül feltételeztük, hogy az ideologikus szociáldemokrata típus sokkal fontosabb Kelet-, mint Nyugat-Európában.
6. Hipotézis: Feltételeztük, hogy a materialista értéktípus (CL7) koefficiensei negatívan korrelálnak az életkorral, azt az egyszerű tényt tükrözve, hogy életpályájuk elején az embereket jobban foglalkoztatják az egzisztencia-teremtés anyagi kérdései. Ennek az összefüggésnek sokkal erősebbnek kell lennie az egykori szocialista országokban, mind az életszínvonalbeli különbségekből, mind pedig az általános posztkommunista társadalmi hangulatból fakadóan. Hasonlóképpen azt várjuk, hogy a hedonista típus (CL6) gyakrabban fordul elő a fiatal válaszolók körében, de itt azt is feltételezzük, hogy a fő választóvonal nem a keleti és a nyugati, hanem a katolikus és a protestáns országok között húzódik, magasabb életkori együtthatóval a katolikus országokban, minthogy a hedonizmus a szekularizáció egyik komponense.

III. Az útegyütthatók

7. Hipotézis: Az útegyütthatókkal kapcsolatos első feltevésünk az, hogy a modell helyesen van megszerkesztve, s hogy a „közvetlen" koefficiensek magasabbak lesznek, mint a „keresztek". A „közvetlen" koefficiensek Kelet-Európában a modernizációs háttérváltozók, Nyugat-Európában pedig születési háttéregyütthatók esetében magasabbak, amiben a formális iskolázásnak az államszocializmus alatti nagy jelentősége tükröződik (Konrád és Szelényi 1979), valamint a generációs különbségek egyre növekvő relevanciája Nyugaton (Inglehart 1977; Abramson és Inglehart 1992). Ez a hipotézis önmagában szintén nem sok újat mond, de megteremti a lehetőségét a döntő fontosságú végső tesztelésnek.
8. Hipotézis: Egy axiális momentum lehetséges azonosítását, harmadik általános hipotézisünket nagyban alátámasztják a „kereszt" útegyütthatók. Utolsó operacionalizált hipotézisünk szerint, miközben azt várjuk, hogy a magasabb iskolai végzettség és a fiatalabb életkor valamiféle általános hasonlóságot fog mutatni – még a fiatalabb kohorszok magasabb iskolai teljesítményének leszámítása után is, a generációs konfliktusok eseteiben, jelentkezhet egy megfordulás is, amely a magasabb végzettségűek és az idősebbek értékei felé közelít. Ha egy bekövetkező általános axiális válságra vonatkozó hipotézisünk helyes, akkor a „kereszt"-együtthatóknak pozitívaknak kell lenniük az axiális nézetrendszerek összes belső országában: Oroszországban (kommunizmus), Franciaországban (felvilágosodás), Svédországban (szociáldemokrácia) és Angliában (protestantizmus, munkásmozgalom).

Továbbá ugyanezek az együtthatók segíthetnek kijelölni az elkülönülési határokat a különböző kelet-európai országok között, a második általános hipotézissel összhangban. Ennek értelmében a kommunizmus által az egyéni értékekre gyakorolt hatás más és más volt azokban az országokban, ahol a kommunizmust a világháború utáni társadalmi összeomlás periódusában teremtették meg, ahol azt még a háború korai időszakában hozta be egy hódító hadsereg (mint például a balti országokban), vagy ahol érintetlenül megmaradt egy axiális nézetrendszer (mint például Lengyelországban).

Mindezt egybevetve lehetőség nyílik e kulcsfontosságú hipotézis nagyon feszes és pontos megfogalmazására. Arra számítunk, hogy a 21.2. ábrán látható modell „kereszt"-útegyüttható i ($M_B A_V$ és $A_B M_V$) negatívak lesznek, kivéve a következőket: az $M_B A_V$ pozitív lesz Anglia és Svédország esetében (jelezve a születési értéktényezőben rejlő zavarokat a releváns axiális nézetrendszerekkel, a protestantizmussal és a szociáldemokráciával kapcsolatosan), továbbá az $A_B M_V$ pozitív lesz Franciaországban, valamint a balti országok és Lengyelország kivételével az összes kelet-európai ország esetében (jelezve a modernizációs értéktényezőben rejlő zavarokat a releváns axiális nézetrendszerekkel, a felvilágosult racionalitással és a szociáldemokráciával kapcsolatosan).

21.8. Eredmények

A 21.9. táblázat az LVPLS-elemzés eredményeit tartalmazza, a 21.10. táblázat pedig azt mutatja, hogy az illeszkedés mértéke kielégítő volt mind a 24 esetre.[12] A következőkben lépésről lépésre áttekintjük az operacionalizált hipotéziseket, azután levonjuk a szélesebb érvényű következtetéseket.

1. Hipotézis. Az első két hipotézist nem azért fogalmaztuk meg, hogy új és meglepő felismerésekre jussunk, hanem inkább csak azért, hogy ellenőrizzük a modell stabilitását és érzékenységét. Hipotéziseink remekül megfeleltek ennek az elvárásnak. Az első hipotézist, amely szerint az iskolázottsági változó fogja dominálni a modernizációs tényezőt, teljes mértékben alátámasztották az adatok. Az iskolázottsági együttható a legtöbb országban közel volt az egyhez, ami csaknem teljes megfelelésről árulkodik e változó és a latens faktor között.[13] Az urbanizációs együttható az előrejelzésnek megfelelően alacsony volt a legtöbb kelet- és dél-európai országban, magasabb – Norvégia kivételével – északon, és különösen magas a germán országokban (amelyekhez az ebben a tekintetben még ma is fontos történelmi kötődések miatt most odasorolhatjuk Lettországot is), s Ausztria, Brunner klasszikus, *Land and Lordship* című könyvének tárgya volt az egyetlen eset, ahol az urbanizáció fontosabbnak bizonyult az iskolázottságnál.

2. Hipotézis. Ahogy azt feltételeztük, az esetek többségében az életkor a biológiai nemnél sokkal fontosabb változónak bizonyult a születési tényező alakításában. Különösen a protestáns országokban, ahol a nem hatása elhanyagolható volt. A nem ugyanakkor nagy jelentőségű maradt a katolikus országokban (Spanyolország, Portugália, Ausztria és Olaszország) és Kelet-Európában (Moszkva azért lehetett itt kivétel, mert városi minta volt). Ahogy arra számítani lehetett, a nem különösen jelentős volt Lengyelországban, ahol még az életkort is megelőzte e tekintetben. Három, a várakozástól eltérő nemegyüttható is jelentkezett. Először is Lengyelország mellett a nem hatása igen

erős volt még Lettországban is, valószínűleg a speciális demográfiai összetétel miatt. Másodszor: Norvégia volt az egyetlen olyan ország, ahol a koefficiens negatívnak bizonyult, jelezve, hogy itt a nők értékpreferenciái inkább a fiatalokéra, semmint az idősekére hasonlítottak. És Svédország volt a harmadik olyan ország, ahol a nem fontosabbnak bizonyult az életkornál.

Így tehát az első két hipotézis eredményei nemcsak megerősítik a közismert dolgokat, hanem még valami többlettel is szolgálnak. A települési mintákban és a nemi eltérésekben rámutattak két olyan területre, ahol még mindig élnek bizonyos tradicionális, nem axiális megfontolások.

3. Hipotézis. A modernizációs értékcsoport eredményei nemcsak megerősítették a hipotézisünket, hanem mind a Kelet–Nyugat, mind pedig a katolikus–protestáns eltéréseket illetően további adalékokkal is szolgáltak, sőt újabb bizonyítékokkal támasztották alá a felvilágosult racionalitás és a hivatalos szocialista ideológia közti kapcsolatot. Először is, a legtöbb szocialista országnál megjelent a faktorsúly várt struktúrája; ezenközben a posztfelvilágosult (CL4) típusnál, s különösen a felvilágosult racionalista (CL1) típusnál magas és pozitív volt az együttható, a posztvallásosnál (CL2) pedig magas és negatív. Akadtak azonban kivételek is. Lengyelország teljesen eltérő, a portugálhoz és a spanyolhoz nagyon hasonló mintát mutatott. A keletnémet minta itt ugyanúgy viselkedett, mint a nyugatnémet. A CL2 együtthatója pozitív volt Litvániánál. És végül, kisebb eltérések mutatkoztak a standard mintától még Lettországnál és Észtországnál is, és nem sikerült magyarázatot találni a CL4 negatív előjelére Csehszlovákia esetében.

Másodszor: valóban feltűnő Kelet–Nyugat eltérés mutatkozott a „posztvallásos" értéktípus (CL2) együtthatóiban. Míg Kelet-Európában, ahogy azt egy magyarországi próbafelmérés is jelezte e típuson belül, az alacsony iskolai végzettségűek voltak túlsúlyban, addig a nyugat-európai országok többségében ez épp fordítva alakult. Nyugat-Európában a felvilágosodás ideológiai célértékeinek elfogadása, a kereszténység eszközjellegű közösségi értékeinek egyidejű megtartása mellett inkább a magasabb iskolai végzettségű rétegekre volt jellemző. Három kivétel azért akadt: Franciaország, a klasszikus felvilágosodás hazája, és két északi ország, Dánia és Norvégia.

Harmadszor: megtaláltuk a várt Kelet–Nyugat különbségeket a CL1 és CL4 együtthatói között. Keleten, főként az egykori Szovjetunió oroszországi részein a felvilágosult racionalista értéktípus (CL1) aránya a modernizációs értékténye zőn belül magas, esetenként többségi volt, míg Nyugaton, Franciaország és talán Svédország kivételével a posztfelvilágosult típus volt uralkodó.

Az adatok ezenkívül egy további általános mintára is fényt derítettek. A Nyugat-Európában a CL1 faktorsúlya főként a katolikus országoknál (Franciaország, Ausztria, Portugália és Spanyolország) pozitív és viszonylag magas volt, és néhány kivétellel nagyon alacsony, sőt negatív volt a protestáns országoknál (Finnország, Norvégia, Hollandia, Anglia). Így tehát még egy 1990-ben elvégzett értékpreferencia-vizsgálatból is kimutatható, hogy a felvilágosodás ideológiája sokkal mélyebb nyomokat hagyott a katolikus, mint a protestáns országokban.

4. Hipotézis. Ezt az elég evidens hipotézist, amely szerint a vallásosságon belüli életkor-különbségek erősebbek voltak a vallásosabb katolikus országokban, az adatok teljes mértékben alátámasztották. Az ilyen triviális megállapítások azonban nem jelentenek problémát, ugyanis éppen a modell szilárdságát mutatják, és további támogatást nyújtanak más, nem triviális eredmények számára. Továbbá a CL5 majdnem egységesen magas faktorsúlya igen hasznos, mivel az ehhez a tényezőhöz vezető útegyütthatók stabil értelmezését kínálja.

	Isk.	Urb.	Nem	Kor	MM	MA	AM	AA
Oroszország	88	36	57	80	15	–08	02	25
Moszkva	100	NA	15	98	12	–05	07	20
Belorusszia	100	NA	54	87	12	–04	02	25
Lettország	78	55	97	27	12	–05	–10	23
Észtország	98	17	38	90	14	–05	–14	17
Litvánia	100	NA	49	84	10	–17	–10	24
Bulgária	92	16	30	98	21	–15	11	20
Magyarország	99	05	43	89	24	–05	01	24
Csehszlovákia	93	26	26	95	07	–19	03	18
Lengyelország	100	NA	95	31	–01	–19	–11	11
K.-Németo.	83	48	36	89	11	–07	–11	25
Ny.-Németo.	79	52	28	93	05	–12	–0	28
Ausztria	61	72	55	82	11	–23	–11	21
Finnország	100	NA	15	99	08	01	–11	31
Svédország	84	44	75	64	05	17	–07	16
Norvégia	88	33	–26	96	03	–09	–27	14
Dánia	83	45	17	96	10	–10	–12	21
Hollandia	85	47	08	100	00	–14	–17	18
Anglia	100	18	22	98	01	06	–14	31
Belgium	83	46	15	98	04	–09	–15	28
Spanyolo.	91	28	39	90	06	–06	–19	39
Portugália	89	26	45	88	20	–26	–10	26
Olaszország	93	25	58	81	07	–10	–09	32
Franciaország	96	22	32	99	10	–07	04	23

	CL1	CL2	CL4	CL8	CL3	CL5	CL6	CL7	CL9
Oroszország	56	–58	59	NA	13	70	–39	–48	32
Moszkva	85	–8	44	02	05	37	–68	–33	54
Belorusszia	79	–43	44	–70	39	66	–27	–39	43
Lettország	46	–19	87	–46	51	43	14	–65	33
Észtország	23	–82	53	–49	06	31	–20	–52	76
Litvánia	61	40	69	–43	22	77	–46	–34	19
Bulgária	59	–63	50	NA	–05	82	–36	–42	17
Magyarország	57	–39	73	–43	–06	61	–33	–71	02
Csehszlovákia	56	–73	–39	19	–22	85	–22	–41	09
Lengyelország	26	53	80	–42	04	82	–16	–51	–21
K.-Németo.	–07	16	98	–65	–11	89	–26	–37	38
Ny.-Németo.	–17	30	94	–46	09	78	–56	–07	25
Ausztria	49	26	83	–55	–03	87	–45	–19	–07
Finnország	–52	56	65	–35	24	66	–49	–18	50
Svédország	56	67	59	–73	73	–20	–51	–24	–32
Norvégia	–23	–26	94	–07	–04	77	–25	–07	58
Dánia	34	–36	87	–41	19	80	–51	–03	27
Hollandia	–16	56	81	–54	31	61	–40	–24	56
Anglia	00	34	94	–63	09	68	–57	–27	36
Belgium	04	78	62	–06	15	86	–43	–08	22
Spanyolo.	20	49	85	–36	–08	87	–42	–19	14
Portugália	36	56	75	–27	–21	87	–37	–22	11
Olaszország	08	14	99	–35	–32	85	–38	–18	06
Franciaország	89	–45	02	50	13	60	–75	–10	21

21.9. táblázat. 24 európai ország LVPLS-együtthatói

Ország	rms
Oroszország	0,031
Moszkva	0,019
Belorusszia	0,037
Lettország	0,021
Észtország	0,025
Litvánia	0,027
Bulgária	0,019
Magyarország	0,027
Csehszlovákia	0,043
Lengyelország	0,029
Kelet-Németország	0,030
Nyugat-Németország	0,033
Ausztria	0,027
Finnország	0,018
Svédország	0,035
Norvégia	0,033
Dánia	0,032
Hollandia	0,042
Anglia	0,027
Belgium	0,029
Spanyolország	0,035
Portugália	0,042
Olaszország	0,021
Franciaország	0,017

21.10. táblázat. Illeszkedési mértékek az LVPLS-modelleknél

Megjegyzés: az rms cov (e, u) a belső és a külső reziduálisok kovarianciáinak négyzetes középértékének négyzetgyöke (root mean squared covariances of inner and outer residuals).

5. Hipotézis. Itt először is meg kell jegyeznünk, hogy modellünkben a szociáldemokrata értékekkel kapcsolatban specifikációs probléma merül fel. Általában véve ezeket az értékeket, a férfi válaszolók jobban preferálták a női válaszolóknál, de a szociáldemokrácia általános bukása miatt támogatói nagyobb számban találhatók az idősebb kohorszokban. Így aztán az általános életkor–nem relációk megfordulnak. Ez a magyarázata annak, hogy gyakran miért oly alacsony ezeknek az értéktípusoknak a faktorsúlya.

Az a hipotézisünk, amely szerint a klasszikus szociáldemokrata típus (CL3) a fiatalok között visszaszorulóban van ennek az ideológiának a hátországában, különösen Svédországban bizonyult helytállónak. Jelentős veszteségek mutatkoztak Hollandiában, Finnországban és Dániában, valamint az egykori Szovjetunióban is, és valamivel kisebbek Nyugat-Németországban. Az ellenkező póluson azok között az országok között, amelyekben ez az értékorientáció 1990-ben még vonzotta a fiatal férfiakat, valóban ott találtuk az előre jelzett országokat, és csakis azokat.

