FOGALMAK ÉS MÓDSZEREK

VITA LÁSZLÓ

A faktoranalízis közgazdasági alkalmazásának lehetőségeiről

A faktoranalízis a többváltozós elemzéseknek mintegy 5—6 évtizedes múltra visszatekintő ága. Kidolgozása Charles Spearman és Karl Pearson nevéhez fűződik. Mivel a faktoranalízis számos modelljét különböző pszichológiai elméletekre építve dolgozták ki, a faktoranalízist sokáig — igen tévesen — speciális pszichológiai módszerként tartották számon. E nézet téves voltát éppen a módszer széleskörű, egyre több területre kiterjeszkedő alkalmazásai bizonyítják. A faktoranalízis tehát nem speciális pszichológiai, hanem igen széles körben alkalmazható *statisztikai* módszer.

Mivel a faktoranalízis leglényegesebb eleme a jelenségek közötti bonyolult összefüggések minél egyszerűbb formában történő leírása, különösen olyan tudományokban alkalmazható sikerrel, mint a közgazdaságtudomány. A társadalmi-gazdasági jelenségek igen bonyolultan, kölcsönösen összefüggő rendszere ugyanis szinte kimeríthetetlen tárháza a faktoranalízis alkalmazási lehetőségeinek. Annak ellenére, hogy közgazdasági alkalmazásai viszonylag újkeletűek, a faktoranalízis igen hasznos segédeszköze lehet a közgazdasági jelenségeket mélyebben megismerni akaró, azokat modellezni kívánó közgazdászoknak.

Mivel eddig — tudomásom szerint — nem jelent meg részletes magyar nyelvű ismertetés a faktoranalízis módszereiről, e cikk első részében az ezekkel kapcsolatos legfontosabb tudnivalókat foglalom össze, a második részben pedig az első részben ismertetett módszer legérdekesebb közgazdasági alkalmazási lehetőségeire térek ki vázlatosan.

Mielőtt azonban rátérnék magának a módszernek az egzakt matematikai tárgyalására, célszerűnek tartom egy olyan példa előrebocsátását, ami egyrészt képet ad az ezután ismertetendő módszer lényegéről, felveti annak teljes problematikáját, másrészt megkönnyíti az ezután következő matematikai modell megértését is.

Tekintsük feladatunknak bizonyos — mondjuk N — számú ország „gazdasági fejlettség" szerinti rangsorolását. E feladat megoldása során az okozza a legfőbb nehézséget, hogy a „gazdasági fejlettség" rendkívül bonyolult, összetett, közvetlenül nem mérhető jelenség. Bonyolultsága elsősorban abban jut kifejezésre, hogy bár számtalan olyan tényező adható meg, amely többékevésbé szoros kapcsolatban áll a „gazdasági fejlettséggel", s ugyanakkor mérhető is, de ezek egyike sem azonosítható teljes mértékben azzal. Ezért a kitűzött rangsorolási feladat megoldásakor vagy úgy járunk el, hogy egyetlen, általunk a „gazdasági fejlettség" szempontjából a legfontosabbnak ítélt mérhető tényező (a továbbiakban: változó) alapján végezzük el az N ország rangsorolását, vagy valamilyen „komplex mutató" alapján kíséreljük meg azt.

Míg az első esetben hallgatólagosan feltételezzük, hogy a „gazdasági fejlettség" teljes mértékben azonosítható a kiemelt változóval, és ezzel nyilvánvalóan lemondunk a feladat megoldásához rendelkezésre álló információ egy részéről, addig a második esetben egy olyan mutatószám meghatározását tartjuk célunknak, ami a rendelkezésre álló információ minél nagyobb hányadát használja fel a rangsorolási feladat megoldásához. A gyakorlatban mindkét fajta megoldással találkozunk.

Az N ország „gazdasági fejlettség" szerinti rangsorolása a faktoranalízis segítségével a másodiknak említett módon végezhető el. Gyűjtsük össze mindazokat az $X_1, X_2, \ldots, X_n$ mérhető változókat, amelyekről feltételezhető, hogy sztochasztikus kapcsolatban állnak az általunk mérhetővé tenni kívánt „gazdasági fejlettséggel".[1] Ezek az eddigi tapasztalatok szerint különböző ellátottsági és demográfiai mutatók ([2] és [7]). Ha megvizsgáljuk az így összegyűjtött változók különböző országokra vonatkozó értékeit, akkor általában azt tapasztaljuk, hogy az egyes változók értékei nem egymástól függetlenül alakulnak. Ha tehát ismerjük pl. az első változó különböző országokra vonatkozó

$$X_{11}, X_{12}, \ldots, X_{1N}$$

értékeit, akkor ennek ismeretében néhány más, esetleg akár az összes többi változó országonkénti alakulására következtethetünk. E következtetéseink természetesen nem lesznek egyértelműek, hanem csak sztochasztikus jellegűek lehetnek.

Az összegyűjtött változók egymástól való függése azonnal magyarázatot nyer, ha arra gondolunk, hogy a vizsgálatba vont változók maguk is valamilyen változók függvényei lehetnek. Az összegyűjtött $X_1, X_2, \ldots, X_n$ változók egymástól való függősége tehát azzal magyarázható, hogy e változók mindegyike, vagy egy része egy vagy több, számunkra még egyelőre ismeretlen közös tényezőtől függ, amelyeket *közös faktoroknak* nevezünk. A közös faktorok tehát olyan *hipotetikus változók*, amelyek *csak közvetett módon*, a vizsgálatba vont változókra vonatkozó megfigyelések elemzése útján számszerűsíthetők, s jelenlétükre csak *a vizsgált változók egymástól való függéséből* következtethetünk.

Ezzel el is érkeztünk a faktoranalízis kiinduló hipotéziséhez, mely szerint a vizsgálatba vont változók maguk is további változók, az ún. közös faktorok lineáris függvényei, azaz

$$\begin{aligned}
\hat{X}_1 &= a_{11}K_1 + a_{12}K_2 + \ldots + a_{1m}K_m \\
\hat{X}_2 &= a_{21}K_1 + a_{22}K_2 + \ldots + a_{2m}K_m \\
&\cdots\cdots\cdots\cdots \\
\hat{X}_n &= a_{n1}K_1 + a_{n2}K_2 + \ldots + a_{nm}K_m
\end{aligned} \tag{1}$$

Az eredeti változóknak itt természetesen nem a pontos, hanem csak a becslésszerű előállításáról van szó, amire az $\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_n$ jelölések is felhívják a

[1] Itt elvonatkoztatunk a változók összegyűjtésénél fellépő problémáktól.

figyelmet.[2] A változók e felírásában szereplő a_{jp} együtthatók egyik meghatározási módszerét az 1.4 alpontban ismertetem.

Ha meghatározzuk a fenti sémában szereplő a_{jp} együtthatókat, akkor azt tapasztaljuk, hogy találhatók olyan közös faktorok, amelyekre minden változó előállításához szükség van (azaz az adott közös faktor minden egyes változóhoz tartozó együtthatója nullától különböző), de találhatók olyanok is, amelyek egynél több, de nem az összes változó előállításához szükségesek. Az előbbieket *általános*, az utóbbiakat *csoportfaktoroknak* nevezzük.

Eredeti feladatunk megoldása szempontjából nyilván az a kérdés, hogy léteznek-e a „gazdasági fejlettség"-gel kapcsolatban álló $X_1, X_2, \ldots, X_n$ változóknak általános faktorai és ha igen, akkor mennyi ezeknek a száma. Nyilvánvaló ugyanis, hogy abban az esetben, ha csak egy általános faktor létezik, akkor ez a „gazdasági fejlettség" egy komplex mérőszámának tekinthető, s az egyes országokra vonatkozó értékei alapján elvégezhető az országok rangsorolása. Ennek az a magyarázata, hogy eleve olyan változókat vontunk be a vizsgálatba, melyekről feltételezhető, hogy valamilyen kapcsolatban állnak a „gazdasági fejlettséggel", s így az az $X_1, X_2, \ldots, X_n$ változók egy általános faktorának tekinthető.

Az a kérdés, hogy a „gazdasági fejlettség" az $X_1, X_2, \ldots, X_n$ változók *egyetlen* általános faktora-e, már jóval bonyolultabb az előbbinél, s vagy az összegyűjtött változók logikai vizsgálata, vagy az (1) sémában szereplő együtthatók vizsgálata alapján választható meg.[3] Mivel az eddigi kutatások nagymértékben valószínűsítik ezt, a továbbiakban feltételezzük, hogy az $X_1, X_2, \ldots, X_n$ változók egyetlen általános faktora a „gazdasági fejlettség". Tegyük fel, hogy ez az (1) sémában a K_1-gyel jelölt faktor. Ezen kívül természetesen lehet az összegyűjtött változóknak egy vagy akár több csoportfaktora[4] is, ezeknek azonban feladatunk megoldása szempontjából nem tulajdonítunk jelentőséget.