A hipotézis második, az ideológiai szociáldemokrata típusra (CL9) vonatkozó része megerősítést nyert, amennyiben e típusnak a fiatalok körében bekövetkezett feltételezett visszaszorulásáról van szó (kivételt csak a szokásos elkülönülők jelentettek: Lengyelország és Svédország), de nem igazolódott a feltételezett, egyszerű Kelet–Nyugat minta

esetében. Az együttható valóban magas volt az egykori Szovjetunió országainál, és alacsony a katolikus nyugati országoknál, de alacsony a többi kelet-európai ország, és ismét magas a protestáns országok esetében, Svédország szokásos kivételével. Ennek oka valószínűleg az, hogy ez az értéktípus a szociáldemokrata és a kommunista értékek kombinációja, s így jelentősége mindenütt csökkent, ahol e két értékrendszer valamelyike tért vesztett az utóbbi évtizedekben. Ez a tényleges térvesztés magasabb volt az egykori Szovjetunióban, mint más kelet-európai országban, valamint a protestáns, s így erőteljesebben szociáldemokrata Északon mint a katolikus Délen.

6. Hipotézis. Ennek a hipotézisnek az első, a materialista értéktípus (CL7) esetében határozott Kelet–Nyugat különbségeket jósoló része kivétel nélkül megerősítést nyert. Az együttható minden kelet-európai ország, még a szokásos két kivétel, Lengyelország és Litvánia esetében is magasabb volt, mint bármelyik nyugat-európai országnál – bár minden országnál, ismét csak kivétel nélkül, negatív volt. Ez két dologra utal. Először is, amennyiben a materialista értéktípusról van szó, az értékek szorosan követik az anyagi jólétben mutatkozó különbségeket. Másodszor: az Inglehart-féle hipotézist megszorításokkal kell kezelni. Ha ugyanis nem attitűdöket, hanem alapvető értékorientációkat mérünk, akkor a materializmus – egyszerű életciklus-okok következtében – valamivel magasabb lesz a fiatalabb, mint az idősebb generációnál.

Ami a hipotézis második részét illeti, a hedonisztikus típus koefficiensei a várakozásnak megfelelően – Lettország kivételével – szintén negatívak lettek, de az adatokból világosan feltárult egy, az elvárásainktól eltérő minta is. A katolikus, protestáns különbségekhez egyáltalán nem kapcsolódva a hedonizmus a keleti materializmus nyugati ekvivalensének tűnik. Az együtthatók a legtöbb nyugat-európai országnál nagyon magasak, jóval a kelet-európai országok szintje fölött, bár a választóvonal itt nem olyan világos. Ami a kivételeket illeti, keleten találhatók, a lettországi pozitív együttható valószínűleg az erős nemi (gender) hatásnak köszönhető, a moszkvai magas negatív együttható pedig a minta speciális jellegének (itt ugyanis minden válaszoló egy főváros lakója). Nyugaton az alacsony norvég együttható valószínűleg szintén a nemi hatás eredménye, míg a nem várt módon alacsonyabb katolikus országbeli együtthatók arra utalnak, hogy korunkban a hedonizmus inkább a szociáldemokrácia, mint a vallásosság alternatívája.

Úgy tűnhet, hogy a 6. hipotézis eredményei szemben állnak általános hipotézisünkkel. A materialista és a hedonista értéktípus együtthatói csupán a mikroszintű egyéni életcikluselemek és a makroszintű gazdasági jóléti elemek kombinációját tükrözik, s itt nem sok szerepet játszanak az axiális megfontolások. Ezek a negatív eredmények is fontosak azonban, mivel segítenek az értékrendszer azon szegmenseinek azonosításában és izolálásában, amelyek érthető módon függetlenek az axiális megfontolásoktól.

7. Hipotézis. Az útegyüttható megerősítették modellünk specifikációját, bár sokkal inkább Kelet-, mint Nyugat-Európa esetében. Először is, az elhanyagolható lengyelországi fordított értéket (–0,01) leszámítva a „közvetlen” koefficiensek egyike sem volt negatív előjelű. Másodszor, Litvánia, az egykori Csehszlovákia és különösen Lengyelország kivételével a kelet-európai országoknál a „közvetlen” útegyüttható jelentősen magasabb volt, mint a „kereszt”-együttható. Nyugat-Európában ez igaz a születési tényezőre, de a modernizációra általánosságban már nem. Franciaország volt az egyetlen, amely a kelet-európai mintát követte. Harmadszor: ez már azt is jelzi, hogy az iskolázottsági útegyüttható a várakozásnak megfelelően általánosságban véve magasabb volt Keleten, míg az életkor hatása jóval erősebb volt Nyugat-Európában. Ezek az eredmények azt mutatják, hogy a megfelelő háttér- és értékcsoportok közötti szoros kapcsolat csak Kelet-Európa esetében áll fönn, ahol a modernizációs háttérváltozók nagyobb hatással vannak a

modernizációs értéktípusokra, mint a születési háttérváltozók. A modell általános specifikációja az előjeleket illetően, és az erőteljes közvetlen születési hatásnak köszönhetően is ugyanakkor érvényes volt a legtöbb ország esetében.

8. Hipotézis. Ezzel eljutottunk utolsó és messze a legfontosabb hipotézisünkhöz, amelyben egy sor előjel-megfordulást tételeztünk fel a speciális eseteknél. Az eredmények határozottan megerősítették a feltételezéseket. Először is Finnországot (a maga elhanyagolható 0,01-es együtthatójával) leszámítva két olyan országot találtunk (Anglia és Svédország), ahol pozitív volt az életciklus modernizációs háttértényezőjének útegyütthatója ($M_B A_V$). Ugyanilyen előjelváltozás következett be a másik „kereszt"-együtthatónál ($A_B M_V$) is a „tipikus" kelet-európai országokban, majdnem pontosan úgy, ahogy azt általános hipotézisünk előre jelezte. Előjelváltozás mutatkozott az egykori Szovjetunió szívében (Moszkva, Oroszország, Belorusszia), valamint azokban a kelet-európai országokban, amelyek az axiális ellenállás értelmében nem jelentettek kivételt (Bulgária, Magyarország, Csehszlovákia).[14] Lengyelországban és a balti országokban ugyanakkor ilyen előjelváltozás nem következett be, ami azt jelzi, hogy ahol a kommunista rezsimnek sohasem sikerült támogatottságot szereznie, ott az összeomlása sem vezetett generációs konfliktushoz. És végül az egyetlen nyugati ország, amely ugyanazt a mintát mutatta, Franciaország volt, ahogy azt vártuk is. Ez a hipotézis 48 olyan „kereszt"-együtthatót érintett, amely a múltbeli és jelenbeli axiális momentumokkal kapcsolatos kijelentéseinket tesztelte, s e 48-ból csak egy, a Kelet-Németországbeli negatív ABMV-együttható alakult a várakozásainktól eltérően, s ebben az országban, a standard kelet-európai mintától eltérően, nem volt generációs előjelváltozás. Ez a megállapítás azonban aligha meglepő, minthogy az 1989–1990-es változások radikális jellege háttérbe szorította a generációs különbségeket.

21.9. Konklúzió

Az operacionalizált hipotézisek áttekintése után most már visszatérhetünk az általános elméleti vonatkozásokhoz.

1. Az eredmények azt mutatják, hogy az értékek valóban számítanak. Saját mintákkal rendelkeznek, amelyek nem igazodnak az életszínvonalhoz vagy más modernizációs vívmányokhoz. Szoros kapcsolat csak olyan értéktípusoknál marad fönn, amelyek lényegileg kötődnek a jólét, a materializmus vagy anyagiasság és az általános boldogság elvéhez, de az értékrendszer legtöbb területén olyan axiális nézetrendszerek érvényesülnek, amelyek tagolják az értékrendszert. Ami az összes többit illeti, a legnagyobb különbségek a protestantizmus és a kommunizmus hatásából fakadtak. A mi eredményeink így megerősítik az 1982-es Európai Értékvizsgálat megállapításait a katolikus–protestáns osztóvonal meglétéről (Harding és Phillips 1986). Különösen biztató, hogy Lengyelország, mint kiderült, sokkal közelebb áll a nyugat-európai katolikus országokhoz, főként Portugáliához és Spanyolországhoz, mint a többi volt kommunista országhoz. A katolikus–protestáns megosztottságon kívül a nyugat-európai országok között eltéréseket idézett elő még a felvilágosodás racionalizmusának és a szociáldemokráciának a hatása is.

2. Az ellenkezőjét hangoztató állításokkal szemben képesek voltunk kimutatni, hogy a kelet-európai országok közötti fő értékpreferencia-különbségek nem a politikai vagy a gazdasági fejlődés mintáit követik, hanem két axiális körülmény alakította ki őket:

egyrészt egy olyan axiális nézetrendszer fennmaradása, amely képes volt szembehelyezkedni a kommunizmussal (Lengyelországban), és egy általános ellenségesség a kommunizmussal szemben, amely abból fakadt, hogy ezt a rendszert csupán egy külső erőszak teremtette meg, s előzőleg nem következett be a társadalmi rend összeomlása (a balti államokban – főként Litvániában). Ez ahhoz az elég meglepő eredményhez vezet, hogy Oroszország, Belorusszia, Bulgária és Magyarország volt az a négy ország, amely sohasem tért el a standard posztkommunista mintától. Az egyetlen olyan nyugati ország, amely legalábbis a modell modernizációs felében ugyanilyen mintát mutatott, Franciaország volt, a felvilágosodás racionalizmusa és a hivatalos kommunista értékek közötti szoros hasonlóságnak köszönhetően.

3. Utolsó és legellentmondásosabb elméleti elképzelésünk azzal kapcsolatos, hogy korunkat az axiális válság egy momentumaként ismerjük el és azonosítsuk. S bár egy ilyen állítást nem lehet maradéktalanul bizonyítani, az eredmények azt mutatják, hogy mi magunk valóban egy axiális momentumba lépünk be. Adataink szerint az összes axiális nézetrendszer visszaszorulóban van, főként ott, ahol eddig uralkodó pozícióban volt, s a folyamathoz generációs konfliktusok is társulnak. Így például a vallásosság a fiatalabb generáción belül mindenütt alacsonyabb, mint az idősebbek körében, de különösen így van ez a még mindig erősen vallásos régiókban. A felvilágosodás ideológiája és a racionalitás ma a legegyértelműbben a saját szülőhelyén, Franciaországban szorul mind jobban háttérbe. A hivatalos szocialista értékrendszer az egykori Kelet-Európában, s különösen Oroszországban indult gyors bomlásnak, akárcsak egykori erősségeiben és főként Svédországban a szociáldemokrata értékrendszer. És végül nyilvánvaló generációs konfliktusok vannak jelen a puritanizmus és a munkásmozgalom hazájának, Angliának az értékrendszerében is.

Gyakran halljuk, hogy korunkat az ideológia hanyatlása jellemzi. Úgy tűnik azonban, hogy nem egyszerűen a globális, átfogó, makroszintű ideológiák eltűnéséről van szó, hanem azoknak az axiális technikáknak és nézeteknek a felbomlásáról is, amelyek civilizációnk alapját alkották, s amelyek nem a durkheimi és marxi makrotársadalmi szinten, hanem az individuális szubjektumok szintjén működnek, és lelki-szellemi támaszt nyújtanak a distressz időszakaiban. Ezekből a technikákból mára már csak az individualisztikus racionalizmus és a hedonizmus maradt meg. Az erő és biztonság forrása az egyénből áttevődött a gazdasági, társadalmi és politikai alrendszerek magától értetődőnek elfogadott racionális automatizmusaiba. Várnunk kell, míg kiderül, hogy ezek az automatizmusok, a jóléti államnak és a vegyes gazdaságnak ezek a (Sismondi kifejezésével élve) „palliativumai" milyen hosszú ideig tudják késleltetni egy újabb axiális válság eljövetelét.

Függelék: Rokeach Value Test

A) **Célértékek**	**Terminal values**
ROK1 ANYAGI JÓLÉT	A COMFORTABLE LIFE (a prosperous life) (in Hungary: MATERIAL WELL-BEING)
ROK2 BÉKE	A WORLD OF PEACE (free of war and conflict)
ROK3 BOLDOGSÁG	HAPPINESS (contentedness)
ROK4 BÖLCSESSÉG	WISDOM (a mature understanding of life)
ROK5 CSALÁDI BIZTONSÁG	FAMILY SECURITY (taking care of the loved ones)

ROK6 BELSŐ HARMÓNIA	INNER HARMONY (freedom from inner conflict)
ROK7 EGYENLŐSÉG	EQUALITY (brotherhood, equal opportunity for all)
ROK8 MUNKA ÖRÖME	A SENSE OF ACCOMPLISHMENT (lasting contribution) (in Hungary: THE SATISFACTION OF WELL-DONE WORK)
ROK9 VÁLTOZATOS ÉLET	AN EXCITING LIFE (a stimulating, active life)
ROK10 A HAZA BIZTONSÁGA	NATIONAL SECURITY (protection from attack)
ROK11 IGAZI BARÁTSÁG	TRUE FRIENDSHIP (close companionship)
ROK12 IGAZI SZERELEM	MATURE LOVE (sexual and spiritual intimacy)
ROK13 ÉLVEZETES ÉLET	PLEASURE (an enjoyable, leisurely life)
ROK14 EMBERI ÖNÉRZET	SELF-RESPECT (self-esteem)
ROK15 SZABADSÁG	FREEDOM (independence, free choice)
ROK16 A SZÉPSÉG VILÁGA	A WORLD OF BEAUTY (beauty of nature and the art)
ROK17 TÁRSAD. MEGBECSÜLÉS	SOCIAL RECOGNITION (recpect, admiration)
ROK18 ÜDVÖZÜLÉS	SALVATION (saved, eternal life)
B) Eszközértékek	**Instrumental values**
ROK19 ALKOTÓ SZELLEMŰ	IMAGINATIV (daring, creative)
ROK20 BÁTOR, GERINCES	COURAGEOUS (standing up for your beliefs)
ROK21 ELŐÍTÉLETEKTŐL MENT.	BROADMINDED (open-minded)
ROK22 ENGEDELMES	OBEDIENT (dutiful, respectful)
ROK23 ÉRTELMES	INTELLECTUAL (intelligent, reflexive)
ROK24 FEGYELMEZETT	SELF-CONTROLLED (restrained, selfdisciplined) (in Hungary: DISCIPLINED)
ROK25 FELELŐSSÉGTELJES	RESPONSIBLE (dependable, reliable)
ROK26 HATÉKONY	CAPABLE (competent, effective)
ROK27 JÓKEDELYŰ	CHEERFUL (lighthearted, joyful)
ROK28 LOGIKUS GONDOLKODÁSÚ	LOGICAL (consistent, rational)
ROK29 MEGBOCSÁTÓ	FORGIVING (willing to pardon others)
ROK30 ÖNÁLLÓ	INDEPENDENT (self-reliant, selfsufficient)
ROK31 SEGÍTŐKÉSZ	HELPFUL (working for the welfare of others)
ROK32 SZAVAHIHETŐ	HONEST (sencere, truthful)
ROK33 SZERETETTEL TELJES	LOVING (affectionate, tender)
ROK34 TISZTA	CLEAN (neat, tidy)
ROK35 TÖREKVŐ	AMBITIOUS (hard-working, aspiring)
ROK36 UDVARIAS	POLITE (sourteous, well-mannered)

Note: in three cases, the re-translation of the word used in the Hungarian version into English is also given.