Mivel az eredeti feladatunk megoldásához végső soron a K_1 általános faktor előállítására van szükség, az eddigi gondolatmenetet megfordítva azt állítjuk, hogy ha minden egyes változó előállítható a közös faktorok — s köztük a K_1 általános faktor — felhasználásával, akkor az egyes közös faktorok is előállíthatók kell hogy legyenek a ténylegesen megfigyelt változók segítségével. Legyen ez az előállítás a K_1 általános faktor esetében a következő:

$$\hat{K}_1 = \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_n X_n,$$

melynek részleteire a 2.1 alpontban térünk ki.

Tekintettel a $\hat{K}_1$ előállításában szereplő változók mérhetőségére, egy eredetileg nem mérhető változót — a „gazdasági fejlettséget" — egy mérhető változóval közelítettünk, amelynek értéke minden egyes országra nézve meghatározható, s ezek alapján elvégezhető az országok „gazdasági fejlettség" szerinti rangsorolása.

[2] A modell pontosabb megfogalmazására az 1.1. alpontban térek ki.

[3] Maga az (1) séma megoldása ugyanis nem tételezi fel feltétlenül a közös faktorok előzetes ismeretét. Itt csak a megértés megkönnyítése érdekében indultunk ki a változók közötti kapcsolat logikai elemzéséből.

[4] Például: hasonló földrajzi adottságok, hasonló demográfiai helyzet stb.

Az előbb vázolt eljárással szemben természetesen felmerülhet egy olyan — egyébként teljesen jogos — ellenvetés, hogy hogyan tulajdoníthatunk közgazdasági tartalmat egy olyan mesterséges változónak, amely esetleg a legkülönfélébb jellegű változók lineáris kombinációjaként áll elő. E kérdéssel a 2.1 alpontban foglalkozom részletesebben.

1. A módszer rövid ismertetése

1.1 *A faktoranalízis modellje és alapfogalmai*

Tegyük fel, hogy egy N elemből álló statisztikai sokaságot egyidejűleg n számú valószínűségi változó[5] (mennyiségi ismérv) szerint vizsgálunk. Ha az egyes változókat X_j-vel ($j = 1, 2, \ldots, n$) jelöljük, akkor az előbbi megfogalmazás azt jelenti, hogy a vizsgált statisztikai sokaság minden egyes egységére vonatkozóan feljegyezzük az X_j változók X_{ji} ($i = 1, 2, \ldots, N$) értékeit, és ezek felhasználásával végezzük el a sokaság elemzését.

A további tárgyalást és formulákat nagymértékben leegyszerűsíthetjük azzal, hogy az eredetileg megfigyelt X_{ji} értékek helyett a

$$z_{ji} = \frac{X_{ji} - \bar{X}_j}{s_j} \qquad \begin{pmatrix} j = 1, 2, \ldots, n \\ i = 1, 2, \ldots, N \end{pmatrix} \tag{1.1.1}$$

ún. *standardizált értékekkel* dolgozunk, ahol

$$\bar{X}_j = \frac{1}{N} \sum_{i=1}^{N} X_{ji} \qquad (j = 1, 2, \ldots, n)$$

az X_j változó átlaga,

$$s_j = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (X_{ji} - \bar{X}_j)^2} \qquad (j = 1, 2, \ldots, n)$$

pedig az X_j változó szórása. A z_{ji} értékeket felvevő változókat z_j-vel jelöljük, és *standardizált változóknak* nevezzük. E standardizált változók átlaga nulla, szórása pedig egy.

A faktoranalízis abból a hipotézisből indul ki, hogy minden egyes standardizált változó további hipotetikus változók, az ún. *faktorok* lineáris függvényeként írható fel. Ez a hipotézis matematikailag a

$$\mathbf{z} = \mathbf{A}\mathbf{f} = \mathbf{A}_k \mathbf{k} + \mathbf{A}_u \mathbf{u} \tag{1.1.2}$$

lineáris modell segítségével írható fel ahol

$\mathbf{z} = [z_1, z_2, \ldots, z_n]^*$ — a standardizált változók oszlopvektora,

$\mathbf{k} = [K_1, K_2, \ldots, K_m]^*$ — az ún. *közös faktorok* oszlopvektora,

$\mathbf{u} = [U_1, U_2, \ldots, U_n]^*$ — az ún. *egyedi faktorok* oszlopvektora,

[5] E valószínűségi változók együttes eloszlására vonatkozóan nem teszünk semmiféle megkötést. A további tárgyalás során az egyszerűség kedvéért mindig e valószínűségeloszlás empirikus jellemzőit használjuk a megfelelő elméleti jellemzők helyett.

$\mathbf{A}_k = [a_{jp}]$ $(j = 1, 2, \ldots, n;\ p = 1, 2, \ldots, m)$ — a közös faktorokra vonatkozó együtthatók — az ún. *közös faktorsúlyok* — $n \times m$ típusú matrixa,

$\mathbf{A}_u = \langle a_1, a_2, \ldots, a_n \rangle$ — az egyedi faktorokra vonatkozó együtthatók diagonális matrixa,

$$\mathbf{A} = [\mathbf{A}_k, \mathbf{A}_u]\ ;\ \mathbf{f} = [\mathbf{k}, \mathbf{u}]^*,$$

m pedig a közös faktorok száma.

Ezt a lineáris modellt *faktorsémának* is szokás nevezni, a benne szereplő **A** matrix pedig az ún. *sémamatrix*.

Az (1.1.2) modellt egy z_j változóra részletesen felírva a

$$z_j = a_{j1} K_1 + a_{j2} K_2 + \ldots + a_{jm} K_m + a_j U_j \qquad (j = 1, 2, \ldots, n) \tag{1.1.3}$$

alakú speciális regressziós egyenletekhez jutunk. E regressziós egyenletek specialitása egyrészt abban áll, hogy a bennük szereplő független változók olyan *közvetlenül nem mérhető standardizált változók*, melyekről csak közvetve — az általunk megfigyelt z_j változókra vonatkozó z_{ji} megfigyeléseken keresztül — nyerhetünk információt, s így az (1.1.3)-ban szereplő ismeretlen a_{jp} és a_j faktorsúlyok nem határozhatók meg a regressziós elemzés szokásos módszereivel. Másrészt itt olyan regressziós egyenletekről van szó, amelyek az ún. maradéktagot (hibatagot) *önálló változó*, az ún. egyedi faktor formájában tartalmazzák, tehát melyekre nézve a többszörös korrelációs együttható értéke egy.

A faktoranalízis modellje még egy másik szempontból is eltér az ún. regressziós modellektől. Míg ugyanis a regressziós modellek egy-egy *realizációja* az (1.1.2)-höz igen hasonló

$$\mathbf{y} = \mathbf{X}\,\mathbf{b} + \mathbf{e}$$

formában írható fel, addig a faktoranalízis (1.1.2) modelljének egy *realizációja*

$$\mathbf{Z} = \mathbf{A}\,\mathbf{F} = \mathbf{A}_k\,\mathbf{K} + \mathbf{A}_u\,\mathbf{U} \tag{1.1.4}$$

alakú, ahol

$$\mathbf{Z} = [z_{ji}] = \begin{bmatrix} z_{11}\, z_{12} \ldots z_{1N} \\ z_{21}\, z_{22} \ldots z_{2N} \\ \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot \\ z_{n1}\, z_{n2} \ldots z_{nN} \end{bmatrix}$$

az n számú standardizált változónak a vizsgált sokaság egyes egységeinél megfigyelt z_{ji} értékeit tartalmazó $n \times N$ méretű matrix, amit a továbbiakban a *megfigyelések matrixának* nevezünk,

$$\mathbf{K} = [K_{pi}] \quad \text{és} \quad \mathbf{U} = [U_{ji}]$$

pedig a közös, illetve egyedi faktoroknak a sokaság egységeire vonatkozó

K_{pi}, illetve U_{ji} értékeiből álló $m \times N$, illetve $n \times N$ méretű matrixok[6] és $\mathbf{F} = [\mathbf{K}, \mathbf{U}]^*$. Az (1.1.2) modell egy (1.1.4) realizációjának z_{ji} eleme részletesen kiírva tehát a következő:

$$z_{ji} = a_{j1} K_{1i} + a_{j2} K_{2i} + \ldots + a_{jm} K_{mi} + a_j U_{ji} \qquad \begin{matrix} (j = 1, 2, \ldots, n; \\ i = 1, 2, \ldots, N) \end{matrix}$$

A faktoranalízis modelljének (1.1.2) felírásából látható, hogy az abban szereplő faktorok két nagy csoportba sorolhatók. A modellben egyrészt szerepelnek olyan faktorok, amelyek egynél több változó leírásához szükségesek, másrészt olyanok is, amelyekre csak egy változó leírásához van szükség. Az előbbieket *közös faktoroknak* (K_p), az utóbbiakat *egyedi faktoroknak* (U_j) nevezzük. Az egy változó leírásához szükséges közös faktorok számát az adott változó *komplexitásának* nevezzük, ami az adott változó bonyolultságának kifejezője. A modellben szereplő közös faktorok számáról (m) feltesszük, hogy az jóval kisebb a megfigyelt változók számánál (n). Részletesebb vizsgálatára a következő alpontban térek ki.