Jegyzetek

1. Mai szemmel nézve a lista kissé különösnek tűnhet. A vizsgálat idején azonban Kelet- és Nyugat-Németország még két önálló államként létezett, s külön kezelésük elméleti okokból is indokolt volt; Csehszlovákia akkor még nem vált ketté, s mérete miatt nem lehetett megosztani a mintát; a kutatók ugyanakkor Moszkva esetében egy külön minta létrehozása mellett döntöttek.
2. Az axiális korról lásd Jaspers (1953), Mumford (1952) és Eisenstadt (szerk. 1986). A koegzisztenciát illetően: gyakorlatilag egyidejűség figyelhető meg az indiai Buddha (Kr. e. 560–483), a kínai Konfuciusz (Kr. e. 551–479) és a görög Héraklitosz (Kr. e. 550–480) között.
3. Bár eddig nem sok figyelmet fordítottak rá, ez a fogalom központi szerepet játszik *A világvallások gazdasági etikája* fontos „Bevezetésében" (Gerth és Mills 1948), ahol a szó többször is előfordul, s mindig fontos kontextusban. Norbert Elias volt az a társadalomtudós, aki jelentőségéhez méltó módon foglalkozott a fogalommal, s négyszer is használta *The Court Society* című doktori értekezésében, a Bevezetés záró bekezdéseiben (39–40. oldal).
4. Ezt az aspektust tükrözi az újabban jelentős figyelmet kapott weberi *Lebensführung* fogalom (Hennis 1988; Schluchter 1989). A véletlen folytán ugyanezt a szót adta egyik könyvének címéül Lewis Mumford (1952), aki Webertől függetlenül dolgozott ki egy saját koncepciót az axiális korszakról.
5. Nagyon hasonló elgondolással jelentkezett Norbert Elias, aki az örvény metaforáját használta módszertani főművében (Elias 1987, 45–49). Ezzel összefüggésben érdemes megjegyezni, hogy a híres váltókezelő-hasonlatot Weber csak 1920-ban tette közzé, a „Bevezetés" második kiadásában, amikor korábbi írására reflektált az első világháború utáni stresszben.
6. Az érték vagy a szubjektivitás „típusainak" kutatása szorosan kapcsolódik a „lenyomatképződés" („stamping") fogalmához, mivel a „lenyomatképződés" (stamping) és a „tipizálás" (typing) szó gyakorlatilag szinonimák.
7. A tipizálás és az osztályozás nem ugyanaz. A tipológia Webernek (1949) volt a központi témája, aki hangsúlyozta, hogy minden típust egyénileg kell kidolgozni. Az osztályozás viszont Durkheimnek (1963) volt a fő témája, aki az egész hálózatra helyezte a hangsúlyt. A klaszterelemzés az osztályozás, nem pedig a tipizálás módszere, de egy osztályozási eljárás reprodukálhatja több különböző jelölési (stamping) vagy tipizálási folyamat rétegeit.
8. Az 1990-es WVS-ből átvett kérdések a következők voltak: V116C, V116E és V116F (a barátok, a politika és a vallás relatív fontossága az életben); V336 (részvétel a vallási szertartásokon); V453 (egy sor gyermeknevelési elv: függetlenség, felelősség, fantázia, takarékosság, kitartás, vallási hit, engedelmesség); V532-5 (egy sor Inglehart-féle érték: több beleszólás a kormányzásba, harc az áremelkedés ellen, szólásszabadság, gazdasági stabilitás, humánusabb társadalom, eszmék *kontra* pénz); V565 (és egy sor tétel az engedékenységi teszthez: állami segélyek igénylése, bliccelés a tömegközlekedésben, adócsalás, talált pénz megtartása, házasságon kívüli kapcsolat, a törvényes korhatár alatti szex, prostitúció, abortusz, válás, harcos rendőrség, eutanázia, valamint emberölés önvédelemből). Az alkalmazott eljárást a gyanúsítás (version of imputation) egyik változatának tekinthetjük (Little és Rubin 1987, 21–39).
9. A LISREL-ről információkkal szolgál Long (1983). A LISREL-t a faktorelemzés és az útelemzés kombinációjának tekinthetjük, amelyben a közönséges regressziós

elemzéstől eltérően az útegyütthatók a latens változók (tényezők) között szerepelnek.

10. A két Írországot ki kellett hagyni a további elemzésből, minthogy a válaszolóknak több mint a fele egyetlen értéktípusba került.
11. Ellenkező szemléletet kínál Reisinger et al. (1994).
12. Minthogy az értéktípusok dichotóm változók, nem alkalmazhattuk a tulajdonképpeni LISREL-modellezési programot, minthogy a maximum likelihood becslés feltételezi a normalitást, ezért az LVPLS-módszert (Latent Variables Path Analysis with Partial Least Squares estimation), ezt az igen robusztus programot használtuk. A modell becslése elég ingatag lábakon állt, mivel a Rokeach-típusok nemcsak dichotóm változók, hanem a megoszlásaik is nagyon aránytalanok voltak, s a típusok egyikébe való besorolódás 0 értéket vont maga után az összes többinél. Ennélfogva az értéktípusok közötti „közvetlen" korrelációk eltorzultak, ezért nullával helyettesítettük őket a korrelációs mátrixban. Következésképpen az értéktípusok faktorsúlya pusztán a kapcsolatok struktúrájának a háttérváltozókkal mutatkozó hasonlóságán alapult. Továbbá néhány országban, ahol nagyon kevés válaszoló tartozott a CL8-hoz, a modell instabillá vált. Komparatív célokból ezért a végső modell minden együtthatóját a CL8 nélküli sorozatokból vettük.
    Mindazonáltal a tájékoztatás kedvéért a CL8 elérhető együtthatóit visszahelyeztük a táblázatba.
13. Egyes országokban sajnos semmilyen kódok nem álltak rendelkezésre az urbanizációval kapcsolatban.
14. Bár egyes országokban (Oroszország, Belorusszia, Magyarország és Csehszlovákia) a pozitív koefficiens nagyon kicsi volt, a fordított érték mégis figyelmet érdemel, minthogy az összes többi országban legalább $-0,10$ körül mozgott.

IRODALOM

Abramson, P. R.–Inglehart, R. (1992): Generational replacement and value change in eight West-European societes. *British Journal of Political Science* (22), 183–228.

Aigner, O. J.–Goldberger, A. S. (1977): *Latent Variables in Socio-economic Models. Amsterdam,* North-Holland, 383.

Akaike, H.: Factor Analysis and AIC. *Psychometrika* (52/3), 317–332.

Aldenderfer, M. S.–Blashfield, R. K. (1984): *Cluster Analysis.* Beverly Hills (Cal.), Sage.

Alker, H. R.–Deutsch, K. W.–Soetzel, A. H. (1973): *Mathematical Approaches to Politics.* Elsevier Scientific, 475.

Ambrosi, K.–Hansohm, J. (1987): Ein dynamischer Ansatz zur Repräsentation von Objekten. In: *Operations Research Proceedings 1986.* Berlin, Springer-Verlag.

Anderberg, M. R. (1973): *Cluster Analysis for Applications.* New York, Academic Press.

Andersen, E. B. (1980): *Discrete Statistical Models with Social Science Applications.* Amsterdam, North-Holland, 383.

Anderson, A. J. B. (1971): Ordination methods in ecology. *Journal of Ecology.* (59), 713–726.

Anderson, J. C.–Gerbing, D. W. (1984): The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika* (49), 155–173.

Anderson, T. W. (1954): An estimation of parameters in latent structure analysis. *Psychometrika* (19), 1–10.

Anderson, T. W. (1984): *An introduction to multivariate statistical analysis.* (2nd szerk.). New York, Wiley.

Anderson, T. W. (1987): *The teaching of practical statistics.* John Wiley and Sons, 199.

Anderson, T. W.–Rubin, H. (1956): Statistical inference in factor analysis. Third Berkeley Symp. *Math. Statist and Prob.* (5), 111–150.

Andorka R. (1982): *A társadalmi mobilitás változásai Magyarországon.* Budapest, Gondolat, 327.

Andrews, D. F.–Herzberg, A. M. (1985): *Data.* New York, Springer-Verlag.

Arató M. (é. n.): *Fejezetek a matematikai statisztikából számítógépes alkalmazásokkal.* Számítógépalkalmazási Kutató Intézet Közleményei (22). Budapest, Számítógépalkalmazási Kutató Intézet.

Backhaus, W.–Menzel, R.–Kreissl, S. (1987): Multidimensional scaling of color similarity in bees. *Biol. Cybern.* (56), 293–304.

Barlow, R. E.–Bartholomew, D. J.–Bremner, J. M.–Brunk, H. D. (1972): *Statistical Inference under Order Restrictions.* London, Wiley.

Barnett, S. (1990): *Matrices: Methods and Applications.* Oxford, Oxford University Press.

Bartholomew, D. J. (1980): Factor analysis for categorical data. *J. Roy. Statist. Soc.* (42), 293–321.

Bartholomew, D. J. (1981/a): *Mathematical Methods in Social Science.* Chichester, Wiley.

Bartholomew, D. J. (1981/b): Posteriori analysis of the factor modell. *Br. J. Math. Statist. Psychol.* (34), 93–99.

Bartholomew, D. J. (1983): Latent variable models for ordered categorical data. *Journal Econometrics* (22), 229–243.

Bartholomew, D. J. (1984): The foundations of factor analysis. *Biometrika* (71), 221–232.

Bartholomew, D. J. (1987): *Latent Variable Models and Factor Analysis.* London, Oxford University Press, 193.

Bartlett, M. S. (1950): Tests of significance in factor analysis. *Br. J. Psychol. (Statistical Sect.)* (3) 77–85.

Bartlett, M. S. (1953): *Factor analysis in psychology as a statistician sees it.* Uppsala Symph. Psychol. Factor Analysis. Uppsala, Almqvist and Wiksell, 23–34.

Bénasséni, J. F.–Hays, W. L. (1993): Perturbational aspects in correspondence analysis. *Computational Statistics & Data Analysis* (15), 393–410.

Bennett, J. M. (1987): Influential observations in multidimensional scaling. In: R. M. Heiberger (szerk.): *Proceedings of the* 19th *Symposium on the Interface (Computer Science and Statistics). Am. Stat. Assoc.,* 147–154.

Bentler, P. M. (1980): Multivariate analysis with latent variables: causal modelling, *Annu. Rev. Psychol.* (31), 419–456.

Bentler, P. M. (1982): Linear systems with multiple levels and types of latent variables. In: K. G. Jöreskog–H. Wold (szerk.): *Systems Under Indirect Observation.* Amsterdam, North-Holland.

Bentler, P. M. (1986): Structural Modeling and Psychometrika: An Histirical Perspective on Growth and Achievements. *Psychometrika* (51/1, March), 35–51.

Bentler, P. M.–Bonett, D. G. (1980): Significance test and goodness-of-fit in the analysis of covariance structures. *Psyhological Bulletin* (88), 588–606

Bentler, P. M.–Lee, S. Y. (1978): Statistical aspects of a three-mode factor analysis model. *Psychometrika* (43), 343–352.

Bentler, P. M.–Lee, S. Y. (1979): A statistical development of three-mode factor analysis. *British Journal of Mathematical and Statistical Psychology* (32), 87–104.

Bentler, P. M.–Weeks, D. G. (1978): Restricted multidimensional scaling models. *Journal of Mathematical Psychology* (17), 138–151.

Bentler, P. M.–Weeks, D. G. (1980): Linear structural equations with latent variables. *Psychometrika* (45), 290–308.

Benzécri, J. P. (1992): *Correspondence Analysis Handbook.* New York, Marcel Dekker.

Berge, J. M. F. (1977): Orthogonal Procrustes rotation for two or more matrices. *Psychometrika* (42), 267–276.

Berge, J. M. F. (1983): A generalization of Verhelst's solution for a constrained regression problem in ALSCAL and related MDS-algorithms. *Psychometrika* (48), 631–638.

Berge, J. M. F.–Knol, D. L. (1984): Orhogonal rotations to maximal agreement for two or more matrices of different column orders. *Psychometrika* (49), 49–55.

Berge, J. M. F.–de Leeuw, J.–Kroonenberg, P. M. (1987): Some additional results on principal components analysis of three-mode data. By means of alternating least squares algorithms. *Psychometrika* (52), 183–191.

Berge, J. M. F.–Nevels, K. (1977): A general solution to Mosier's oblique Procrustes problem. *Psychometrika* (42), 593–600.

Blalock, H. M. (1961): Correlation and causality: the multivariate case. *Social Forces* (39), 246–251.

Blalock, H. M. (1963): Making causal inferences for unmeasured variables from correlations among indicators. *Amer. J. Sociol.* (69), 53–62.

Bloxom, B. (1965): *Induviduald ifferences in multidimensional scaling. Princeton University Educational Testing Service Research Bulletin.* Princeton (N. J.) Princeton University Press, 68–145.

Bloxom, B. (1978): Constrained multidimensional scaling in *N* Spaces. *Psychometrika* (43), 283–319.

Bock, R. D.–Aitkin, M. (1981): Marginal maximum likelihood estimation of item parameters: application of an *EM* algorithm. *Psychometrika* (46), 443–459.

Bock, R. D.–Lieberman, M. (1970): Fitting a response model for *n* dichotomously scored items. *Psychometrika* (35), 179–197.

Boolem, K. (1986): Sample size and Bentler and Bonett's nonnormed fit index. *Psychometrika* (51), 375–377.

Boomsma, A. (1985): Nonconvergence, improper solutions, and starting values in Lisrel maximum likelihood estimation.*Psychometrika* (50), 229–242.

Borg, I. (1977): Geometric representation of individual differences. In: J. C. Lingoes (szerk.): *Geometric Representations f Relational Data.* Ann Arbor (Mich.), Mathesis.

Borg, I.–Lingoes, J. C. (1980): A model and algorithm for multidimensional scaling with external constraints on the distances. *Psychometrika* (45), 25–38.

Borg, I.–Lingoes, J. (1987): Multidimensional Similarity Structure Analysis. Springer-Verlag, 390.

Boxom, B. (1968): A note on invariance in three-mode factor analysis.*Psychometrika* (3), 347–350.

Brady, H. E. (1985): Statistical consistency and hypothesis testing for non-metric multidimensional scaling. *Psychometrika* (50), 509–537.

Bricker, C.–Tooley, R. V.–Crone, G. R. (1976): *Landmarks of mapmaking: An Illustrated Survey of Maps and Mapmakers.* New York, Thomas Y. Cromwell.

Brokken, F. B. (1983): Orthogonal Procrustes rotation maximizing congruence. *Psychometrika* (48), 343–349.

Brown, M. B.–Benedetti, J. (1977): On the mean and variance of the tetrachoric correlation coefficient. *Psychometrika* (42), 347–355.

Brown, S. R. (1980): Political Subjectivity: Applications of *Q* Methodology in Political Science. New Haven (Conn.), Yale University Press.

Browne, M. W. (1967): On oblique Procrustes rotation. *Psychometrika* (32), 125–132.

Browne, M. W. (1968): A comparison of factor analytic techniques. *Psychometrika* (33), 267–333.

Browne, M. W. (1982): Covariance structures in Topics in Applied Multivariate Analysis. (D. M. Hawkins szerk.) Cambridge, Cambridge University Press.

Browne, M. W. (1984): Asymptotically distribution-fre methods for the analysis of covariance structures. *Br. J. Math. Statist. Psychol.* (37), 62–83.

Brunner, O. (1992): *Land and Lordship: Structures of Governance in Medieval Austria.* University Park (Pa.), University of Pennsylvania Press.

Bryant E. H.–Atchley, W. R. (1975): *Multivariate Statistical Methods Within-Groups Covariation.* Dowden, Hutchinson and Ross, 436.

Burt, R. S. (1976): Positions in Networks. *Social Forces* (55/1), 93–122.

Büyükkurt, B. K.–Büyükkurt, M. D. (1990): Robustness and smallsample properties of the estimators of probabilistic multidimensional scaling (PROSCAL). *Journal of Marketing Research* (27), 139–149.

C. Lif, N. (1966): Orthogonal rotation to congruence. *Psychometrika* (31), 33–42.

Cailliez, F. (1983): The analytical solution of the additive constant problem. *Psychometrika* (48), 305–308.

Carmone, F. J.–Green, P. E.–Robinson, P. J. (1986): TRICON: an IBM 360/65 program for the triangularisation of conjoint data. *Journal of Marketing Research* (5), 219–220.

Carroll, J. D. (1968): *Generalisation of Canonical Correlation Analysis to three or more sets of Variables.* Proceedings of the 76th Annual Convention of the APA, 227–228

Carroll, J. D. (1972/a): Individual Differences and Multidimensional Scaling. In R. N. Separd–A. K. Romney–S. B. Nerlove (szerk.). *Multidimensional Scaling: Theory and Applications in the Behavioural Sciences* (Vol. 1.) New York, Seminar Press.

Carroll, J. D. (1972/b): Review of Delbeke. *Psychometrika* (35), 178–281.

Carroll, J. D. (1974): *Samon ethodological advences in INDSCAL.* Psychometric Society, Stanford, Mimeo.

Carroll, J. D.–Arabie, P. (1980): Multidimensional scaling. *Ann. Rev. Psychol*, (31), 607–649.

Carroll, J. D.–Chang, J. J. (1964): A General Index of Nonlinear Correlations and its Application to the Problem of Relating Physical and Psychological Dimensions. (Unpublished paper.) Laboratories, Murray Hill (N. J.).

Carroll, J. D.–Chang, J. J. (1967): *Relating Preference Data to Multidimensional Scaling Solutions via a Generalisation of Coombs' Unfolding Model.* Murray Hill (N. J.), Mimeo.

Carroll, J. D.–Chang, J. J. (1970): Analysis of individual differences in multidimensional scaling via an *n*-way generalization of "Eckart-Young" decomposition. *Psychometrika* (35), 283–319.

Carroll, J. D.–Chang, J. J. (1971): *An Alternative Solution to the "metric unfolding problem".* Paper presented at the Psychometric Society Meeting. (April). St. Louis (Miss.).

Carroll, J. D.–Chang, J. J. (1972): *IDIOSCAL (Individual Differences In Orientation Scaling): A generalization of INDSCAL allowing IDIOSYCRATIC reference systems as well as an analytic approximation to INDSCAL.* (Manuscript.) Bell Laboratories, Murray Hill (N. J.) Presented at 1972 Spring meeting of Psychometric Society. Princeton (N. J.) March.

Carroll, J. D.–Chang, J. J. (1973): *Models and Algorithmus for Multidimensional Scaling,* Bell Laboratories, Mimeo.

Carroll, J. D.–Pruzansky, S.–Kruskal, J. B. (1980): CANDELINC: A general approach to multidimensional analysis of many-way arrays with linear constraints on parameters. *Psychometrika* (45), 3–24.

Carroll, J. D.–Wish, M. (1975/a): Models and Methods for three wey multidimensional scaling. In: R. C. Atkinson, D. H. Krantz, Luce–P. Suppes (szerk.): *Contemporary Methods in Mathematical Psychology.* San Francisco, W. H. Freeman.

Carroll, J. D.–Wish, M. (1975/b): Multidimensional perceptual models and measurement methods. In: E. C. Carteret–M. P. Friedman (szerk.): *Handbook of Perception.* (Vol. 2): (Ch. 5: *Individual Differences in Perception*). New York, Academic Press.

Caroll, R. J. (1953): An analytical solution for approximating simple structure in factor analysis. *Psychometrika* (18), 23–38.

Caroll, R. J. (1988): *Transformation and weighting in regression.* Chapmann and Hall, 249.

Catell, R. B. (1952): *Factor Analysis.* New York, Harper and Row.

Cermak, G. W.–Cornillo, P. C. (1976): Multidimensional analyses of judgments about trafic noise. *Journal of the Acoustical Society* (59/6), 1412–1420.

Chang, J. J. (1962): *How to use Paramap.* Murray Hill (N. J.) Bell Telephone Laboratories, Mimeo.

Chang, J.-J.–Caroll, J. D. (1968): Carroll: *How to use PROFIT, a computer program for Property Fitting by optimizing nonlinear or linear correlation.* Murray Hill (N. J.), Bell Laboratories, Mimeo.

Chang, C. L.–Lee, R. C. T. (1973): A heuristic relaxation method for non-linear mapping in cluster analysis. *I. E. E. E. Trans. On Systems, Man. and Cybernetics* (3), 197–200.

Chatfield, C.–Collins, A. J. (1980): *Introduction to Multivariate Analysis.* London, Chapman and Hall.

Chikán, A.–Füstös, L.–Paprika, Z. (1987): *Analysis of Multicriteria Decision on Capital Allocation by Multivariate Statistics.* In: (Gy. Meszéna szerk.) Papers on applications (I.) Department of Mathematics, Karl Marx University of Economics. Budapest (May), 95–133.

Choulakian, V. (1988): Exploratory analysis of contingency tables by log-linear formulation and generalizations of correspondence analysis. *Psychometrika* (53), 235–250.

Cliff, N. (1968): The "idealized individual" interpretation of individual differences in multidimensional scaling. *Psychometrika* (33), 225–232.

Cliff, N. (1987): Analysing multivariate data. San Diego, Harcourt Brace, Jovanovich, 494.

Cliff, N.–Girard, R.–Green, R. S.–Kehoe, J. F.–Doherty, L. M. (1977): INTERSCAL: A TSO FORTRAIN IV program for subject computer interactive multidimensional scaling. *Educational and Psychological Measurement* (37), 185–188.

Clogg, C. C. (1979): Some latent structure models for the analysis of Likert-type data. *Social Sci. Res.* (8), 287–301.

Clogg, C. C. (1980): New developments in latent structure analysis. In: D. J. Jackson–E. F. Borgatta (szerk.): *Factor Analysis and Measurement in Sociological Research.* Beverly Hills (Ca.), Sage, 215–246.

Clogg, C. C.–Goodman, L. A. (1985): Simultaneous Latent Structure Analysis in Several Groups in Sociological Methodology, 81–110.

Cohran, W. G. (1963): *Sampling Techniques.* John Wiley and Sons, 413.

Coleman, J. S. (1964): *Introduction to Mathematical Sociology.* New York, Free Press.

Cooley, P. L. (1973): *Return distributions: A multidimensional analysis of institutional investor perception and preference.* (Unpublished Ph. D. Thesis.). The Ohio State University.

Coombs, C. H. (1950): Psychological scaling without a unit of measurement. *Psychol. Rev.* (27), 148–158.

Coombs, C. H. (1964): *A Theory of Data.* New York, Wiley.

Coombs, C. H.–Kao, R. C. (1960): On a connection between factor analysis and multidimensional unfolding. *Psychometrika* (25), 219–231.

Cooper, L. G. (1972): A new solution to the additive constant problem in metric multidimensional scaling. *Psychometrika* (37), 311–321.

Cormack, R. M. (1971): A review of classification (with Discussion). *J. R. Statist. Soc., A.* (134), 321–367.

Cornelius, F. T. (1973): *The predictive validity of multidimensional spaces.* (Unpublished Ph. D. thesis), Texas Christian University.

Corradino, C. (1990): Proximity structure in a captive colony of Japanese monkeys (*Macaca fuscata fuscata*): an application of multidimensional scaling. *Primates* (31), 351–362.

Coury, B. G. (1987): Multidimensional scaling as a method of assessing internal conceptual models of inspection tasks. *Ergonomics* (30), 959–973.

Cox, C. R.–Hinkley, D. V. (1974): *Theoretical Statistics.* Chapman and Hall, 511.

Cox, D. R.–Dakes, D. (1984): *Analysis of Survival Data Monographs on Statistics and Applied Probability.* Chapman and Hall, 201.

Cox, M. A. A.–Cox, T. F. (1992): Interpretation of stress in nonmetric multidimensional scaling. *Statistica Applicata* (4), 611–618.

Cox, T. F.–Cox, M. A. A. (1990): Interpreting stress in multidimensional scaling. *J. Statist. Comput. Simul.* (37), 211–223.

Cox, T. F.–Cox, M. A. A. (1991): Multidimensional scaling on a sphere. *Commun. Statist.* (20), 2943–2953.

Cox, T. F.–Cox, M. A. A.–Branco, J. A. (1992): Multidimensional scaling for n-tuples. *Brithis Journal of Mathematical and Statistical Psychology* (44), 195–206.

Coxon, A. P. M. (1971): Occupational Attributes: Constructs and Structure. *Sociology* (5), 335–354.

Coxon, A. P. M.–C. L. Jones (1974): Applications of multidimensional scaling, techniques in the analysis of survey data. In: C. J. Payne and C. O'Muircheartaigh (szerk.): *Survey Analysis.* London, Wiley.

Coxon, A. M. P.–C. L. Jones (1978): *The images of occupational prestige.* London, Macmillan.

Cramer, E. M. (1974): On Browne's solution for oblique Procrustes rotation. *Psychometrika* (39), 159–163.

Cramér, H. (1966): *Mathematical Methods of Statistics.* Princeton (N. J.), Princeton University Press, 574.

Cristoffersson, A. (1975): Factor analysis of dichotomized variable. *Psychometrika* (40), 5–32.

Critchley, F. (1978): Multidimensional scaling: a short critique and a new method. In: L. C. A. Corsten–J. Hermans (szerk.), *COMPSTAT 1978.* Vienna, Physica-Verlag.

Cronbach, L. J. (é. n.): Internal Consistency of Tests: Analysis Old and New. *Psychometrika* (53/1), 63–70.

Cronbach, L. J. (1951): Coefficiant alpha and the internal structure of test. *Psychometrika* (16), 297–334.

Cseh-Szombathy L. (1979): *Családszociológiai problémák és módszerek.* Budapest, Gondolat, 402.

Cseh-Szombathy L.–Ferge Zs. (1975): *A szociológiai felvétel módszerei.* Budapest, Közgazdasági és Jogi Könyvkiadó.

Cseh-Szombathy L.–Léderer P. (szerk.) (1973): *Az empirikus szociológiai következtetés statisztikai alapjai.* Budapest, Tankönyvkiadó, 280.

D'Agoston, R. B.–Stephens, M. A. (1986): Goodness-of-fit Techniques. *Statistics: textbooks and monographs* (Vol. 68.), Marcel Dekker, 560.

Davidson, J. A. (1972): A geometric analysis of the unfolding model: nondegenerate solutions. *Psychometrika* (3), 193–216.

Davidson, J. A. (1973): A geometrical analysis of the unfolding model: general solutions. *Psychometrika* (38), 305–336.

Davidson, M. L. (1974): Fitting a set of points to a space defined by a second set. (Unpublished Ph. D. Thesis.) University of Illinois at Urbana-Champaign.

Davidson, M. L. (1983): *Multidimensional scaling.* New York, Wiley, 242.

Davidson, M. L.–Jones, L. E. (1976): A similarity-attraction model for predicting sociometric choice from perceived group structure, *J. Parson and Social Psychol.*

Davies, P. M.–Coxon, A. P. M. (1983): *The MDS(X) User Manual.* University of Edinburgh, Program Library Unit.

Davies, P. M.–Coxon, A. P. M. (szerk.) (1982): *Key Tests in Multidimensional Scaling.* London, Heinemann Educational Books, 347.

Delbeke, L. (1968): *Construction of preference spaces.* Louvain, University of Louvain Press.

De Leeuw, J. A. (1977/a): Applications of convex analysis to multidimensional scaling. In: J. R. Barra, F. Brodeau, G. Romier–B. van Cutsen. (szerk.): *Recent Developments in Statistics.* Amsterdam, North Holland, 133–145.

De Leeuw, J. (1977/b): Correctness of Kruskal's algorithms for monotone regression with ties. *Psychometrika* (42), 141–144.

De Leeuw, J. (1984): Differentiabilility of Kruskal's stress at a local minimum. *Psychometrika* (49), 111–113.

De Leeuw, J. (1988): Convergence of the majorization method for multidimensional scaling. *Journal of Classification* (5), 163–180.

De Leeuw, J. (1992): *Fitting distances by least squares.* (Unpublished report.)

De Leeuw, J.–Heiser, W. (1977): Convergence of correction matrix algorithms for multidimensional scaling. In: J. C. Lingoes (szerk.): *Geometric Representations of Relational Data.* Ann Arbor (Mich.), Mathesis Press.

De Leeuw, J.–Heiser, W. (1980): Multidimensional scaling with restrictions on the configuration. In: P. R. Krishnaiah (szerk.): *Multivariate Analysis* (V.). Amsterdam, North Holland.

De Leeuw, J.–Heiser, W. (1982): Theory of multidimensional scaling. In: P. R. Krishnaiah–L. N. Kana. (szerk.): *Handbook of Statistics* (Vol. 2). Amsterdam, North Holland, 285–316.

De Leeuw, J.–Van der Heijden, P. G. M. (1988): Correspondence analysis of incomplete contingency tables. *Psychometrika* (53), 223–233.

De Leeuw, J.–Stoop, I. (1984): Upper bounds for Kruskal's stress. *Psychometrika* (49), 391–402.

De Leeuw, J.–Young, F. W.–Takene, Y. (1976): Additive structure in qualitative data: an alternating least squares method with optimal scaling feature. *Psychometrika* (41), 471–503.

Derry, W. D.–Lewis-Beck, M. S. (1986): *New Tools for Social Scientists.* Sage.

Desbarat, J. M. (1976): Semantic structure and perceived environment. *Geographical Analysis* (8/4), 453–467.

Desarbo, W. S.–Caroll, J. D. (1985): Three-way Metric Unfolding Via Alternating Weighted Least Squares. *Psychometrika* (50, 3, September), 275–300.

Diday, E.–Simon, J. C. (1976): Clustering analysis. In: K. S. Fu (szerk.): *Communication and Cybernetics 10 Digital Pattern Recognition.* Berlin, Springer-Verlag.

Digby, P. G. N.–Kempton, R. A. (1987): *Multivariate Analysis of Ecological Communities.* London, Chapman and Hall.

Diggle, P. J. (1983): *Statistical Analysis of Spatial Point Patterns.* London, Academic Press.

Dillon, W. R.–Goldstein, M. (1987): *Multivariate Analysis Methods and Applications.* John Wiley and Sons, 587.

Dobson, A. J. (1983): *Introduction to Statistical Modelling.* London, Chapman and Hall.

Draper, N. R.–Smith, H. (1966): *Applied Regression Analysis.* John Wiley and Sons, 407.

Durkheim, E.–Mauss, M. (1963): *Primitive Classification.* Chicago, Chicago University Press.

Eaton, M. L. (1983): *Multivariate Statistics: A Vector Space Approach.* John Wiley and Sons, 512.

Ehrlich É.–Pártos Gy.–Csorbe M.–Szarvas P. (1977): *Fejlettségi szintek, arányok, szerkezetek.* Budapest, Országos Tervhivatal Tervgazdasági Intézet Kiadványa (I–II. füzet).

Eisenstadt, S. N. (szerk.) (1986): *The Origins and Diversity of Axial Age Civilisations.* New York, State University of New York Press.

Ekman, G. (1954): Dimensions of colour vision. *Journal of Psychology* (38), 467–474.

Elias, N. (1994): *The Civilizing Process.* Oxford, Blackwell.

Elias, N. (1983): *The Court Society.* Blackwell, Oxford.

Elias, N. (1987): *Involvement and Detachment.* Oxford, Blackwell.

Éltető Ö.–Ziermann M. (1961): *Matematikai statisztika.* Budapest, Tankönyvkiadó.

Everitt, B. S. (1974): *Cluster Analysis.* London, Heinemann.

Everitt, B. S. (1977/a): An Introduction to Latent Variable Models. *Monographs on Statistics and Applied Probability.* Chapmann and Hall, 128.

Everitt, B. S. (1977/b): The Analysis of Contingency Tables. *Monographs on Statistics and Applied Probability.* Chapmann and Hall, 128.

Fagot, R. F.–Mazo, R. M. (1989): Association coefficients of identity and proportionality for metric scales. *Psychometrika* (54), 93–104.

Fawcett, C. D. (1901): A second study of the variation and correlation of the human skull, with special reference to the Naqada crania. *Biometrika* (1), 408–467.

Feinberg, S. E. (1980): *The Analysis of Cross-Classified Categorical Data.* (Mass.) The Massachusetts Institute of Technology, 198.

Fenton, M.–Pearce, P. (1988): Multidimensional scaling and tourism research. *Annals of Tourism Research* (15), 236–254.

Fielding, A. (1977): Latent Structure Models. 125–57., In: C. Payne–C. A. O'Muircheartaigh (szerk.): *The Analysis of Survey Data.* (Vol. I.): *Exploring data Structures.* New York, John Wiley and Sons, 125–157.

Finney, D. J. (1977): *Probit Analysis.* Cambridge University Press, 333.

Fischer, G. H. (1938): Logistic latent trait models with linear constraints. *Psychometrika* (48), 3–26.

Fisher, R. A. (1940): The precision of discriminant functions. *Ann. Eugen.* (10), 422–429.

Fitzgerald, L. F.–Hubert, L. J. (1987): Multidimensional scaling: some possibilities for counseling psychology. *Journal of Counseling Psychology* (34), 469–480.

Fless, J. L. (1973): *Statistical Methods For Rates and Proportions.* New York, John Wiley and Sons.

Fletcher, R.–Powell, M. J. D. (1963): A rapidly convergent descent method for minimization. *Comput. J.* (2), 163–168.

Forgácsné Kovács E.–Törökné Matits Á. (1986): *A gazdasági adatrendszerek struktúrájának elemzése.* Budapest, Tankönyvkiadó, 132.

Formann, A. K. (é. n.): Latent Class Models for Nonmonotone Dichotomous Items. *Psychometrika* (53/1), 45–63.

Formann , A. K. (1978/a): The Latent class analysis of polytomous data. *Biometrical Journal* (20), 755–771.

Formann, A. K. (1978/b): A note on parameter estimation for Lazarsfeld's latent class analysis. *Psychometrika* (43), 123–126.