A közös faktorok újabb két csoportba sorolhatók. Az első csoportba azok a közös faktorok tartoznak, melyekre minden egyes változó lineáris előállításához szükség van. Ezeket *általános faktoroknak* nevezzük. A másik csoportba az egynél, több, de nem az összes változó előállításához szükséges, *csoportfaktoroknak* nevezett közös faktorok kerülnek.

A faktoranalízis modelljének ismeretében már megfogalmazhatjuk a faktoranalízis feladatát, célját. A faktoranalízis feladata kettős: az egyik a közös faktorokra vonatkozó a_{jp} közös faktorsúlyok becslése, a másik pedig maguknak a faktoroknak az előállítása. Ez utóbbi feladat természetesen csak az $\mathbf{A}_k$ matrix ismeretében oldható meg.

Mielőtt áttekintenénk a faktoranalízis két alapfeladatának megoldására szolgáló módszereket, meg kell ismerkednünk a faktoranalízis alapfogalmaival és az (1.1.2) modell legfontosabb tulajdonságaival is.

Könnyen belátható ([6], 13. old.), hogy a j-edik standardizált változó σ_j^2-tel jelölt szórásnégyzete az (1.1.2) modell alapján a következőképpen írható fel:

$$\sigma^2 = \mathbf{a}_j^* \, \mathbf{\Phi} \, \mathbf{a}_j$$

ahol $\mathbf{a}_j^*$ az $\mathbf{A}$ teljes sémamatrix j-edik sora, $\mathbf{a}_j$ pedig ugyanez oszlopvektor formában felírva, és

$$\mathbf{\Phi} = \begin{bmatrix} \mathbf{\Phi}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{E}_n \end{bmatrix} \tag{1.1.6}$$

ahol $\mathbf{\Phi}_k = [r_{K_p K_q}]$ a közös faktorok m-edrendű korrelációs matrixa.[7]

A $\mathbf{\Phi}$ matrix azért írható fel az (1.1.6) módon, mert az U_j egyedi faktorokról minden esetben feltesszük, hogy egymással is és a közös faktorokkal is páronként korrelálatlanok, azaz $r_{K_p U_j} = 0$ és $r_{U_i U_j} = 0$, ha $i \neq j$. A közös faktorok ezzel szemben lehetnek egymással korreláltak és páronként korrelálatlanok is

[6] A **K** és **U** matrixok elemei természetesen nem figyelhetők meg közvetlenül.
[7] $\mathbf{E}_n$ itt az n-edrendű egységmatrixot jelöli.

Ha a közös faktorok páronként korrelálatlanok, akkor az (1.1.5) kifejezés a következőképpen egyszerűsödik:

$$\sigma_j^2 = \mathbf{a}_j^* \mathbf{a}_j = \sum_{p=1}^{m} a_{jp}^2 + a_j^2 = h_j^2 + a_j^2 \qquad (j = 1, 2, \ldots, n), \tag{1.1.7}$$

mely eredmény a következőképpen értelmezhető. Minden változó szórásnégyzete két részre bontható: az egyik rész az adott változó szórásnégyzetének a közös faktorok által *együttesen megmagyarázható része*, amit az adott változó *kommunalitásának* (h_j^2) nevezünk, a másik rész pedig az adott változó szórásnégyzetének az egyedi faktor által megmagyarázható része (a_j^2), amit az adott változó *egyediségének* szokás nevezni. Csak utalni kívánok rá, hogy ez utóbbi részt — általában becslésszerűen — további két részre: a modell adott megválasztásából, specifikációjából származó *specifikációs hibára*, és a mérési pontatlanságokból adódó *maradék* hibára szokás felbontani. Mivel minden z_j standardizált változó σ_j^2 szórásnégyzetének értéke 1, (1.1.7) felírható a

$$h_j^2 + a_j^2 = 1$$

módon is. Ez a tény a magyarázata annak, hogy a megoldás során elegendő csak a közös faktorsúlyokat meghatározni.

A faktoranalízisben fontos szerepet játszik a

$$V_p = \sum_{j=1}^{n} a_{jp}^2 \qquad (p = 1, 2, \ldots, m) \tag{1.1.8}$$

mennyiség is, ami azt mutatja, hogy a p-edik közös faktor milyen mértékben járul hozzá az összes vizsgált változó szórásnégyzetéhez.

Bizonyos esetekben — elsősorban a faktorok elnevezésének megválasztásakor, ami a kapott megoldás értelmezésének fontos mozzanata — szükség van az egyes változók és a faktorok közötti korrelációs együtthatók, azaz az

$$s_{jp} = r_{z_j K_p} \qquad (j = 1, 2, \ldots n;\ p = 1, 2, \ldots m)$$

és az $r_{z_j U_j}$ értékek ismeretére. Ezeket az értékeket *struktúra-együtthatóknak*, a belőlük felépülő

$$\mathbf{S} = [\mathbf{S}_k, \mathbf{S}_u] = \begin{bmatrix} s_{11} s_{12} \ldots s_{1m} & a_1\ 0 \ldots 0 \\ s_{21} s_{22} \ldots s_{2m} & 0\ a_2 \ldots 0 \\ \cdot\ \cdot\ \cdot & \cdot\ \cdot\ \cdot \\ \cdot\ \cdot\ \cdot & \cdot\ \cdot\ \cdot \\ \cdot\ \cdot\ \cdot & \cdot\ \cdot\ \cdot \\ s_{n1} s_{n2} \ldots s_{nm} & 0\ 0 \ldots a_n \end{bmatrix}$$

matrixot pedig sruktúramatrixnak nevezzük, amely az **A** sémamatrixhoz hasonlóan két részből: a közös faktorokra vonatkozó struktúraegyütthatókat tartalmazó $\mathbf{S}_k$ blokkból, és az egyedi faktorokra vonatkozó együtthatókat tartalmazó $\mathbf{S}_u = \mathbf{A}_u$ blokkból tevődik össze.

Bebizonyítható ([6], 32—34. old.), hogy az **S** struktúramatrix és az **A** sémamatrix között az

$$\mathbf{S} = \mathbf{A}\,\mathbf{\Phi}, \text{ illetve } \mathbf{S}_k = \mathbf{A}_k\,\mathbf{\Phi}_k \tag{1.1.9}$$

összefüggés áll fenn, ahol Φ az (1.1.6) korrelációs matrix. Ebből az is látható, hogy páronként korrelálatlan közös faktorok esetén

$$\mathbf{S} = \mathbf{A},$$

mivel ebben az esetben $\Phi = \mathbf{E}_{n+m}$. Ebből az eredményből látható, hogy abban az esetben, ha a megoldás páronként korrelálatlan közös faktorokból áll, elegendő csak az **A** sémamatrixot meghatározni. Ezzel szemben, ha a modellben korrelált közös faktorokat is megengedünk, akkor a megoldásnak mind a séma-, mind a struktúramatrixot tartalmaznia kell. Röviden összefoglalva: a séma a változóknak a faktorokból való összetevődését mutatja, a struktúra pedig a változók és a faktorok közötti korrelációs együtthatókat tartalmazza.

Az (1.1.2) modell, illetve annak (1.1.4) realizációja alapján nemcsak az egyes változók szórásnégyzetének felbontására nyílik lehetőség, hanem lehetővé válik az egyes változók közötti ún. *reprodukált korrelációs együtthatók* meghatározása is, melyek alapján megvizsgálható a faktoranalízis modelljének valósághűsége. Mielőtt megadnánk a reprodukált korrelációs együtthatók definícióját, néhány új fogalmat kell bevezetnünk.

Ismeretes, hogy a megfigyelt változók közötti korrelációs együtthatókból felépülő korrelációs matrix (**R**), a vizsgált változókra vonatozó megfigyelések felhasználásával az

$$\mathbf{R} = \frac{1}{N}\mathbf{Z}\mathbf{Z}^* \tag{1.1.10}$$

módon állítható elő.[8] Ezt a továbbiakban *teljes korrelációs matrixnak* nevezzük. Ha a teljes korrelációs matrix diagonális elemeiből rendre levonjuk az egyes változók egyediségét, akkor az

$$\mathbf{R}_h = \mathbf{R} - \mathbf{A}_u^2 \tag{1.1.11}$$

redukált korrelációs matrixhoz jutunk. A redukált korrelációs matrix (1.1.11) definíciójából jól látható, hogy annak diagonális elemei éppen az egyes változók kommunalitásai.