Formann, A. K. (1982): Linear logistic latent class analysis. *Biometrical Journal* (24), 171–190.

Formann, A. K. (1985): Constrained latent class models: Theory and applications. *British Journal of Mathematical and Statistical Psychology* (38), 87–111.

Foucault, M. (1976): *Discipline and Punish.* New York, Vintage.

Foucault, M. (1980–1986): *The History of Sexuality* (3 vols). New York, Vintage, 86.

Fuller, E. L.–Hemmerle, W. J. (1966): Robustness of the maximum – likehood estimation procedure in factor analysis. *Psychometrika* (31), 255–266.

Fuller, W. A. (1987): *Measurment Error Models.* John Wiley and Sons, 440.

Füstös L. (1979): *Szociológiai kutatások sokváltozós matematikai statisztikai módszerei.* Budapest, MTA Szociológiai Kutató Intézet Kiadványai, 220.

Füstös L. (1980): *Methods of Measuring Characteristics of Distributions.* Budapest, Center for Value Sociology, Institute for Sociology, The Hungarian Academy of Sciences, Institute for Culture, 60.

Füstös L. (1980–1983): *A sokdimenziós skálázás módszerei: MINISSA, INDSCAL, PREFMAP,PROFIT, PARAMAP, MRSCAL, HICLUS, MINIRSA, MDPREF, UNICON.* Módszertani füzetek (10), Budapest, MTA Szociológiai Kutató Intézet, 341.

Füstös L. (1982): *Klaszterelemzés.* Módszertani füzetek. Budapest, MTA Szociológiai Kutató Intézet, 54.

Füstös L. (1983): *Lineáris egyenletrendszerek általánosított modelljei: LISREL, LVPLS.* Módszertani füzetek. Budapest, MTA Szociológiai Kutató Intézet, 161.

Füstös L. (1984): *Három kívánság* (I–II. kötet.) Módszertani füzetek. Budapest, MTA Szociológiai Kutató Intézet, 98.

Füstös L. (1985): *Loglineáris modell kontingenciatáblák elemzésére.* Módszertani füzetek. Budapest, MTA Szociológiai Kutató Intézet, 54.

Füstös L. (1988/a): *Az adatelemzés statisztikai módszerei.* Módszertani füzetek. MTA Szociológiai Kutató Intézet, 587.

Füstös L. (1988/b): *"Értéktérkép" (16 ország értéktérképe a gyermeknevelési elvek tükrében).* Budapest, Módszertani füzetek. MTA Szociológiai Kutató Intézet, 44.

Füstös L. (1988/c): *Az exploratív faktorelemzés módszerei.* Budapest, Módszertani füzetek. MTA Szociológiai Kutató Intézet, 55.

Füstös L. (1988/d): *The methods of exploratory factor analysis.* Módszertani füzetek. Budapest, MTA Szociológiai Kutató Intézet, 51.

Füstös L.–Kovács E. (1989): *A számítógépes adatelemzés statisztikai módszerei.* Budapest, Tankönyvkiadó, 587.

Füstös L.–Könyves Tóth I. (1988): *Gyermeknevelési elvek. A magyar társadalom és Kővágóőrs: egy helyi társadalom értéktérképének összehasonlító vizsgálata.* Módszertani füzetek. Budapest, MTA Szociológiai Kutató Intézet, 88.

Füstös L.–Manchin R.–Tóth K. (1981): *SZOCPROG 1.3 verzió. Társadalomstatisztikai programrendszer.* Módszertani füzetek. Budapest, MTA Szociológiai Kutató Intézet, 5.

Füstös L.–Meszéna Gy.–Simonné Mosolygó N. (1977): Cluster Analysis. *Szigma* (10/3).

Füstös L.–Meszéna Gy.–Simonné Mosolygó N. (1982): A sokdimenziós skálázás egyes újabb módszerei (I.). *Szigma* (15/3).

Füstös L.–Meszéna Gy.–Simonné Mosolygó N. (1983): *Bevezetés az adatelemzés sokváltozós módszereibe.* Budapest, Tankönyvkiadó, 265.

Füstös L.–Meszéna Gy.–Simonné Mosolygó N. (1985): *Bevezetés az adatelemzés sokváltozós módszereibe.* Budapest, Akadémiai Kiadó, 265.

Füstös L.–Meszéna Gy.–Simonné Mosolygó N. (1986): A sokdimenziós skálázás egyes újabb módszerei (III.) *Szigma* (19/1–2).

Füstös L.–Meszéna Gy.–Ress S.–Simonné Mosolygó N. (1978–1988): Strukturális kapcsolatok általános lineáris modellje (LISREL). *Szigma* (20/1).

Füstös, L.–Meszéna, Gy.–Ress, S.–Simonné Mosolygó, N. (1987): LISREL. Das allgemeine Lineare Modell von Strukturgleichungen. In: *Faktoranalyse* (Beitrage zum 2.) Jena, Anwenderseminar in Jena, Friedrich Shiller Universität, 33–69.

Füstös L.–Mónus Z. (1984): *Bevezetés az adatelemzésbe, Szocprog-PC1.1' verzió, Társadalomstatisztikai programrendszer személyi számítógépre.* Módszertani füzetek. Budapest, MTA Szociológiai Kutató Intézet, 283.

Füstös L.–Paprika Z. (1982): *Innováció nemzetközi mércével.* Módszertani füzetek. Budapest, MTA Szociológiai Kutató Intézet, 25.

Gert, H.–Mills, C. W. (szerk.) (1948): *From Max Weber.* London, Routledge.

Gert, H. (1982): The reception of Max Weber's work in American sociology. In: J. Bensman–A. J. Vidich–N. Gerth (szerk.): *Politics, Character and Culture.* Westport (Conn.), Greenwood Press.

Gibson, W. A. (1959): Three multivariate models: factor analysis, latent structure analysis and latent profile analysis. *Psychometrika* (24), 229–252.

Gilula, Z.–Ritov, Y. (1990): Inferential ordinal correspondence analysis: motivation, derivation and limitations. *International Statistical Review* (58), 99–108.

Girard, R. A.–Cliff, N. (1976): A Monte Carlo evaluation of interactive multidimensional scaling. *Psychometrika* (41), 43–64.

Gleason, T. C. (1969): *Multidimensional scaling of sociometric data.* (Ph. D. Thesis). Ann Arbor (Mich.), University of Michigen.

Gold, E. M. (1973): Metric unfolding: data requirement for unique solution and clarification of Schönemann's algorithm. *Psychometrika* (38), 555–569.

Golderg, D.–Coombs, G. H. (1964): *Some applications of unfolding theory to fertility analysis in Emerging Techiques in Population Research.* Proceedings of the 1962 Annual Conference of the Milban Memorial Fund, New York.

Goodall, D. W. (1967): The distribution of the matching coefficient. *Biometrics* (23), 647–656.

Goodman, L. A. (1974/a): The analysis of systems of qualitive variable when some of the variables are unobservable. (Part I.): A modofied latent structure approach. *American Journal of Sociology* (79), 1179–1259.

Goodman, L. A. (1974/b): Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biomtrika* (61), 215–231.

Goodman, L. A. (1978): *Analyzing Qualitative/Categorical Data. Long-linear models and Latent-Structure Analysis.* London, Addison-Wesley, 467.

Goodman, L. A. (1984): *The Analysis of Cross-Classified Data. Having Ordered Categories.* Cambridge (Mass.) Harvard University Press, 414.

Gordon, A. D. (1981/a): *Classification.* London, Chapman and Hall.

Gordon, A. D. (1981/b): Constructing dissimilarity measure. *Journal of Classification.* (7), 257–269.

Gorsuch, R. L. (1983): *Factoranalysis.* London, Lawrence Erbaum Associates.

Gower, J. C. (1966): Some distance properties of latent root and vector methods in multivariate analysis. *Biometrika* (53), 325–338.

Gower, J. C. (1971): A general coefficient of similarity and some of its properties. *Biometrics* (27), 857–874.

Gower, J. C. (1975): Generalized Procrustes analysis. *Psychometrika* (40), 33–51.

Gower, J. C. (1977): The analysis of asimmetry and orthogonality. In: J. R. Barra *et al.* (szerk.): *Recent Developments in Statistics.* Amsterdam, North Holland.

Gower,. J. C. (1980): Some Characterisations of matrix multidimensional scaling methods. *J. R. Statist. Soc. a* többi adat is kellene.

Gower, J. C. (1984): Multivariate analysis: ordination, multidimensional scaling and allied topics. In: E. H. Lloyd (szerk.), *Handbook of Applicable Mathematics.* (Vol 5.) New York, Wiley.

Gower, J. C. (1985): Measures of similarity, dissimilarity and distance. In: S. Kotz, N. L. Johnson–C. B. Read (szerk.): *Encyclopedia of Statistical Sciences.* (Vol 5.), 397–405.

Gower, J. C. (1990): Fisher's optimal scores and multiple correspondence analysis. *Biometrics* (46), 947–961.

Gower, J. C.–Legendre, P. (1986): Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification* (3), 5–48.

Green, B. F. (1951): A general solution of the latent class model of latent structure analysis and latent profil analysis, *Psychometrika* (6), 151–166.

Green, B. F. (1952): The orthogonal approximation of an oblique structure in factor analysis. *Psychometrika* (17), 429–440.

Green, R. S.–Bentler, P. M. (1979): Improving the effeciency and effectiveness of interactively selected MDS data designs. *Psychometrika* (44), 115–119.

Green, P. E.–Carmone, F. J. (1969): Multidimensional scaling: an introduction and comparison of nonmetric unfolding techniques. *Journal of Marketing Research* (6), 330–341.

Green, P. J.–Sibson, R. (1978): Computing Dirichlet tesselations in the plane. *Computer J.* (21), 168–173.

Greenacre, M. J. (1984): *Theory and Applications of Correspondence Analysis.* London, Academic Press.

Greenacre, M. J. (1988): Correspondence Analysis of multivariate categorical data by weighted least-squares. *Biometrika* (75), 457–467.

Greenacre, M. J.–Browne, M. W. (1986): An efficient alternating least-squares algorithm to perform multidimensional unfolding. *Psychometrika* (51), 241–250.

Greenacre, M. J.–Hastie, T. (1987): The geometrical interpretation of correspondence analysis. *JASA* (82), 437–447.

Greenacre, M. J.–Underhill, L. G. (1982): Scaling a data matrix in a low dimensional Euclidean space. In: D. M. Hawkins (szerk.): *Topics in Applied Multivariate Analysis.* Cambridge (Mass.), Cambridge University Press, 183–268.

Groenen, P. J. F. (1993): *The Majorization Approach to Multidimensional Scaling: Some Problems and Extensions.* Leiden, DSWO Press.

Gruvaeus, G. T.–Jöreskog, K. G. (1970): A computer program for minimizing a function of several variables. *Educational Testing Service Research Bulletin* (RB-70-14).

Guttman, L. (é. n.): A new approach to factor analysis: The Radex Chapter: In: P. F. Lazarsfeld (szerk.): *Mathematical Thinking in the Social Sciences.* New York, Columbia University Press, 258–348.

Guttman, L. (1941): The quantification of a class of attributes: a theory and method of scale construction. In: Horst *at al.* (szerk.): *The Prediction of Personal Adjustment.* New York, Social Science Research Council, 319–348.

Guttman L. (1955): The determinancy of factor score matrices with implications for other basic problems of common factor theory. *Br. J. Statist. Psychol.* (8), 65–82.

Guttman, L. (1968): A general nonmetric technique for finding the smallest coordinate space for a configuration of points. *Psychometrika* (33), 469–506.

Haberman, S. J. (1979): *Analysis of Qualitive Data.* (Vol. 2): New Developments, Academic Press, 612.

Hagenaars, J. A. (1988): *Latent Structure models with Direct Effects Between Indicators: Local Dependence Models in Sociological Methods and Research* (16/3), Sage, 379–405.

Hankiss, E.–Manchin, R.–Füstös, L.–Szakolczai, Á. (é. n.): *Continuity and Break: The Analysis of the Value System of Hungarian Society, 1930–1978.* Budapest, MTA Szociológiai Intézet.

Hankiss, E.–Manchin, R.–Füstös L. (1980): *The Role of Goals in Peorle's Lives.* (Az UNESCO kiadványai), Budapest–Paris, 103.

Hankiss, E.–Manchin, R.–Füstös, L. (1981): *Cross-National QOL Research, An Outline for a Conceptual Framework.* (Az UNESCO kiadványai.) Budapest–Paris, 102.

Hankiss E.–Manchin R.–Füstös L.–Szakolczai Á. (1982/a): *Folytonosság és Szakadás.* Értékszociológiai Műhely kiadványai. Budapest, MTA Szociológiai Kutató Intézet, 604.

Hankiss E.–Manchin R.–Füstös L.–Szakolczi Á. (1982/b): Kényszerpályán? Értékszociológiai Műhely kiadványai, Budapest, MTA Szociológiai Kutató Intézet, 382.

Hankiss, E.–Manchin, R.–Füstös, L.–Szakolczai, Á. (1982/c): *Modernization of Value Systems.* (Symp. on Cultural Indicators), Budapest–Vienna.

Hankiss, E.–Manchin, R.–Füstös, L.–Szakolczai, Á. (1982/d): *The Role of Values in Various Cultural Contexts.* (Az UNESCO kiadványai.) Budapest–Paris, 84.

Hankiss, E.–Manchin, R.–Füstös, L. (1982/e): *The Role of Values and Value Deficiencies in Primary Health Care Recording Systems.* In: M. Lipkin–K. Kupka (szerk.): *Psychosocial Factors.* New York, Affecting Health.

Hansohm, J. (1987): *DMDS dynamic multidimensional scaling.* (Report.) Augsburg, University of Augsburg.

Harding, S.–Phillips, D.–Fogarty, M. (1986): *Contrasting Values in Western Europe: Unity, Diversity and Change.* London, Macmillan.

Harman, H. H. (1976): *Modern factor analysis.* Chicago, University of Chicago Press.

Harris, C. W. (1962): Some Rao-Guttmann relationships. *Psychometrika* (27), 247–263.

Harris, C. W. (1963): Canonical factor models for the description of change. In: C. W. Harris (szerk.): *Problems in measuring change.* Madison (Vis.), University of Visconsin Press.

Harris, C. W. (1967): On factors and factor scores. *Psychometrika* (32), 363–379

Harris, C. W.–Kaiser, H. F. (1964): Oblique factor analytic solutions by orthogonal transformations. *Psychometrika* (29), 347–362.

Harshman, R. A. (1970): Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis. *UCLA Working Papers in Phonetics* (16), 1–84.

Harshman, R. A. (1972): Determination and proof of minimum uniqueness conditions for PARAFAC-1. UCLA *Working Papers in Phonetics* (22).

Harshman, R. A. (1978): *Models for analysis of asymmetrical relationships among N objects or stimuli.* Paper presented at the First Joint Meeting of the Psychometric Society and the Society of Mathematical Psychology. Ontario, Hamilton.

Harsman, R. A.–Berenbaum, S. A (1981): Basic concepts underlying the PARAFAC-CANDECOMP three-way factor analysis model and its application to longitudinal data. In: D. H. Eichorn, J. A. Clausen, N. Haan, M. P. Honzik,–P. H. Mussen (szerk.): *Present and Past in Middle Life.* New York, Acadamic Press.

Harshman, R. A.–Lundy, M. E. (1984/a) Data preprocessing and extended PARAFAC model. In: J. G. Law, C. W. Snyder, J. A. Hattie,–R. P. McDonald (szerk.), *Research Methods for Multimode Data Analysis.* New York, Praeger, 216–284.

Harshman, R. A.–Lundy, M. E. (1984/b) The PARAFAC model for three-way factor analysis and multidimensional scaling. In: H. G. Law, C. W. Snyder, J. A. Hattie,–R. P. McDonald (szerk.), *Research Methods for Multimode Data Analysis.* New York, Praeger, 122–215.

Hartigan, J. A. (1967): Representation of similarity matrices by trees. *J. Am. Statist. Ass.* (62), 1140–1158.

Hawkins, D. M. (1980): *Identification of Outhiers, Monographs on Applied Probability and Statistics.* London–New York, 188.

Hays, W. L.–Bennett, J. F. (1961): Multidimensional unfolding: determining configuration from complete rank order preference data. *Psychometrika* (26), 221–238.

Healy, M. J. R. (1986): *Matrices for Statistics.* Oxford, Clarendon Press.