Tegyük fel azután, hogy a vizsgált változók halmazára nézve helyes az (1.1.2) módon megfogalmazott hipotézis. Ez egyben azt is jelenti, hogy

$$\mathbf{Z} = \mathbf{A}_k\mathbf{K} + \mathbf{A}_u\,\mathbf{U} \tag{1.1.12}$$

áll fenn, Ez az összefüggés természetesen csak az $\mathbf{A}_k$ matrix ismeretében írható fel. Helyettesítsük ezután (1.1.12)-t (1.1.10)-be Így a következő eredményre jutunk:

$$\mathbf{R}_r = \frac{1}{N}(\mathbf{A}_k\,\mathbf{K} + \mathbf{A}_u\,\mathbf{U})(\mathbf{A}_k\,\mathbf{K} + \mathbf{A}_u\,\mathbf{U})^* = \tag{1.1.13}$$

$$= \mathbf{A}_k\frac{\mathbf{K}\mathbf{K}^*}{N}\mathbf{A}_k^* + 2\,\mathbf{A}_k\frac{\mathbf{K}\mathbf{U}^*}{N}\mathbf{A}_u + \mathbf{A}_u\frac{\mathbf{U}\mathbf{U}^*}{N}\mathbf{A}_u$$

[8] Ez a tény egyszerűen a korrelációs együtthatók definíciájából és a vizsgált változók standardizáltságából következik.

Vegyük most figyelembe, hogy $\frac{\mathbf{KK}^*}{N}$, illetve $\frac{\mathbf{UU}^*}{N}$ nem más, mint a közös, illetve egyedi faktorok korrelációs matrixa, $\frac{\mathbf{KU}^*}{N}$ pedig a közös és egyedi faktorok közötti korrelációs együtthatókból álló matrix. Mivel azonban az egyedi és közös faktorokról, illetve az egyedi faktorokról is feltettük a páronkénti korrelálatlanságot (1.1.13) az

$$\mathbf{R}_r = \mathbf{A}_k \frac{\mathbf{KK}^*}{N} \mathbf{A}_k^* + \mathbf{A}_u^2 = \mathbf{A}_k \boldsymbol{\Phi}_k \mathbf{A}_k^* + \mathbf{A}_u^2 = \mathbf{A} \boldsymbol{\Phi} \mathbf{A}^* \tag{1.1.14}$$

alakba megy át, ahol $\boldsymbol{\Phi}$ az (1.1.6) korrelációs matrix. Az $\mathbf{R}_r$-rel jelölt matrix az ún. *reprodukált korrelációs matrix*. Elnevezését az indokolja, hogy az (1.1.14) és az (1.1.13) egyenlőségek teljesülésének egyaránt az (1.1.2) hipotézis helyes sége a feltétele, itt tehát a teljes korrelációs matrix (1.1.2) modellen keresztüli visszaszámolásáról, reprodukálásáról van szó.

Ezek után kézenfekvő az (1.1.2) modell valósághűségének ellenőrzése is. Képezzük ugyanis az

$$\overline{\mathbf{R}} = \mathbf{R} - \mathbf{R}_r \tag{1.1.15}$$

reziduális matrixnak nevezett különbséget. Mivel az $\mathbf{R} = \mathbf{R}_r$ egyenlőség teljesülésének feltétele az (1.1.2) s következésképpen az (1.1.12) pontos teljesülése, a reziduális matrix elemeinek nagysága lehetővé teszi a faktoranalízis kiinduló hipozézise valósághűségének vizsgálatát.

A gyakorlatban általában — később ismertetendő okokból — megelégszünk a redukált korrelációs matrix ($\mathbf{R}_h$) reprodukálásával is. Ilyenkor az (1.1.14)-nek csak a közös faktorokra vonatkozó

$$\mathbf{R}_{r|h} = \mathbf{A}_k \boldsymbol{\Phi}_k \mathbf{A}_k^* \tag{1.1.16}$$

részét tekintjük, ami (1.1.9) felhasználásával az

$$\mathbf{R}_{r|h} = \mathbf{S}_k \mathbf{A}_k^* = \mathbf{A}_k \mathbf{S}_k^*$$

módon is felírható. Ha a közös faktorok páronként korrelálatlanok, akkor (1.1.16) az

$$\mathbf{R}_{r|h} = \mathbf{A}_k \mathbf{A}_k^* \tag{1.1.17}$$

alakba megy át, amit a faktoranalízis *Thurstone-féle alaptételének* szokás nevezni.

Az egyedi faktorok figyelmen kívül hagyása esetén a reziduális korrelációs matrix természetesen a redukált korrelációs matrix és $\mathbf{R}_{r|h}$ különbségeként adódik.

Sajnos annak megítélésére, hogy a reziduális korrelációs matrix elemei szignifikánsan különböznek-e a nullától, nem állnak rendelkezésre minden megoldási módszer esetén alkalmazható statisztikai próbák. Ehelyett különböző, gyakorlati tapasztalatokon alapuló kritériumokat szoktak adni ennek eldöntésére. Egy ilyen kritérium például a következő

$$\sigma_{\overline{rr}} \leq \frac{1}{\sqrt{N}}, \tag{1.1.18}$$

ahol $\sigma_{\bar{r}\bar{r}}$ a reziduális korrelációs matrix $\bar{r}_{jk}$ $(j, k = 1, 2, \ldots n)$ elemeiből számított szórás. Az (1.1.18) teljesülése esetén elfogadjuk az (1.1.2) formában kifejezett hipotézist, ellenkező esetben pedig elvetjük azt, és új modell szám-szerűsítésével kísérletezünk.

1.2 *A kommunalitásokról*

Ebben az alpontban egyrészt a kommunalitások és a közös faktorok számának kapcsolatával, másrészt pedig a kommunalitások becslésével foglalkozunk.

A *faktoranalízis* feladata úgy is megfogalmazható, hogy minél kevesebb számú közös faktor segítségével írjunk le egy adott változóhalmazt. Ennek érdekében minden egyes eredetileg megfigyelt változót két egymással korrelálatlan változó összegére bontjuk fel a

$$z_j = z'_j + z''_j \qquad (j = 1, 2, \ldots, n) \tag{1.2.1}$$

módon ahol

$$z'_j = a_{j1}K_1 + a_{j2}K_2 + \ldots + a_{jm}K_m \text{ és } z''_j = a_jU_j \quad (j = 1, 2, \ldots, n)$$

és eltekintünk a z_j változók z''_j komponensétől. A z_j változók z'_j komponenseit a továbbiakban *redukált változóknak*, az azokra vonatkozó megfigyeléseket tartalmazó

$$\mathbf{Z}' = [z'_{ji}] \qquad (j = 1, 2, \ldots, n, \quad i = 1, 2, \ldots, N)$$

matrixot pedig — melynek elemei természetesen nem figyelhetők meg közvetlenül — a *megfigyelések redukált matrixának* nevezzük. Könnyen belátható, hogy a redukált változók korrelációs matrixa az (1.1.11) redukált korrelációs matrix, ami felírható az

$$\mathbf{R}_h = \frac{1}{N}\mathbf{Z}'\mathbf{Z}'^* \tag{1.2.2}$$

módon is. Az (1.2.2) alapján azt is mondhatjuk, hogy a minél egyszerűbb faktoranalitikus megoldás előállítása érdekében lemondunk az eredetileg megfigyelt z_j változók szórásnégyzetének, vagy más kifejezéssel élve a z_j változó által tartalmazott információmennyiség egy részéről.[9] A z_j változók szórásnégyzetének e figyelmen kívül hagyott része az a_j^2 egyediség.

Ez a magyarázata annak, hogy a kommunalitások értéke és a közös faktorok száma között szoros kapcsolat van, aminek pontosabb megfogalmazására csak az (1.1.2) modell geometriai interpretációjának megadása után térhetünk ki.

A megfigyelések $\mathbf{Z}$ matrixának

$$z_i^* = [z_{j1}, z_{j2}, \ldots, z_{jN}) \qquad (j = 1, 2, \ldots, n)$$

sorai az N-dimenziós euklideszi tér (jelölése R^N) n pontjának tekinthetők. Ezt az N-dimenziós teret, melynek koordinátatengelyei a vizsgált statisztikai sokaság (minta) egy-egy elemét reprezentálják, *mintatérnek* szokás nevezni.

[9] Amennyiben a z_j változó által tartalmazott információmennyiséget annak σ^2 szórásnégyzetével mérjük.

Az is nyilvánvaló, hogy e pontok egy a mintatérbe beágyazott legfeljebb n-dimenziós altér, az ún. *változótér* pontjainak is tekinthetők. Ez a szemlélet annak felel meg, hogy most a megfigyelések matrixának oszlopait tekintjük, a változótér dimenziójára vonatkozó megállapítás pedig egyszerűen abból következik, hogy a **Z** matrix rangja legfeljebb n, hiszen általában $n \ll N$ teljesül. A változótér koordinátatengelyei az egyes z_j változóknak felelnek meg.

Ha a z_j változók *lineárisan függetlenek*, azaz a **Z** matrix sorai lineárisan függetlenek, akkor a változótér n dimenziós, minden más esetben n-nél kisebb dimenziójú. A változók ilyen értelemben vett lineáris függetlensége azonban nem zárja még ki azok páronkénti korreláltságát. A változók páronkénti korreláltságából viszont az következik, hogy legalább egy változó *közelítőleg* előállítható az összes többi lineáris függvényeként. Másképpen fogalmazva a változók páronkénti korreláltsága esetén legalább az egyiküknek az összes többire vonatkozó többszörös korrelációs együtthatója elég nagy. Geometriailag ez annyit jelent, hogy a változótér pontjainak helyzetét kisebb-nagyobb mértékben megváltoztatva elérhetjük azt, hogy mind az N megváltoztatott helyzetű pont egy n-nél kisebb dimenziósszámú altérben helyezkedjen el. A faktoranalízis feladata tehát geometriailag fogalmazva az, hogy megkeresse azt a legkisebb dimenziószámú ún. közös *faktorteret*, amely még a változótér minden módosított helyzetű pontját tartalmazza. E módosított helyzetű pontoknak a változótér koordinátatengelyeire vonatkozó koordinátáit a megfigyelések redukált matrixának oszlopai, az egyes változóknak a közös faktortér közös faktorokat reprezentáló koordinátatengelyeire vonatkozó koordinátái pedig az $\mathbf{A}_k$ sémamatrix soraiból olvashatók ki. Míg a változótér dimenziószámát a **Z** matrix rangja, addig a közös faktortér dimenziószámát, melyet a továbbiakban m-mel jelölünk, a **Z'** matrix rangja határozza meg.

Eddigi fejtegetéseinket a megfigyelések **Z** matrixára alapoztuk. Mivel azonban a faktoranalízis alapadatait a legtöbbször a vizsgált változók között megfigyelt korrelációs együtthatók képezik, ezután következő állításainkat célszerűbb a teljes, illetve redukált korrelációs matrixra alapozni. Ennek lehetőségét az adja meg, hogy mind a teljes, mind a redukált korrelációs matrix ún. Gram-féle matrix, melynek rangja megegyezik a **Z**, illetve **Z'** matrix rangjával. ([10], 124. old.)

Mielőtt megfogalmaznánk a közös faktortér dimenziószámára vonatkozó tételt, megjegyezzük, hogy a *faktoranalízis* és a *komponenselemzés*[10] egységes tárgyalhatósága érdekében a továbbiakban a teljes korrelációs matrixot speciális redukált korrelációs matrixnak, a megfigyelések matrixát pedig a megfigyelések speciálisan redukált matrixának tekintjük.

A közös faktortér dimenziószáma (m) az alábbi tétel alapján határozható meg:

Ha a redukált korrelációs matrix rangja m, akkor a benne szereplő korrelációs együtthatókat teljes mértékben reprodukáló, lineárisan független közös faktorok legkisebb száma m, azaz a közös faktorok tere legalább m dimenziós ([6], 64. old.)

Mivel a redukált korrelációs matrix rangja a kommunalitások alkalmas megválasztása esetén *kisebb* a teljes korrelációs matrix rangjánál, tételünk szerint a közös faktorok számára nézve mindig teljesül az

$$1 \leq m \leq n$$

[10] A komponenselemzés modelljét az 1.3 alpontban ismertetem.

egyenlőtlenség. A közös faktorok számára vonatkozó hipotézis helyessége az (1.1.18) kritériummal ellenőrizhető, bár egyes faktoranalitikus megoldásokra vannak ennél jóval egzaktabb kritériumok is [16].

Annak, hogy az (1.1.18) kritérium nem teljesül, két oka lehet. Az egyik az, hogy *kevés* közös faktort vettünk be (1.1.2) modellünkbe, a másik pedig az, hogy nem igaz az (1.1.2) modellbe foglalt *linearitási hipotézis*. Arról, hogy a két ok közül melyik forog fenn, csak úgy lehet meggyőződni, hogy a közös faktorok számát növeljük, és ismételten ellenőrizzük, hogy teljesül-e (1.1.18).

Mivel a később ismertetendő faktoranalitikus megoldási módszerek egy része feltételezi a kommunalitások előzetes ismeretét, röviden foglalkoznunk kell a kommunalitások becslésével is. Az eddig elmondottakból következik, hogy elvileg olyan *maximális* kommunalitások meghatározása lenne a cél, amelyek *minimalizálják* a redukált korrelációs matrix rangját azon feltétel mellett, hogy $\mathbf{R}_h$ pozitív definit. Ez ugyanis annak a biztosítéka, hogy minimális számú közös faktorral magyarázzuk meg változóink szórásnégyzetének maximális hányadát, azaz a változók minimális számú közös faktorral való leírása minimális információveszteséggel járjon. E feltételes szélsőértékfeladat azonban csak igen szigorú, a gyakorlatban szinte sohasem teljesülő feltételek mellett oldható meg. Ezért a gyakorlatban csak ezen elvi optimumot közelítő becslésekről lehet szó.

A kommunalitások részleges vagy teljes módon becsülhetők. Míg a részleges becslések a teljes korrelációs matrix nem-diagonális elemeinek csak egy részét, addig a teljes becslések a nem-diagonális elemek összességét használják fel.

A legegyszerűbb *részleges* becslési eljárás az, amikor a h_j^2 kommunalitásnak a z_j változó és az azzal legjobban korreláló változó közötti korrelációs együtthatót tekintjük, azaz

$$h_j^2 = \max_{i \neq j} r_{ji} \tag{1.2.3}$$

Egy másik, ugyancsak e csoportba tartozó becslési eljárás a

$$h_j^2 = \frac{r_{jk} r_{jl}}{r_{kl}} \tag{1.2.4}$$

uń. triádok alkalmazása, ahol a z_k és z_l a z_j-vel legjobban korreláló két változó.

Az egyik legegyszerűbb teljes becslés a következő:

$$h_j^2 = \frac{n}{n-1} \frac{\left(\sum_{k=1}^{n} r_{jk}\right)^2}{\sum_{k=1}^{n} \sum_{l=1}^{n} r_{kl}} \qquad (k \neq j; k \neq l) \tag{1.2.5}$$

Végül a kommunalitások „lehető legjobb" becslésének az adott z_j változó összes többi változóra vonatkozó többszörös korrelációs együtthatójának négyzetét tekintik, ami az

$$R_{j(n-1)}^2 = 1 - \frac{1}{r^{jj}} \tag{1.2.6}$$

módon határozható meg, ahol r^{jj} a teljes korrelációs matrix *inverzének* j-edik diagonális eleme. Ennek az a magyarázata, hogy Dwyer bebizonyította az

(1.2.7) $$R^2_{j(n-1)} \leq h^2_j \qquad (j = 1, 2, \ldots, n)$$

egyenlőtlenséget és azt, hogy (1.2.7) akkor teljesül egyenlőség formájában, ha a redukált korrelációs matrix rangja minimális.[11]

1.3 *A faktoranalízis megoldási módszerei*

A faktoranalitikus módszerek áttekintése előtt célszerű a *faktoranalízis* és a *komponenselemzés* közötti kapcsolat tisztázása. A komponenselemzés a faktoranalízis azon speciális esete, amelyben minden egyes változó kommunalitása egy. Ez azt jelenti, hogy minden változó teljes szórásnégyzetét meg kívánjuk magyarázni a közös faktorokkal, azaz a redukált korrelációs matrix helyett a teljes korrelációs matrixból indulunk ki. Ez azt jelenti, hogy az (1.1.2) modell helyett a

$$\mathbf{z} = \mathbf{A}_k \cdot \mathbf{k}$$

modellből indulunk ki, ami *csak* közös faktorokat tartalmaz. E módszer nem a változók számának csökkentését, hanem olyan új változók bevezetését tűzi ki célul, amelyen páronként ortogonálisak és egyre csökkenő mértékben járulnak hozzá az eredeti változók szórásnégyzetéhez. Ez igen hasonló az alapvető faktorok módszerének célkitűzéséhez, aminek részletes ismertetésére a következő alpontban térek ki.