Hearn, D.–Baker, M. P. (1986): *Computer Graphics.* Prentice-Hall, International, 352.

Hefner, R. A. (1958): *Extensions of the law of comparative judgement to discriminable and multidimensional stimuli.* (Doctoral dissertation.) University of Michigan.

Heiser, W. J. (1991): A generalized majorization method for least squares multidimensional scaling of pseudodistances that may be negative. *Psychometrika* (56), 7–27.

Hennis, W. (1988): *Max Weber: Essays in Reconstruction.* London, Allen.

Hettmansperger, T. P.–Thomas, H. (1973): Estimation of *J* scales for unidimensional unfolding. *Psychometrika* (38), 269–284.

Heywood, H. B. (1931): On finite sequences of real numbers. Proc. Roy. Soc., (Ser. A/134), 486–510.

Hidy P.–Kovács E. (1986): *A lokális döntések természetéről.* Budapest, Művelődéskutató Intézet.

Hill, M. O. (1973): Reciprocal averaging: An eigenvector method of ordination. *J. Ecol.* (61), 237–251.

Hill, M. O. (1974): Correspondence analysis: a neglected multivariate method. *Appl. Statist.,* (23), 340–354.

Hirschfeld, H. O. (1935): A connection between correlation and contingency. *Cambridge Phil. Soc. Proc.* (31), 520–524.

Hoppál, M.–Szecskő T. (1987): *Értékek és változások* (1–2. kötet.) Budapest, Tömegkommunikációs Kutatóközpont, 277.

Horan, C. B. (1969): Multidimensional scaling: combining observation when individuals have different perceptual structure. *Psychometrika* (34, 2, pt.1), 139–165.

Horst, P. (1935): Measuring complex attitudes. *J. Social Psychol* (6), 369–374.

Horst, P. (1965/a): *Factor Analysis of Data Matrices.* New York, Holt, Rinehart and Winston.

Horst, P. (1965/b): Factor Analysis of Data Matrix. New York, Holt, Rinehart and Vinston, 1965.

Hotelling, H. (1933): Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* (24), 417–441, 498–520.

Howe, W. G. (1955): *Some Contributions to Factor Analysis.* Oak Ridge, Oak Ridge National Laboratory.

Ihara, M.–Kano, Y. (1986): A New Estimator of the Uniquesness in Factor Analysis. *Psychometrika* (51/4, December), 563–566.

Inglehart, R. (1977): *The Silent Revolution.* Princeton (N. J.) Princeton University Press.

Jackson, D. A., Somers, K. M.–Harvey, H. H. (1989): Similarity coefficients: measures of co-occurence and association or simply measures of occurence. *Am. Nat.* (133), 436–453.

Jackson, D. N.–Messick, S. J. (1963): Individual differences in social perception. *British Journal of Social Clinical Psychology* (2), 1–10.

Jackson, M. (1989): *Michael Jackson's Malt Whisky Companion: A Connoisseur's Guide to the Malt Whiskies of Scotland.* London, Dorling Kindersley.

Jánossy F. (1963): *A gazdasági fejlettség mérhetősége és új mérési módszerei.* Budapest, Közgazdasági és Jogi Könyvkiadó.

Jánossy F. (1975): *A gazdaságfejlődés trendvonaláról.* Budapest, Magvető.

Jánossy, L. (1965): *A valószínűségelmélet alapjaés néhány alkalmazása – különös tekintettel mérési eredmények kiértékelésére.* Budapest, Tankönyvkiadó, 206.

Jánossy, L. (1968): *Mérési eredmények kiértékelésének elmélete és gyakorlata.* Budapest, Akadémiai Kiadó, 527.

Jardine, N.–Sibson, R. (1971): *Mathematical Taxonomy.* London, Wiley.

Jaspers, K. (1953): *The Origin and Goal of History.* New Haven (Conn.), Yale University Press.

Jennrich, R. I.–Robinson, S. M. (1969): A Newton-Raphson algorithm for maximum likelihood factor analysis. *Psychometrika* (34), 111–123.

Johnson, R. A.–Wichern, D. W. (1982): *Applied Multivariate Analysis.* Englewood Cliffs (N. J.), Prentice-Hall.

Johnson, Stephen C. (1967): *A Simple Cluster Statistic.* (Unpublished paper.) Murray Hill (N. J.), Bell Laboratories.

Johnston, J. N. (1976): Typology formation across socio-economic indicators. *SociologicaEconomics* (10/4), 167–171.

Jones, L-E.–Young, F. W. (1971): *A LongitudinaIndividual Differences Scaling of the L. L. Thurstone Psychometric Laboratory.* University of North Carolina.

Jones, R. A.–R. D. (1973): Ashmore: The structure of intergroup perception. *J. Pers. and Soc. Psichol.* (25), 428–438.

Jones, R. A.–Rosenberg, S (1974): Structural representations of naturalistic descriptions of personality. *Multiv. Beh. Res.* 217–230.

Jonson, S. C. (é. n.): *A simple clustering statistic.* Mimeo, Bell Laboratories.

Jöreskog, K. G. (1967): Some contributions to maximum likelihood factor analysis. *Psychometrika* (32), 443–482.

Jöreskog, K. G. (1969): A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika* (34), 183–220.

Jöreskog, K. G. (1970): A general method for analysis of covariance structures. *Biometrika* (57), 239–251.

Jöreskog, K. G. (1971/a): Simultan Factor Analysis in Several Popualtions. *Psychometrika* (36), 409–426.

Jöreskog, K. G. (1971/b): Statistical Analysis of Congeneric Test. *Psychometrika* (36), 109–133.

Jöreskog, K. G. (1973): General Method for Estimations a Linear Structural Equation Systems. In: A. S. Goldberger–O. D. Duncan (szerk.): *StructuraEquation Models in the SociaSciences.* New York, Seminar Press, 85–122.

Jöreskog, K. G. (1977): Structural equation models in the social sciences: Specification, estimation and testing. In: P. R. Krishnaiah (szerk.): *Applications of Statistics.* Amsterdam, North-Holland.

Jöreskog, K. G. (1979): Basic ideas of factor and component analysis. In: K. G. Jöreskog–D. Sorbom (szerk.): *Advances in Fantos Analysis and StructuraEquation Models.* Cambridge (Mass), Abt Books.

Jöreskog, K. G.–Goldberger, A. S. (1972): Factor analysis by generalized least squares. *Psychometrika* (37), 243–260.

Jöreskog, K. G.–Sörbom, D. (1977): Statistical models and mathods for analysis of longitudinal data. In: D. J. Aigner–A. S. Goldberger (szerk.): *LatenVariables in Socioeconomic Models.* Amsterdam, North-Holland, 285–325.

Kaiser, H. F. (1958): The varimax criterion for analytic rotation in factor analysis. *Psychometrika* (23), 187–200

Kaiser, H. F. (1963): Image analysis. In: C. W. Harris (szerk.): *Problems in measuring change.* Madison (Wis.), Unicersity of Wisconsin Press.

Kelly, M. J.–Wooldridge, L.–Hennessy, R. T.–Vreuls, D.–Barneby, S. F.–Cotton, J. C.–Reed, J. C. (1979): *Aircombat maneuvering performance measurement. Williams Air Force Base, AZ: Flying Training Division.* Air Force Human Resources Laboratory (NAVTRAEQUIPCEN IH 315/AFHRL-TR-79-3).

Kendall, D. G. (1971): Seriation from abundance matrices. In: Hodson et al. (szerk.): *Mathematics in the Archaelogical and HistoricalSciences.* Edinburgh, Edinburgh University Press.

Kendall, D. G. (1977): On the tertiary treatment of ties. Appendix to Rivett, B. H. P., Policy selection by structural mapping. *Proc. R. Soc. Lond.* (354), 422–423.

Kendall, M. G. (1975): *A Course in Multivariate Analysis.* London, Griffin.

Kendall, M. G.–Babington Smith, B. (1950): Factor analysis. *J. Roy. Statist. Soc.* (12), 60–94.

Kendall, M. G.–Lawley, D. N. (1956): The principles of factor analysis. *J. Roy. Statist. Soc.,* (119), 83–84.

Kiers, H. A. L. (1989): An alternating least squares algorithm for fitting the two- and three-way DEDICOM model and the IDIOSCAL model. *Psychometrika* (54), 515–521.

Kiers, H. A. L.–ten Berge, J. M. F.–Takane, Y.–de Leeuw, J. (1990): A generalization of Takane's algorithm for DEDICOM. *Psychometrika* (55), 151–158.

Kiers, H. A. L.–Krijnen, W. P. (1991): An efficient algorithm for PARAFAC of three-way data with large numbers of observation units. *Psychometrika* (56), 147–152.

Kindler J.–Papp O. (1977): *Komplex rendszerek vizsgálata.* Budapest, Műszaki Könyvkiadó, 262.

Kish, L. (1987): *StatisticalDesign for research.* John Wiley and Sons, 267.

Klahr, D. (1969): A Monte Carlo investigation of the statistical significance of Kriskal's nonmetric scaling procedure. *Psychometrika* (34), 319–330.

Kleijen, J. P. C. (1987): *Statistical Tools for Simulation Practitioners.* New York–Basel, Marcel Dekker, 429.

Koemler, K. (1986): Goodness-of-fit test for log-linear models in sparse contingency tables. *J. of the Amer. Stat. Asso.* (81). 483–493.

Kolosi, T. (1984): Státusz és réteg. In: *Rétegződésmodell-vizsgálat* (III.) Budapest, MSZMP KB Társadalomtudományi Intézet, 280.

Kolosi, T.–Rudas, T. (1988): *Empirikus problémamegoldás a szociológiában.* Budapest, OMIKK–TÁRKI, 213.

Konrád, Gy.–Szelényi, I. (1979): *The Intellectuals on Road to Class Power.* New York, Harcourt. (Magyarul: *Az értelmiség útjaz osztályhatalomhoz.* 1989. Budapest, Gondolat. 328.)

Koopman, R. F. (1978): On Bayesian estimation in unrestricted factor analysis. *Psychometrika* (43), 109–110.

Korth, B.–Tudker, L. R. (1976): Procrustes matching by congruence coefficients. *Psychometrika* (41), 531–535.

Koschat, M. A.–Swayne, D. F. (1991): A weighted Procrustes criterion. *Psychometrika* (56), 229–239.

Krane, W. R.–McDonald, R. P. (1978): Scale invariance and the factor analysis of correlation matrices. *Br. J. Math. Statist. Pasychol.* (31), 218–228.

Kristof, W.–Wingersky, B. (1971): Generalization of the orthogonal Procrustes rotation procedure to more than two matrices. In: *Proceedings.* 79th *Annual Convention of the American Psychological Association*, 81–90.

Kroonenberg, P. M. (1981): *Users Guide to TUCKALS3. A program for Three-Mode Principal Component Analysis WEP-REEKS, WT 81-6RP.* Leiden, The Netherlands, University of Leiden.

Kroonenberg, P. M. (1983): *Three-Mode Principal Components Analysis.* Leiden, The Netherlands, DSWO Press.

Kroonenberg, P. M.–de Leeuw, J. (1980): Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika* (45), 69–97.

Kruskal, J. B. (1964/a): Multidimensional Scaling by optimising goodness of fit to a nonmetric hypotheses. *Psychometrika* (29), 1–27, 115–129.

Kruskal, J. B. (1964/b): Nonmetric multidimensional scaling: a numerical method. *Psychometrika* (29), 115–129.

Kruskal, J. B. (1966): Geometric Interpretation of Diagnostic Data from a Digital Machine. *Bess Syst. Tech. J.* (45), 1299–1338.

Kruskal, J. B. (1971): Monotone regression: continuity and differentiability properties. *Psychometrika* (36), 57–62.

Kruskal, J. B. (1972): *A brief description of the "clasical" method of multidimensional scaling.* Bell Telephone Laboratories, Mimeo.

Kruskal, J. B. (1978): Factor analysis and principal components: Bilinear methods. In: W. H-Kruskal–J. M. Tanur (szerk.): *International Encyclopedia of Statistics.* New York, Free Press.

Kruskal, J. B.–Carroll, J. D. (1968): *Geometric models and badness-of-fit functions from Multivariate Analysis.* (Vol. 2.), New York, Academic Press.

Kruskal, J. B.–Wish, M. (1978): *Multidimensional Scaling.* Beverly Hills (Ca.), Sage.

Krzanowski, W. J. (1988): *Principles of Multivariate Analysis: A User's Perspective.* Oxford, Clarendon Press.

Krzanowski, W. J. (1993): Attribute selection in correspondence analysis of incidence matrices. *Appl. Statist.* (42), 529–541.

Langeheine, R. (1982): Statistical evaluation of measures of fit in the Lingoes-Borg Procrustean individual differences scaling. *Psychometrika* (47), 427–442.

Langron, S. P.–Collings, A. J. (1985): Perturubation theory for generalized Procrustes analysis. *J. R. Statist. Soc. B.* (47), 277–284.

Lapointe, F. J.–Legendre, P. (1994): A classification of pure malt Scotch whiskies. *Appl. Statist.* (43), 237–257.

Law, H. G.–Sbyderm, C. W.–Hattie, J. A.–McDonald, R. P. (szerk.) (1984): *Reearch Methods for Moltimode Data Analysis.* New York, Praeger.

Lawley, D. N.–Maxwell, A. E. (1971): *Factor Analysis as a Statestical Method.* London, Butterworths, 153.

Lawson, W. J.–Ogg, P. J. (1989): Analysis of phenetic relationships among populations of the avian genus Batis (Platysteirinae) by means of cluster analysis and multidimensional scaling. *Biom. J.* (31), 243–254.

Lazarsfeld, P. F.–Henry, N. W. (1968): *Latent Structure Analysis.* Boston, Houghton-Mifflin.

Lazarsfeld, P. F.–Obershall, A. R. (1965): Max Weber and empirical social research. *American Sociological Review* (30), 185–199.

Lee, H. B.–Comrey, A. L. (1978): An empirical comparison of two minimum residual factor extraction methods. *Multivariate Behavioral Res.* (13), 497–507.

Lee, S.-Y. (1981): A Bayesian approach to confirmatory factor analysis. *Psychometrika* (46), 153–176

Lee, S.-Y. (1984): Multidimensional scaling models with inequality and equality constraints. *Commun. Statist.-Simula. Computa.* (13), 127–140.

Lee, S.-Y.–Bentler, P. M. (1980): Functional relations in mutidimensional scaling. *British Journal of Mathematical and Statistical Psychology* (33), 142–150.

Leik, R. K.–Meeker, B. T. (1975): *Mathematical Sociology.* Prentice-Hall, 242.

Levine, D. M. (1978): A Monte Carlo study of Kruskal's variance based measure on stress. *Psychometrika* (43), 307–315.

Lingoes, J. C.–Borg, I. (1976): Procrustean individual differences scaling: PINDIS. *Journal of Marketing Research* (13), 406–407.

Lingoes, J. C.–Borg, I. (1977): Procrustean individual differences scaling: PINDIS. *Sozial Psychologie* (8,) 210–217.

Lingoes, J. C.–Borg, I. (1978): A direct approach to individual differences scaling using increasingly complex transformations. *Psychometrika* (43), 491–519.

Lingoes, J. C.–Roskam, E. E. (1927): An algorithm for multidimensional scaling of data by metric transformation of data. *Program Bulletin No. 23. Math. Psychol. Group,* Holland, Department of Psychology, University of Nijmegen.

Lingoes, J. C.–Roskam, E. E. (1969/a): A Comparison of principles for algorothm construction in nonmetric scaling. *Mich. Math. Psychol. Program,* Techical Report, 69–72.

Lingoes, J. C.–Roskam, E. E. (1969/b): Data theory and algorihms for nonmetric scaling, (Part I and Part II), Technical report. (Math. Psychol. Group.) Department of Psychology. Holland, University of Nijmegen.

Lingoes, J. C.–Roskam, E. E. (1969/c): Minissa-1: a Fortran-IV (g) Program for the Smallest Space Analysis of Square Symmetric Matrices. *Behavioral Science* (15), 204–205.

Lingoes, J. C.–Roskam, E. E. (1970/a): Fortran-e (g) Program for the Smallest Space Analysis of Square Symmetric Matrices. *Behavioral Science* (15), 204–205.

Lingoes, J. C.–Roskam, E. E. (1970/b): The method of triads for nonmetric multidimensional scaling. *Nederlands Tijdschrift Voor De Psychologie* (25), 404–417.