Az ezután következő rendszerezés csak a *faktoranalitikus* megoldásokra vonatkozik. E megoldások két nagy csoportba, a *közvetlen megoldások* csoportjába és az ún. *leszármaztatott* megoldások csoportjába sorolhatók. Ez utóbbiak az előbbiekből ortogonális transzformációval származtatott megoldások, míg az előbbiek közvetlenül a redukált korrelációs matrix alapján határozhatók meg. A leszármaztatott megoldásoknak az a céljuk, hogy egy valamilyen közvetlen módszerrel már meghatározott **A** sémamatrixból kiindulva egy adott tulajdonságokkal rendelkező, többnyire az **A**-nál egyszerűbb szerkezetű sémamatrixú megoldást szolgáltassanak. A Thurstone-féle ún. *egyszerű struktúrák* is ilyen leszármaztatott megoldások, melyekre az jellemző, hogy sémamatrixuk a lehető legnagyobb számú zérust tartalmazza. Ilyen „egyszerű struktúrákhoz" többnyire csak páronként korrelált közös faktorok megengedésével lehet eljutni. A leszármaztatott megoldások léte azt bizonyítja, hogy nem létezik egyértelmű faktoranalitikus megoldás, mert bármely adott megoldáson egy ortogonális transzformációt vérehajtva újabb megoldáshoz jutunk.

A közvetlen megoldások két szempont szerint rendszerezhetők. Az első szempont a megoldás alapját képező modell változóinak *komplexitása*. Eszerint egy-, két- stb. faktoros megoldásokról beszélhetünk.

A rendszerezés másik szempontja az, hogy a megoldás igényeli-e a kommunalitások előzetes becslését vagy nem. A kommunalitások előzetes becslését igénylő megoldások csoportjába tartozik az *alapvető faktorok módszere* (Principal Factor Solution), melynek részletes ismertetésére a következő alpontban térek ki, valamint az ún. *centroid* módszer, mely az előbbihez közeleső eredményeket szolgáltat. E módszert eredetileg az alapvető faktorok módszerének nagy számításigénye miatt dolgozták ki, jelentősége azonban az elektronikus számító-

[11] Lásd: DWYER, P. S.: The Contribution of An Orthogonal Factor Solution to Multiple Correlation (Psychometrika 4 (1939)).

gépek megjelenése óta csökkent. E módszereket igen sok szerző csak előzetes, kiinduló megoldásoknak tekinti, melyek aztán valamilyen leszármaztatott megoldás alapját képezik. Végül e csoportba tartoznak az ún. *többfaktoros* megoldások is, amelyek egymást „átfedő" csoportfaktorokat is tartalmazhatnak, tehát egy változó leírásában egynél több csoportfaktor is szerepelhet.

A kommunalitások előzetes becslését nem igénylő megoldásokat csak a megoldások rendszerezésének első szempontja szerint szokás megkülönböztetni.

Ugyancsak a közvetlen megoldások közé tartozik a faktorsúlyok Lawley-tól származó, maximum likelihood módszeren alapuló becslése, ami azonban szigorúan véve nem tartozik a közvetlen megoldások egyik nagy csoportjába sem.

Itt kívánom végül megjegyezni, hogy véleményem szerint nem sok értelme van a páronként korrelált közös faktorokat tartalmazó megoldásoknak, mert ez igen megnehezíti a kapott eredmények értelmezhetőségét. Részben ez az oka annak is, hogy a faktoranalízis számos megoldási módszere közül csak az alapvető faktorok módszerét ismertetem részletesen.

1.4 *Az alapvető faktorok módszere*

A faktoranalízis módszerei kétféle módon alkalmazhatók. Az első esetben rendelkezünk valamilyen előzetes, *a priori* modellel a vizsgált jelenségre vonatkozóan, s ilyenkor a faktoranalízis feladata e modell paramétereinek becslése, s az a priori modellel felírt hipotézis helyességének ellenőrzése. Ez utóbbi célra részben az (1.1.18) kritérium, részben pedig különböző matematikai statisztikai próbák állnak rendelkezésre. A második esetben nincs semmiféle elképzelésünk az általunk vizsgált jelenség modelljére vonatkozóan, s éppen e modell megkeresésére a faktoranalízis célja. A faktoranalízis megoldási módszerei közül e célra véleményem szerint az alapvető faktorok módszere a legalkalmasabb, s elsősorban ezért tartom szükségesnek e módszer részletes ismertetését. Ezt indokolja továbbá az előző alpont végén említett tény is, melyből számos a 2. részben ismertetendő előny származik.

Az alapvető faktorok módszere olyan közös faktorok meghatározását tűzi ki célul, amelyek

1. a lehető legnagyobb mértékben járulnak hozzá az összes változó

$$H^2 = \sum_{j=1}^{n} h_j^2 = \sum_{p=1}^{m} V_p \tag{1.4.1}$$

teljes kommunalitásához,

2. a lehető legjobban megközelítik a reprodukálandó redukált korrelációs matrixot ($\mathbf{R}_h$-t), és

3. páronként ortogonálisak.

Mivel az ezután következő eljárás lényege az lesz, hogy a közös faktorokat egyenként, az összes változó H^2 teljes kommunalitásához való hozzájárulásuk nagyságának csökkenő sorrendjében határozzuk meg, meg kell vizsgálni azt, hogy az egyes közös faktorok milyen szerepet játszanak az első két kritérium teljesülésében. Nyilvánvaló, hogy ilyen szempont szerint az első kritérium az (1.1.8)-cal definiált V_p ($p = 1, 2, \ldots, m$) mennyiségek egyenkénti maximalizálását jelenti.

A második kritériummal kapcsolatban először is azt jegyezzük meg, hogy a redukált korrelációs matrix „lehető legjobb" megközelítésén a legkisebb négyzetek módszere értelmében vett közelítést értjük. Ami az egyes közös faktoroknak a második kritérium teljesülésében betöltött szerepét illeti, könnyen belátható, hogy a faktoranalízis (1.1.17)-tel felírt Thurstone-féle alaptétele felírható az

$$\mathbf{R}_{r|h} = \sum_{p=1}^{m} \mathbf{a}_p \mathbf{a}_p^* = \sum_{p=1}^{m} \mathbf{Q}_p \tag{1.4.2}$$

alakban is, ahol $\mathbf{a}_p$ az $\mathbf{A}_k$ közös sémamatrix p-edik oszlopa, és $\mathbf{a}_p \mathbf{a}_p^* = \mathbf{Q}_p$. Ez a felbontás pedig éppen az egyes közös faktoroknak a redukált korrelációs matrix[12] reprodukálásában betöltött szerepét mutatja.

Mint látni fogjuk, a három fenti kritérium közül az első és az utolsó el is hagyható, mert azok teljesülése a második teljesüléséből automatikusan következik. Első lépésként ezért egy olyan K_1 közös faktort keresünk, ami a lehető legnagyobb mértékben résztvesz a redukált korrelációs matrix reprodukálásában, azaz amelyre nézve az

$$S_1 = \sum_{j=1}^{n} \sum_{k=1}^{n} (r_{jk} - a_{j1} a_{k1})^2 \qquad (r_{jj} = h_j^2)$$

eltérés-négyzetösszeg minimális. Az egyszerűség kedvéért írjuk fel S_1-et az

$$S_1 = tr\,[(\mathbf{R}_h - \mathbf{a}_1 \mathbf{a}_1^*)(\mathbf{R}_h - \mathbf{a}_1 \mathbf{a}_1^*)] = tr(\mathbf{R}_h^2) - 2\,\mathbf{a}_1^* \mathbf{R}_h \mathbf{a}_1 + (\mathbf{a}_1^* \mathbf{a}_1)^2 \tag{1.4.3}$$

módon,[13] ahol $\mathbf{a}_1$ az $\mathbf{A}_k$ közös sémamatrix első oszlopa.

Elvégezve (1.4.3) $\mathbf{a}_1$ szerinti deriválását, és a deriváltat $\mathbf{0}$-val egyenlővé téve, majd az így kapott egyenletet rendezve:

$$(\mathbf{R}_h - \lambda_1 \mathbf{E})\mathbf{a}_1 = \mathbf{0}, \tag{1.4.4}$$

ahol $\lambda_1 = \mathbf{a}_1^* \mathbf{a}_1$. Arra az eredményre jutottunk tehát, hogy a 2. kritériumnak elegettevő $\mathbf{a}_1$ vektor az $\mathbf{R}_h$ matrix egyik sajátvektora. Az (1.4.4)-ből adódó

$$\mathbf{R}_h \mathbf{a}_1 = \lambda_1 \mathbf{a}_1 \tag{1.4.5}$$

összefüggést figyelembe véve (1.4.3) az

$$S_1 = tr(\mathbf{R}_h^2) - \lambda_1^2$$

alakba megy át. Tekintettel arra, hogy egy szimmetrikus pozitív definit, illetve pozitív szemidefinit matrix sajátértékei pozitív, illetve nem-negatív valós értékek, az S_1 eltérés-négyzetösszeg csak abban az esetben lehet minimális, ha λ_1 az $\mathbf{R}_h$ matrix legnagyobb sajátértéke. Az $\mathbf{a}_1$ vektort tehát az $\mathbf{R}_h$ legnagyobb sajátértékéhez tartozó sajátvektorok halmazából kell kiválasztanunk, amit az első kritérium figyelembevételével tehetünk meg. Ugyanis a maximalizálandó V_1 mennyiség éppen λ_1 gyel egyenlő, amiből az következik, hogy az

[12] Abban az esetben, ha komponenselemzést végzünk, a redukált korrelációs matrix szerepét a teljes korrelációs matrix veszi át.