Lingoes, J. C.–Roskam, E. E. (1971): Multidimensional scaling of conditional similarity data. Program Bulletin (20), *(Math. Psychol. Group),* Holland, Department of Psychology, University of Nijmegen.

Lingoes, J. C.–Roskam, E. E. (1972): Multidimensional scaling by metric transformation of data. *Nederlands Tijdschrift Voor De Psychologie* (27) 486–508.

Lingoes, J. C.–Roskam, E. E. (1973/a.): A mathematical and empirical analysis of two multidimensional scaling algorithms. *Psychometrika Monograph Supplement* (38/4) Part 2.

Lingoes, J. C.–Roskam, E. E. (1973/b.): A mathematical and empirical study of two multidimensional scaling algorithms. *(Mich. Math. Psychol. Program)*, Technical Report.

Lingoes, J. C.–Roskam, E. E. (1977): Nonmetric data analysis, general methodology and techniques with brief descriptions of Mini-programs. Report (75) (Ma.) 13. Department of Psychology, University of Nijmegen.

Lissitz, R. W.–Schönemann, P. H.–Lingoes, J. C. (1976): A solution to the weighted Procrustes problem in which the transformation is in agreement with the loss function. *Psychometrika* (41), 547–550.

Little, R. J. A.–Rubin, D. B. (1987): *Statistical Analysis with Missing Data.* New York, John Wiley.

Loehlin, I. C. (1987): *Latent Variable Models: An Introduction to Factor, Path and Structural Analysis.* Hollsdale (N. J.), Erlbaum, 273.

Long, J. S. (1983): *Covariance Structure Models: An Itroduction to LISREL.* Beverly Hills (Cal.).

MacCallum, R. C. (1976): Effects on INDSCAL of non-orthogonal perceptions of object space dimensions. *Psychometrika* (41), 177–188.

MacCallum, R. C. (1977a): Effects of conditionality on INDSCAL and ALSCAL weights. *Psychometrika* (42), 297–305.

MacCallum, R. C. (1977b): A Monte Carlo investigation of recovery of structure by ALSCAL. *Psychometrika* (42), 401–428.

MacCallum, R. C. (1978): Recovery of structure in incomplete data by ALSCAL. *Psychometrika* (44), 69–74.

MacCullen, R. (1983): A comparison of factor analysos programs in SPSS, BMDP, and SAS. *Psychometrika* (48), 223–231.

MacCallum, R. C.–Cornelius III, E. T. (1977): A Monte Carlo investigation of recovery of structure by ALSCAL. *Psychometrika* (42), 401–428.

Mardia, K. V. (1978): Some properties of classical multidimensional scaling. *Commun. Statist. Theor. Meth.* (A/7), 1233–1241.

Mardia, K. V.–Kent, J. T.–Bibby, J. M. (1982): *Multivariate Analysis.* London, Academic Press.

Marton Á. (1982): Robusztusság a statisztikában. *Statisztikai Szemle* (60, 8–9), 905–909.

Marton Á.–Vincze I. (1983): A matematikai statisztika a gazdasági és társadalmi jelenségek vizsgálatában. *Statisztikai Szemle* (61/1), 43–58.

Masters, G. N. (1985): A Comprasion of Latent Trait and Latent Class Analysis of Likert-type Data. *Psychometrika* (50, 1, March), 69–82.

Masters, G. N.–Wright, B. D. (1984): The essential process in a family of measurement models. *Psychometrika* (49), 529–544.

Maxwell, A. E. (1977): *Multivariate Analysis in Behavioural Rearch.* London, Chapman and Hall.

McCullagh, P.–Nelder, J. A. (1983): *Generalized Linear Models, Monographs on Statistics and Applied Probability.* London, Chapman and Hall, 261.

McCutcheon, A. (1967): *Latent Class Analysis.* Beverly Hills (Ca.), Sage.

McDonald, R. P. (1974): The measurement of factor indeterminancy. *Psychometrika* (39), 203–222.

McDonald, R. P. (1985): *Factor Analysis and Related Methods.* Lawrenc Erlbaum Associates, 259.

McDonald, R. P.–Burr, E. J. (1967): A Comparsion of four Methods of Constructing Factor Score, *Psychometrika* (32), 381–401.

McElwain, D. W.–Keats, J. A. (1961): Multidimensional unfolding: some geometrical solutions. *Psychometrika* (26), 325–332.

Mead, A. (1992): Review of the development of multidimensional scaling methods. *The Statistician* (41), 27–39.

Messick, S. J. (1956): Perception of Social Attitudes. *Journal of Abnormal Social Psychology* (52), 57–69.

Messick, S. J.–Abelson, R. P. (1956): The additive constant problems in multidimensional scaling. *Psychometrika* (21), 1–17.

Meszéna Gy. (szerk.) (1984): *Sztochasztikus módszerek a döntéselőkészítésben.* Budapest, Tankönyvkiadó, 252.

Meszéna Gy.–Zimmermann M. (1981): *Valószinűségelmélet és matematikai statisztika.* Budapest, Közgazdasági és Jogi Könyvkiadó, 554.

Mill, M. O. (1974): Correspondence Analysis, A neglected multivariate method. *Applied Statistics* (23), 340–354.

Miller, J. E.–Shepard, R. N.–Chang, J.-J. (1964): An analytical approach to the interpretation of multidimensional scaling solutions. Paper Presented at 1964 meeting of A. P. A *Abstract in American Psychologist* (19), 579–580.

Mislevy, R. J. (1984): Estimating latent distributions. *Psychometrika* (49), 359–381.

Mislevy, R. J. (1985): Estimating of latent group effects. *J. Am. Statist. Assoc.* (80), 993–997.

Mislevy, R. J. (1986): Recent developments in the factor analysis of catagorical variables. *J. of Educational Statistics* (11), 3–31.

Mooijaart, A. (1983): Two kinds of factor analysis for ordered categorical variables. *Multivariate Behavioural Res.* (18), 423–441.

Mooijaart, A. (1985): Factor analysis for non-normal variables. *Psychometrika* (50), 323–342.

Mooijaart, A.–Commandeur, J. J. F. (1990): A general solution of the weighted orthonormal Procrustes problem. *Psychometrika* (55), 657–663.

Móri F. T.–Székely J. G. (szerk.) (1986): *Többváltozós statisztikai analízis.* Budapest, Műszaki Könyvkiadó, 393.

Morrison, D. F. (1967): *Multivariate Stetistical Methods.* New York, McGraw-Hill.

Mulaik, S. A. (1986): Factor analysis and Psychometrika: major developments. *Psychometrika* (51), 23–33.

Mulaik, S. A. (1872): *The Foundations of Factor Analysis.* New York, McGraw-Hill.

Mulaik, S. A.–McDonald, R. P. (1978): The effect of additional variables on factor indeterminacy in models with a single ccommon factor. *Psychometrika* (43), 177–192.

Mumford, L. (1952): *The Conduct of Life.* London, Secker and Warburg.

Munck, I. M. E. (1979): Model Building in Comparative Eduction, Applications of the LISREL Method to Cross-National Survey Data. *International Association for the Evaluation of Educational Achievement IEA Monograph Studies* (10). Stockholm, Almqvist and Wiksell International, 199.

Mundruczó Gy. (1979): *Sztochasztikus modellek közgazdasági alkalmazásának kérdései, különös tekintettel a mérésh ibákra.* (Kandidátusi értekezés.) Budapest, BKE, 155.

Mundruczó Gy.–Kerékgyártó Gy.-né (1975): *Alkalmazott regressziószámítás.* Budapest, Tankönyvkiadó, 193.

Muthén, B. (1976): Contributions to factor analysis of dichotomous variable. *Psychometrika* (43), 551–560.

Muthén, B. (1979): A structural probit model with latent variables. *J. Amer. Statist. Ass.* (74), 807–811.

Muthén, B. (1984): A General structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika* (46), 115–132.

Muthén, B.–Christoffersson, A. (1981): Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika* (46), 485–500.

Muthén, B.–Kaplan, D. (1985): A comporison of some methodogics for the factor analysis of non-normal likert variables. *British Journal of Mathematical and Statistical Psychology* (38), 171–189.

Namboodiri, N. K.–Carter, L. F.–Blalock, H. M. (1975): *Applied Multivariate Analysis and Experimental Design.* McGraw-Hill, 688.

New Geographical Digest (1986): London, George Philip.

Nietzsche, F. (1967): *On the Genealogy of Morals.* New York, Vintage, New York.

Nishisato, S. (1980): *Analysis of Categorical Data: Dual Scaling and its Applications.* Toronto, University of Toronto Press.

Nygren, T. E. (1975): *Individual differences in perceptions of political candidates.* (Unpublished M. Sc. Thesis.) University of Illinois at Urbana-Champaign.

Oestreich, G. (1982): *Neostoicism and the Early Modern State.* Cambridge, Cambridge University Press.

O'Hare, D. (1976): Individual differences in perceived similarity and preference for visual art: a multidimansional scaling analysis. *Perception and psychophysics* (20), 445–452.

Olsham, K. M. (1971): *Ther multidimensional structure of person perception in children.* (Unpublished Ph. D. Thesis.) New Brunswick (N. J.) Rutgers University, The State University, The State University of New Jersey.

O'Muircheartaigh, C. A.–Payne, C. (szerk.) (1977): *Analysis of Survey Data.* (Vol. 1): *Exploring Data Structures* (A. Filding: *Latent Structure Models:* Chapter 5, 125–157), (Vol. 2): *Model Fitting.* John Wiley és Sons, 273, 255.

Orlóci L. (1975): *Multivariate Analysis in Vegetation Research.* The University of Western Ontario, The Hague, Dr. W. Junk B. V., 276.

Orlóci, L.–Kenekel, N. C. (1983): *Introdution to Data Analysis with Applications in Poulation and Community Biology. UWO Biology (*224/a, 352/b). London–Ontario, The Univeriuty of Western Ontario.

Pack, P.–Jolliffe, I. T. (1992): Influence in correspondence analysis. *Appl. Statist.* (31), 365–380.

Pan, G.–Harris, D. P. (1991): A new multidimensional scaling technique based upon associations of triple objects – Pijk and its application to the analysis of geochemical data. *Mathematical Geology* (6), 861–886.

Parsons, T. (1967): *The Structure of Social Action.* New York, The Free Press.

Peay, E. R. (1988): Multidimensional rotation and scaling of configurations to optimal agreement. *Psychometrika* (53), 199–208.

Perreault, W. D.–F. A. (1976): Russ: Physical distribution service in industrial purchase decisions. *Journal of Marketing* (40/2), 3–10.

Pirenne, H. (1939): *Mohamed and Charlemagne.* London, Allen.

Plackett, R. L. (1981): *The Analysis of Categorical Data.* London, Griffin.

Plewis, J. (1985): *Analysing Change, Measurement and Explanation Using Longitudinal Data.* John Wiley és Sons, 182.

Polzella, D. J.–Teid, G. R. (1989): Multidimensional scaling analysis of simulated air combat maneuvering performance data. *Aviat. Space, Environ. Mszerk.* (60), 141–144.

Preece, P. F. W. (1976): Mapping cognitive structure – Comparison of methods. *Journal of Educational Psychology* (68/1) 1–8.

Prékopa A. (1972): *Valószínűségelmélet műszaki alkalmazásokkal.* Budapest, Műszaki Könyvkiadó.

Prékopa A.–Éltető Ö. (1961): *Matematikai jegyzetek* (IV. rész.) Budapest, Matemetikai Stetisztika, Statisztikai Kiadó.

Ramsay, J. O. (1977): Maximum likelihood estimation in multidimensional scaling. *Psychometrika* (42), 241–266.

Ramsay, J. O. (1978/a): Confidence regions for multidimensional scaling analysis. *Psychometrika* (43), 145–160.

Ramsay, J. O. (1978/b): MULTISCALE: *Four Programs of Multidimenstional Scaling by the Method of Maximum Likelihood.* Chicago, International Educational Services.

Ramsay, J. O. (1980): Some small sample results for maximum likelihood estimation in multidimensional scaling. *Psychometrika* (45), 141–146.

Ramsay, J. O. (1982): Some statistical approaches to multidimensional scaling data. *J. R. Statist. Soc., A.* (145), 285–312.

Rao, C. R. (1955): Estimacion and tests of sugnificance in factor analysis. *Psychometrika* (20), 93–111.

Rash, G. (1980): *Probabilistic models for some intelligence and attainments test* (2nd szerk.). Chicago, University of Chicago Press.

Reeb, M. (1959): How People see Jobs: a Multidimensional Analysis. *Occup. Psychol.* (33), 1–17.

Reisinger, W. M.–Miller, A. H.–Hesli, V. L.–Maher, K. H. (1994): Political Values in Russia, Ukraine and Lithuania: sources and impliccations for democracy. *British Journal of Political Science* (24), 183–223.

Rényi A. (1954): *Valószínűségszámítás.* Budapest, Tankönyvkiadó.

Richardson, M. W. (1938): Multidimensional psychophysics. *Psychological Bulletin*, 35.

Richardson, M.–Kuder, G. F. (1933): Making a rating scale that measures. *Personnel J.*, (12), 36–40.

Ripley, B. D. (1981): *Spatial Statistics.* New York, Wiley.

Rivett, B. H. P. (1981): Policy selection by structural mapping. Proc. R. Soc. Lond. (354), 407–423.

Roberts, G.–Martyn, A. L.–Dobson, A. J.–McCarthy, W. H. (1981): Tumour thickness and histological type in malignant melanoma in New South Wales, Australia. 1970–1976. *Pathology* (13), 763–770.

Rokeach, M. (1973): *The Nature of Human Values.* New York, The Free Press.

Roskam, E. E. (1959): *Data Theory and Algoritms for Nonmetric Scaling* (I–II) (stencil) Psychology Laboratory, Mathematishe Psychologie, University of Nijmegen, The Netherlands.

Roskam, E. E. (1969): Data theorie en metrische analyse. NSZERK. *Tijdschrift Voor Psychologie* (25), 15–54, 66–82.

Roskam, E. E. (1872): An algorithm for multidimensional scaling by metric transformation of data. *Nijmegen: Program Buletin* (23/106), 659–660.

Roskam, E. E. (1975): *Nonmetric data analysis: General methodology and technique with brief descriptions of miniprograms.* Report No. 75-MA-13, Nijmegen, The Netherlands Psychology Laboratory, Mathematische Psychologie, University of Nijmegen.

Ross, J.–Cliff, N. (1964): A generalization of the interpoint distance mode. *Psychometrika* (29), 167–176.

Rost, J. (1985): A Latant Class Model for Rating Data. *Psychometrika* (50/1, March), 37–49.

Saaito, T. (1978): The problem of the additive constant and eigenvalues in metric multidimensional scaling. *Psychometrika* (43), (193–201.

Sammon, J. W. (1969): A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers* (18), 401–409.

Saris, W. E.–Stronkhorst, L. M. (1984): *Causal modelling in nonexperimental research: An introduction to the LISREL approach.* Amsterdam, Sociometric Research Foundation.

Schiffman, S. S.–Reynolds, M. L.–Young, F. W. (1981): *Introduction to Multidimensional Scaling: Theory, Methods and Applications.* New York, Academic Press.

Schlesinger, I. M.–Guttman, L. (1969): Smallest Space Analysis of Inteligence and Achievement Task. *Psyhological Bulletin* (71), 95–100.

Schluchter, W. (1989): *Rationalism, Religion and Domination.* Berkeley (Cal.), University of California Press.

Schmidt, C. F. (1972): Multidimensional scaling analysis of the printed media's explanations of the riots of the Summer of 1967. *J. Pers. And Soc. Psychil.* (24), 59–67.

Schobert, R. (1979): *Die Dynamisierung komplexer Marktmodelle mit Hilfe von Verfahren der mehrdimensionalen Skalierung*, Berlin, Duncker and Humbolt.

Schoenberg, I. J. (1935): Remarks to Maurice Fréchet's article "Sur la définition axiomatique d'une classe d'espaces vectoriels distanciés applicables vectoriellement sur l'espace de Hilbert". *Ann. Math.* (36), 724–732.

Schönemann, P. H. (1966): A generalized solution of the orthogonal procrustes problem. *Psychometrika* (21), 1–10.

Schönemann, P. H. (1970): On metric multidimensional unfolding. *Psychometrika* (35), 349–366.