[13] A $tr(\mathbf{A})$ jelölés az $\mathbf{A}$ kvadratikus matrix nyomát jelenti. Egy kvadratikus $\mathbf{A}$ matrix nyomán az $\mathbf{A}$ matrix diagonális elemeinek összegét értjük. Szokásos még a matrix nyomának $Sp(\mathbf{A})$ jelölése is.

első két követelménynek együttesen elegettevő $\mathbf{a}_1$ vektor az $\mathbf{R}_h$ matrix legnagyobb λ_1 sajátértékéhez tartozó $\sqrt{\lambda_1}$ hosszúságú sajátvektor. Vegyük észre, hogy az $\mathbf{a}_1$ ilyen megválasztása esetén a harmadik követelmény is teljesül, mert a sajátvektor definíciója értelmében $\mathbf{a}_1 = \mathbf{0}$ nem állhat fenn.

Az $\mathbf{A}_k$ matrix második, $\mathbf{a}_2$ oszlopának meghatározása az $\mathbf{a}_1$ meghatározásához képest azzal a különbséggel történik, hogy $\mathbf{R}_h$ szerepét az

$$\mathbf{R}_1 = \mathbf{R}_h - \mathbf{a}_1 \mathbf{a}_1^* \tag{1.4.6}$$

ún. *első reziduális korrelációs matrix* veszi át. Az előzőekhez teljesen hasonlóan belátható, hogy az első két kritériumot egyszerre kielégítő $\mathbf{a}_2$ vektor az $\mathbf{R}_1$ matrix λ_2 vel jelölt legnagyobb sajátértékéhez tartozó $\sqrt{\lambda_2}$ hosszúságú sajátvektor, ahol $\lambda_2 = \mathbf{a}_2^* \mathbf{a}_2$. Könnyen igazolható, hogy az így kapott $\mathbf{a}_2$ vektor ortogonális az előbbi $\mathbf{a}_1$ vektorra.

Az $\mathbf{A}_k$ oszlopai tehát, most már általánosan fogalmazva, a következőképpen határozhatók meg. Definiáljuk az s-edik reziduális korrelációs matrixot az

$$\mathbf{R}_s = \mathbf{R}_h - \sum_{p=1}^{s}{}' \mathbf{a}_p \mathbf{a}_p^* = \mathbf{R}_h - \sum_{p=1}^{s}{}' \mathbf{Q}_p \tag{1.4.7}$$

$$(s = 0, 1, \ldots m-1)$$

előírással, ahol $\mathbf{R}_0 = \mathbf{R}_h$. Az $\mathbf{A}_k$ matrix $(s+1)$-edik oszlopa ekkor az

$$S_{s+1} = \operatorname{tr}\left[(\mathbf{R}_s - \mathbf{a}_{s+1}\mathbf{a}_{s+1}^*)(\mathbf{R}_s - \mathbf{a}_{s+1}\mathbf{a}_{s+1}^*)\right] \tag{1.4.8}$$

$$(s = 0, 1, 2, \ldots, m-1)$$

eltérés-négyzetszöget minimalizáló $\mathbf{a}_{s+1}$ vektor. Az $\mathbf{a}_1$ meghatározásához teljesen hasonlóan eljárva belátható, hogy $\mathbf{a}_{s+1}$ az $\mathbf{R}_s$ matrix legnagyobb, $\lambda_{s+1} = \mathbf{a}_{s+1}^* \mathbf{a}_{s+1}$ sajátértékéhez tartozó, $\sqrt{\lambda_{s+1}}$ hosszúságú sajátvektor. Teljes indukcióval bebizonyítható, hogy az egymást követő

$$\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_m$$

vektorok páronként ortogonálisak.

Az $\mathbf{a}_p$ $(p = 1, 2, \ldots, m)$ vektorok páronkénti ortogonalitására támaszkodva az is könnyen bebizonyítható, hogy az előbbi eljárás során adódó

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$$

sajátértékek mindegyike egyben az $\mathbf{R}_h$ redukált korrelációs matrixnak is sajátértéke, s így nincs szükség az eljárás során az egymást követő reziduális korrelációs matrixok meghatározására.

Az utolsó $\mathbf{R}_m$ reziduális matrixról a sajátértékek négyzetösszegére vonatkozó ismert tétel ([1], 224. old.) felhasználásával bebizonyítható, hogy értéke $\mathbf{0}$, azaz a fenti eljárás teljes mértékben reprodukálja a redukált korrelációs matrixot.

Ezzel az általunk kitűzött feladatot matematikailag megoldottnak tekinthetjük. *Kaiser* szerint az $\mathbf{A}_k$ közös sémamatrixnak csak az $\mathbf{R}_h$ matrix 1-nél nagyobb sajátértékeihez tartozó oszlopait érdemes meghatározni, mert az ezekhez tartozó közös faktorok az egyes változók kommunalitását csaknem teljes mértékben megmagyarázzák. Ily módon eljárva a közös faktorok száma az eredeti változók számának általában 1/6—1/3-a.

A faktoranalízis módszerét ismertető rész befejezéseként a páronként korrelálatlan közös faktorokat tartalmazó faktoranalitikus megoldás megadásának szokásos táblázatos formáját ismertetem.

Változó (j)	Sémaegyütthatók				Kommunalitás		
	1	2	...	m	Eredeti h_j^2	Számított $\sum_{p=1}^{m} a_{jp}^2$	Különbség $h_j^2 - \sum_{p=1}^{m} a_{jp}^2$
1	a_{11}	a_{12}	...	a_{1m}			
2	a_{21}	a_{22}	...	a_{2m}			
.	.	.		.			
.	.	.		.			
.	.	.		.			
n	$\mathbf{a}_{n1}$	$\mathbf{a}_{n2}$	...	a_{nm}			
Összesen	X	X	...	X			
A faktorok hozzájárulása (V_p)	$\mathbf{a}_1^*\mathbf{a}_1$	$\mathbf{a}_2^*\mathbf{a}_2$	...	$\mathbf{a}_m^*\mathbf{a}_m$	X	X	X
Az eredeti teljes kommunalitás %-ában							

2. A közgazdasági alkalmazás lehetőségeiről

A faktoranalízis közgazdasági alkalmazásának *lehetőségét* elsősorban a közgazdasági jelenségek bonyolultsága, összetett volta adja meg. A faktoranalízis alkalmazása ugyanis elsősorban olyan esetekben vezethet hasznos eredményre, amikor a vizsgálatba vont változók szoros sztochasztikus kapcsolatban állnak egymással, kölcsönösen függenek egymástól. Ellenkező esetben, tehát független változók esetén csak egyszerűen arról van szó, hogy egy sor bonyolult számítás után magukat az eredeti változókat kapjuk vissza közös faktorokként. Az ilyen felesleges számítások azonban a vizsgált változók korrelációs matrixának előzetes vizsgálata alapján elkerülhetők.

A faktoranalízis legjellegzetesebb közgazdasági alkalmazásai az alábbi három csoportba sorolhatók ([11]):

1. a változók számának csökkentése,
2. osztályozási (csoportosítási) feladatok megoldása,
3. indexsúlyozási feladatok megoldása.

Az egyes alkalmazási területekkel itt elsősorban elméleti szempontból foglalkozom, a konkrét gyakorlati alkalmazásokkal kapcsolatban a téma egyre jobban bővülő irodalmára utalok. Az egyes alkalmazási területek részletesebb ismertetése előtt külön alpontban foglalkozom a faktorok értelmezésével.

2.1 *A faktorok előállítása és értelmezése*

Ebben az alpontban először a faktorok változókkal történő előállítását ismertetem. Ekkor célszerű az alábbi két eset megkülönböztetése.

1. *Komponenselemzés:* Ekkor a közös faktorok száma n, s azok egyértelműen előállíthatók a z_j változók lineáris kombinációjaként a

$$\mathbf{k} = \mathbf{A}_k^{-1}\mathbf{z} = \mathbf{A}_k^{*}\mathbf{z} \tag{2.1.1}$$

formula alapján.[14] Mint az már ismeretes, a komponenselemzés esetében nincsenek egyedi faktorok.

2. *Faktoranalízis:* Tekintettel arra, hogy ebben az esetben a megoldás egyedi faktorokat is tartalmaz, nem lehet szó a faktorok egyértelmű előállításáról. Itt a faktorok előállításának egyik legegyszerűbb módját, az ún. *regressziós módszert* ismertetem. ([6] és [16]) Ebben az esetben — a levezetések mellőzésével —

$$\hat{\mathbf{f}} = \mathbf{S}^{*}\mathbf{R}^{-1}\mathbf{z} = \Phi\,\mathbf{A}^{*}\,\mathbf{R}^{-1}\mathbf{z}, \tag{2.1.2}$$

ami páronként korrelálatlan közös faktorok esetén az

$$\hat{\mathbf{f}} = \mathbf{A}^{*}\,\mathbf{R}^{-1}\mathbf{z} \tag{2.1.3}$$

alakba megy át. Az $\hat{\mathbf{f}}$ jelölés mindkét esetben arra utal, hogy itt a faktorok egy lehetséges becsléséről van szó.