Schönemann, P. H.–Carroll, R. M. (1970): Fitting one matrix to another under choice of a central dilation and a rigid motion. *Psychometrika* (35), 245–256.

Schönemann, P. H.–Wand, M. (1972): Some new results on factor indeterminancy, *Psychometrika* (37), 61–91.

Searla, S. R. (1971): *Linear Models.* John Wiley és Sons, 532.

Seber, G. A. F. (1984): *Multivariate Observations.* John Wiley és Sons, 686.

Seligson, M. A.–J. A. Booth (1976): Political participation in Latin America – Agenda for research. *Latin American Research* (11/3), 95–119.

Shepard, R. N. (1962/a): The analysis of proximities: multidimensional scaling with an unknown distance function (I.) *Psychometrika* (27), 125–140.

Shepard, R. N. (1962/b): The analysis of proximities: multidimensional scaling with an unknown distance function (II.) *Psychometrika* (27), 219–246.

Shepard, R. N. (1974): Representation of Structure in similatrities data: problems and prospects. *Psychometrika* (39), 373–421.

Shepard R. N.–Carroll, J. D. (1966): Parametric Representation of Nonlinear Data Structures. In: P. R. Krishnaiah (szerk.): *Multivariate Analysis.* New York, Academic Press, 561–592,

Sherman, C. R. (1972): Nonmetric multidimensional scaling: a Monte Carlo study of the basic parameters. *Psychometrika* (37), 323–355.

Sibson, R. (1978): Studies in the robustness of multidimensional scaling: Procrustes statistics. *J. R. Statist. Soc.* (40), 234–238.

Sibson, R. (1979): Studies in the robustness of multidimensional scaling; perturbational analysis of classical scaling. *J. R. Statist. Soc.* (41), 217–229.

Sibson, R.–Bowyer, A.–Osmond, C. (1981): Studies in the robustness of multidimensional scaling: Euclidean models and simulation studies. *J. Statist. Comput. Simul.* (13), 273–296.

Sikos T. T. (szerk.) (1984): *Matematikai és statisztikai módszerek alkalmazásának lehetőségei a területek utatásokban.* Budapest, Akadémiai Kiadó, 301.

Singer, B. (1989): Grade of membership representations concepts and problems. In: T. W. Anderson–K. B. Athreya (szerk.): *Festschrift für Samuel Karlin.* New York, Academic Press.

Singson, R. L. (1973): *A multidimensional scaling and unfolding analysis of store image and shoping behavior.* (Ph. D. Thesis.) Seatle (Wash.), University of Washington.

Smith, N. J.–Iles, K. (1988): A graphical depiction of multivariate similarity among sample plots. *Can. J. For. Res.* (18), 467–472.

Sneath, P. H. A.–Sokal, R. R. (1973): *Numerical Taxonomy.* San Francisco, W. H. Freeman.

Snijders, T. A. B.–Dormaar, M.–van Schuur, W. H.–Dijkman-Caes, C.–Driessen, G. (1990): Distribution of some similarity coefficients for dyadic binary data in the case of associated attributes. *Journal of Classification* (7), 5–31.

Sörbom, D. (1974): A General Method for Studying Differences in Factor Means and Factor Structure Between Groups. *British Journal of Mathematical and Statistical Psychology* (27), 229–239.

Spaeth, H. J.–Guthery, S. B. (1969): The use and utility of the monotone criterion in multidimensional scaling. *Multivariate Behavioral Research* (4), 501–515.

Spath, H. (1980): *Cluster Analysis Algorithms, Computers and Their Applications.* Ellis Horwood Limited, 221.

Spence, I. (1970/a): Local minimum solution in nonmetric multidimensional scaling. *Proc. of the Soc. Stats. Section of the American Statist. Assoc.* (13), 365–367.

Spence, I. (1970/b): *Multidimensional scaling: An empirical and theoretical investigation.* (Unpublised Ph. D. Thesis.), Toronto, University of Toronto.

Spence, I. (1972): A Monte Carlo evaluation of three nonmetric multidimensional scaling algorithms. *Psychometrika* (37), 461–486.

Spence, I.–Domoney, D. W. (1974): *Single subject incomplete designs for nonmetric multidimensional scaling. Takane, Y. 1978a: A maximum likelihood method for nonmetric multidimensional scaling (I.)* The case in which all empirical pairwise orderings are independent – theory. *Japanese Psychological Research* (20), 7–17.

Spence, I.–Graef, J. (1974): The determination of the underlying dimensionality of an empirically obtained matrix of proximities. *Multiv. Behav. Res.* (9), 331–342.

Spence, I.–Lewandowsky, S. (1989): Robust multidimensional scaling. *Psychometrika* (54), 501–513.

Spence, I.–Ogilvie, J. C. (1973): A table of expected stress values for random rankings in nonmetric multidimensional scaling. *Multivariate Behavioral Research* (8), 511–517.

Stenson, H. H.–Knoll, R. L. (1969): Goodness of fit for random rankings in Kruskal's nonmetric scaling procedure. *Psychological Bulletin* (71), 122–126.

Sváb J. (1979): *Többváltozós módszerek a biometriában.* Budapest, Mezőgazdasági Kiadó, 221.

Szakolczai, Á. (1987): On the charracteristics of the value choices of students. In: I. Hrubos (szerk.): *The Social Conditions of Engineers and Economists.* Budapest, Department of Sociology University of Economics.

Takács L.–Ziermann, M. (1954): *Valószínűségszámítás.* Budapest, Tankönyvkiadó.

Takane, Y. (1978/a): A maximum likelihood method for nonmetric multidimensional scaling: (I.) The case in which all empirical pairwise orderings are independent – theory. *Japanese Psychological Research* (20), 7–17.

Takane, Y. (1978b): A maximum likelihood method for nonmetric multidimensional scaling: (I.) The case in which all empirical pairwise orderings are independent – evaluation. *Japanese Psychological Research* (20), 105–114.

Takane, Y. (1981): Multidimensional successive categories scaling: a mimum likelihood method. *Psychometrika* (46), 9–28.

Takane, Y.–Young, F. W.–de Leeuw, J. (1977): Nonmetric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features. *Psychometrika* (42), 7–67.

Telegdi L. (1987–1988): Bináris változók struktúrájának vizsgálatáé. *Alkalmazott Matematikai Lapok* (13), 17–42.

Tenbruck, E. H. (1980): The problem of thematic unity in the works of Max Weber. *British Journal of Sociology* (31), 316–351.

Tenenhaus, M.–Young, F. W. (1985): An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika* (50), 91–119.

Ter Braak, C. J. F. (1992): Multidimensional scaling and regression. *Statistica Applicata* (4), 577–586.

Tijssen, R. J. W.–Van Raan, A. F. J. (1989): Mapping co-word structures: a comparison of multidimensional scaling and leximappe. *Scientometrics* (15), 283–295.

Tiku, M. L. (1986): *Tan and Balakrishnan, N.: Robust Inference Statistics: textbooks and monographs* (Vol. 71.) Marcel Dekker, 321.

Tobler, W.–Wineberg, S (1971): A Cappadocian Speculation. *Nature* (231), 39–41.

Tong, S. T. Y. (1989): On nonmetric multidimensional scaling ordination and interpretation of the matorral vegetation in lowland Murcia. *Vegetatio* (79), 65–74.

Torgerson, W. S. (1952): Multidimensional scaling (1.): Theory and method. *Psychometrika* (17), 401–419.

Torgerson, W. S. (1958): *Theory and Method of Scaling.* New York, Wiley.

Tribe, K. (1988): *Governing Economy.* Cambridge, Cambridge University Press.

Tucker, L. R. (1951): *A method for synthesis of factor analytic studies. Personnel Research Section Report No. 984.* Washington D. C., Department of the Army.

Tucker, L. R. (1960): Intra-individual and inter-individual multidimensionality. In: H. Gulliksen–S. Messick (szerk.): *Psychological Scaling: Theory and Applications*. New York, Wiley.

Tucker, L. R. (1963): Implications of factor analysis of three-way matrices for measurement of change. In: C. W. Harris (szerk.): *Problems in Measuring Change.* Madison (Wis.), University of Wisconsin Press.

Tucker, L. R. (1965): Experiments in multi-mode factor analysis. In: A. Anastasi (szerk.): *Testing Problems in Perspective.* Washington D. C.: *American Council on Education,* 1966/a. (Reprinted from proceeding of the 1964 Invitational Conference on Testing Problems.) Princeton (N. J.): Educational Testing Service.

Tucker, L. R. (1966/a): Experiments in multimode factor analysis. In: A. Anastassi (szerk.): *Testing Problems in Perspective* Washington D. C., American Council on Education.

Tucker, L. R. (1966/b): Some mathematical notes on three-mode factor analysis. *Psychometrika* (31), 279–311.

Tucker, L. R. (1967): Three-mode factor analysis of Parker-Feisman complex racking behavior data. *Multivariate Behavioral Research* (2), 139–51.

Tucker, L. R. (1972): Relations between multidimensional scaling and three-mode factor analysis. *Psychometrika* (37), 3–27.

Tucker, L. R.–Messick, S. (1963): An individual differences model for multidimensional scaling. *Psychometrika* (28), 333–367.

Tukey, J. W. (1977): *Exploratory Data Analysis.* Addison-Wesley, 499.

Turner, R. E. (1970): *Perceptuall imensions of salesmen: Multidimensional analysis of calling-allocating and sales-redponse behavior.* (Unpublished Ph. D. Thesis.) Northwestern University. 1970.

Upton, G. J. G. (1978): *Analysis of Cross-tabulated Data.* John Wiley and Sons, 148.

Van der Heijden, P. G. M.–de Falguerolles, A.–de Leeuw, J. (1985): A combined approach to contingency table analysis using correspondence analysis and log-linear analysis. *Appl. Statist.* (38), 249–292.

Van der Heijden, P. G. M.–de Leeuw, J. (1985): Correspondence analysis used complementary to loglinear analysis. *Psychometrika* (50), 429–447.

Van der Heijden, P. G. M.–Meijerink, F. (1989): Generalized correspondence analysis of multi-way contingency tables and multy-way (super-) indicator matrices. In: R. Copi–S. Bolasco (szerk.): *Multiway Data Analysis.* Amsterdam, North-Holland, 185–202.

Van der Heijden, P. G. M.–Worsley, K. J. (1988): Comment on "Correspondence analysis used complementary to loglinear analysis". *Psychometrika* (53), 287–291.

Van der Kloot, W. A.–Kroonenberg, P. M. (1985): External Analysis with Three-Mode Principal Component Models. *Psychometrika* (50/4, December) 479–494.

Van Schuur, W. H. (1977): Preference Mapping. An Introduction to Carroll & Chang's PREFMAP. ECPR publication subtantially reproduced in P. L. U., *Report* (39).

Velicer, W. F. (1977): The empirical comparison of the similarity of principal component, image and factor patterns. *Multivariate Behavioral Research* (12), 3–22.

Verhelst, N. D. (1981): A note on ALSCAL: the extimation of the additive constant. *Psychometrika* (46), 465–468.

Vince I. (1968): *Matematikai statisztika ipari alkalmazásokkal.* Budapest, Műszaki Könyvkiadó.

Voegelin, E. (é. n.): *Order and History.* 1956–1987. (5 Vols.) Baton Rouge (La.), Lousiana State University Press.

Von Neumann, J. (1941): Distribution of the ratio of the mean squere successive difference to the variance. *Am. Math. Stat.* (12), 367–395.

Wagenaar, W. A.–Padmos, P. (1971): Quantitative interpretation of stress in Kruskal's multidimensional scaling technique. *Brit. J. Math. Stat. Psychol.* (24), 101–110.

Ward, J. H. Jr. (1963): Hierarchical grouping to optimise an objective function. Journal of the American Statistical Association (58), 236–244.

Weber, M. (1949): *"Objectivity" in social science and social policy. In on the Metholdology of the Social Sciences.* Glencoe (Ill.), The Free Press.

Weber, M. (1978): *Economy and Society.* Berkeley (Cal.), University of California Press.

Weeks, D. G.–Bentler, P. M. (1982): Restricted multidimensional scaling models for asymmetric proximities. *Psychometrika* (47), 201–208.

Wiley, D. E. (1973): *The identification problem for structural equation models with unmeasured variables, in Sructural Equation Models in the Social Sciences.* (A. S. Goldberger–O. D. Ducan szerk.) New York, Seminar Press.

Wilson, K. R. (é. n.): *Prestige attribution: An application of nonmetric, multidimensional scaling.* (Unpublished Ph. D. Thesis.) Pardue University.

Winsberg, S.–Carroll, J. D. (1989/a): A quasi-nonmetric method for multidimensional scaling of multiway data via a restricted case of an extended Euclidean model. *Psychometrika* (54), 217–229.

Winsberg, S.–Carroll, J. D. (1989/b): A quasi-nonmetric method for multidimensional scaling of multiway data via a restricted case of an extended INDSCAL model. In: R. Coppi, R.–S. Bolasco (szerk.): *Multiway Data Analysis.* Amsterdam, North Holland.

Winsberg, S.–De Soete, G. (1993): A latent class approach to fitting the weighted Euclidean model, CLASCAL. *Psychometrika* (58), 315–330.

Wish, M.–Carroll, J. D. (1974): Applications of individual differences scaling to studies of human percetion and judgment. In: *Carteret and Friedman Handbook of Perception* (Vol. 2.): *Psychophysical Judgment and Measurement.* New York, Acadamic Press, 15.

Wish, M.–Carroll, J. D. (1982): Theory of mutidimensional scaling. In: P. R. Krishnaiah–L. N. Kanal (szerk.): *Handbook of Statistics.* (Vol, 2.) Amsterdam, North Holland, 317–345.

Wish, M.–Deutsch, M.–Biener, L. (1972): Differences in perceived similarity of nations. In: A. K. Romney–R. N. Shepard–S. B. Nerlove (szerk.): *Theory and Applications in the Behavioural Sciences.* (Vol. 2.) New York, Seminar Press, 289–313.

Wold, H. (1966): Estimation of principal components and related models by iterative least squares In: P. Krishnaiah (szerk.): *International Symposium on Multivariate Analysis Dayton Ohio (1965.* New York, Academic Press 391–420.

Woodbury, H. A.–Manton, K. G. (1989): Grade of Membership Analysis of Depression-related Psychiatric Disorders. *Sociological Methods and Research* (18/1,. August), Sage.

Wright, B. D.–Masters, G. N. (1982): *Rating scale analysis.* Chicago, Mesa Press.

Wrigley, N. (1985): *Categorical Data Analysis for Geographers and Environmental Scientists.* Longman, 392.

Yana, H.–Mukherjee, D. N. (é. n.): A Generealized Method of Image Analysis from an Intercorrelation Matrix which may be Singular. *Psychometrika* (52/4), 555–564.

Young, F. W. (1987): *Multidimensional scaling: History, Theory and Applications.* (R. M. Hamer, szerk.) Hillsdale (N. J.), Lawrence Erlbaum.

Young, F. W.–Cliff, N. F. (1972): Interactive scaling with individual subjects. *Psychometrika* (37), 385–415.

Young, G.–Householder, A. (1938): Discussion of a set of points in terms of their mutual distances. *Psychometrika* (3), 19–22.

Young, F. W.–de Leeuw, J.–Takane, Y. (1976): Regression with qualitative and quantitative variables: an alternating least squares method with optimal scaling feature. *Psychometrika* (41), 505–529.

Young, F. W.–De Leeuw, J.–Takane, Y. (1980): Quantifying qualitative data. In: E. D. Lantermann–H. Fegger (szerk.): *In: Similarity and Choice.* Bern, Hans Huber.

Young, F. W.–Null, C. H. (1978): Multidimensional scaling of nominal data: the recovery of metric information with ALSCAL. *Psychometrika* (43), 367–379.

Young, F. W.–Takane, Y.–Lewyckyj, R. (1978): Three notes on ALSCLAL. *Psychometrika* (43), 433–435.

Zegers, F. E. (1986): A family of chance-corrected association coefficients for metric scales. *Psychometrika* (51), 559–562.

Zegers, F. E.–ten Berge, J. M. F. (1985): A family of association coefficients for metric scales. *Psychometrika* (50), 17–24.

Zinnes, J. L.–Griggs, R. A. (1974): Probabilistic, multidimensional unfolding analysis. *Psychometrika* (39), 327–350.

Zinnes, J. L.–MacKay, D. B. (1983): Probabilistic multidimensional scaling: complete and incomplete data. *Psychometrika* (48), 27–48.