Mivel ezen előállítás egyértelműsége nem jelenti feltétlenül azt, hogy egy faktor előállításában csak egy változó szerepel, a faktorok értelmezhetőségének problémáját célszerű mindjárt az $m < n$ esetre vonatkoztatva tárgyalni.

Ha a kapott eredményeket valamilyen jelenség közelítésére, előrejelzésére kívánjuk felhasználni, akkor nincs is feltétlenül szükség a kapott faktorok értelmezésére. Ilyen volt a helyzet bevezető példánk esetében is, amikor elsősorban az általános faktorok számára vonatkozó hipotézis helyessége bírt döntő jelenséggel. Ennek eldöntése után már szinte „magától" adódott a kapott eredmény értelmezése. Ugyanez a helyzet akkor is ha a faktoranalízist egy a priori modell helyességének ellenőrzésére kívánjuk felhasználni. Más a helyzet azonban akkor, ha a kapott eredményeket valamilyen jelenség közgazdasági elemzésére, mélyebb megismerésére kívánjuk felhasználni. Ekkor ugyanis óhatatlanul szembekerülünk a faktorok értelmezésének, interpretálásának problémájával. Ilyen esetekben döntő jelentőségű az, hogy sikerül-e a kapott faktorokat valamilyen közgazdaságilag értelmezhető, esetleg közvetlenül megfigyelhető változóval azonosítani.

A faktorok értelmezésének alapját a faktorok változók segítségével történő (2.1.2) előállítása, és a megoldás szerves részeként adódó *struktúramatrix* vizsgálata képezi. Ha a faktorok (2.1.2) előállítását megvizsgálva azt tapasztaljuk, hogy az egyes faktorok olyan diszjunkt változócsoportok lineáris kombinációi, hogy az egyes csoportokba tartozó változóknak létezik valamilyen lényeges közös jellemzője, s a különböző változócsoportok közös jellemzői mind különbözők, akkor a faktorok rendre a hozzájuk tartozó változócsoportok közös jellemzőivel azonosíthatók.

Ha ez a feltétel nem teljesül, akkor azt a tényt használhatjuk fel a faktorok értelmezésére, hogy tekintettel a változók és faktorok standardizáltságára az egyes faktoroknak a változókból való összetevődését mutató szorzókonstansok nagysága éppen az egyes változók faktorok kialakításában játszott szere-

[14] Az $\mathbf{A}_k^{-1} = \mathbf{A}_k^{*}$ egyenlőség a közös faktorok páronkénti ortogonalitásából következik.

pének fontosságát mutatják. Erre támaszkodva pedig kiválasztható az adott faktor értékeinek alakulását a legdöntőbb mértékben befolyásoló változó, amivel az adott faktor azonosítható.

Tekintettel arra, hogy itt már csak páronként korrelálatlan közös faktorokat tartalmazó megoldásokkal foglalkozunk ez az eljárás azzal egyenértékű, hogy az adott faktort a vele legszorosabban korreláló változóval azonosítjuk.

Ha a faktorok változókkal történő azonosításakor a faktorok és a változók közötti korrelációs együtthatókat vesszük alapul, akkor sok esetben igen hasznos lehet az alapvető faktorok módszerével kapott megoldást *kiinduló megoldásnak* tekinteni, s abból egy ortogonális transzformáció segítségével egy újabb megoldást származtatni. K. A. Schäffer véleménye szerint ([12]) erre a célra a *Kaiser*-től származó *varimax-módszert* érdemes alkalmazni. Az e módszerrel kapott megoldás igen közel áll a már említett „egyszerű struktúrához", ami sokszor igen megkönnyíti a kapott eredmények értelmezését. A varimax módszer ismertetésére nem térek ki, részletes leírása pl. a [6]-ban található meg.

A kapott eredmények értelmezésénél felmerülő nehézségek véleményem szerint elsősorban a közelíteni kívánt közgazdasági jelenségek rendkívül bonyolult természetéből fakadnak. A faktoranalitikus megoldások éppen arra mutatnak rá, hogy a közgazdasági jelenségek annyira összetettek, hogy csak ún. összetett változók segítségével közelíthetők. Az értelmezésüknél fellépő nehézségek ellenére úgy vélem, hogy a faktoranalízis egyrészt a közgazdasági jelenségek modellezésének igen hatékony segédeszköze lehet, másrészt pedig jól felhasználható az egyes jelenségekre adott — sokszor igen semmitmondó — definíciók pontosabbá tételére is. A faktoranalízis ilyen jellegű alkalmazhatóságáról a következő két alpontban lesz szó.

2.2 *A változók számának csökkentése*

A közgazdasági elemzések során ma már egyre gyakoribb az ún. *regressziós modellek* alkalmazása. Ezek közül is a leggyakoribb az

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \tag{2.2.1}$$

lineáris modellek alkalmazása, ahol

$\mathbf{y} = [y, y_2, \ldots, y_N]^*$
— az eredményváltozóra vonatkozó N megfigyelést tartalmazó oszlopvektor

$\mathbf{X} = [x_{ij}]$, $(i = 1, 2, \ldots, N,\ j = 1, 2, \ldots, n)$
— a modellben szereplő n magyarázó változóra vonatkozó megfigyeléseket tartalmazó $N \times n$ méretű matrix

$\beta = [\beta_1, \beta_2, \ldots, \beta_n]^*$
— a modell ismeretlen paramétereit tartalmazó oszlopvektor

$\epsilon = [\varepsilon_1, \varepsilon_2 \ldots \varepsilon_N]^*$
— az ún. hibavektor.

A modellben szereplő $Y, X_1, X_2, \ldots, X_n$ változókról feltesszük, hogy azok standardizált formában adott valószínűségi változók, melyek együttes elosz-

lásának sűrűségfüggvénye $f(Y, X_1, X_2, \ldots, X_n)$, Y feltételes várható értéke pedig

$$E(Y \mid X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = f(x_1, x_2, \ldots, x_n, \beta_1, \beta_2, \ldots, \beta_n)$$

alakú, ahol a β_i-k ismeretlen együtthatók, az x_i-k pedig az egyes valószínűségi változók rögzített értékei. A (2.2.1) modell ezen ún. elméleti regresszió-függvény lineáris közelítése.

A (2.2.1) modellben szereplő β paramétervektor legkisebb négyzetek módszere szerinti becslése,

$$\mathbf{b} = (\mathbf{X}^* \mathbf{X})^{-1} \mathbf{X}^* \mathbf{y}$$

alakú, ami azt jelenti, hogy a (2.2.1) elméleti modellt az

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \tag{2.2.2}$$

formában becsüljük, ahol $\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b}$.

Könnyen belátható, hogy a β_i regressziós együtthatók e becsléseinek varianciája akkor lesz minimális, ha a magyarázó változók páronként korrelálatlanok. Ez azonban a közgazdasági gyakorlatban csak igen ritkán teljesül. Sokkal inkább jellemző a magyarázó változók páronkénti korreláltsága, aminek következményét már az 1.2 alpontban említettük. E jelenség az irodalomban *multikollinearitás* néven ismert.

A multikollinearitás fellépése azonban nemcsak a regressziós együtthatók hibáját növeli meg, hanem megnehezíti a magyarázó változók hatásának szétválasztását is, ami már közgazdasági probléma. A magyarázó változók hatásának szétválaszthatatlansága ugyanis azt jelenti, hogy nem, illetve csak erős fenntartásokkal adható meg a regressziós együtthatók szokásos értelmezése. Minél erősebb fokú a multikollinearitás, annál inkább számolni kell a jelenlétéből adódó káros következménnyel.

Ha a modellben szereplő változókra vonatkozó megfigyelések adott száma (N) mellett növeljük a modell változóinak számát, akkor ezzel párhuzamosan egyre növekszik a multikollinearitás veszélye is. Éppen ezért a változók számát ésszerűen kell megválasztani. A változók ésszerű, „optimális" számát igen nehéz pontosan definiálni. Célszerű azonban optimalitási kritériumnak a többszörös determinációs együttható — a többszörös korrelációs együttható négyzetének — értékét tekinti. Ha azonban a többszörös determinációs együttható értékét minden megkötés nélkül maximalizálnánk, akkor ez a változók számának minden határon túli növelését igényelné, ami viszont a multikollinearitás fokának növekedését is maga után vonná. Ezzel szemben ha a többszörös determinációs együttható értékét azon feltétel mellett maximalizáljuk, hogy a modellbe kerülő magyarázó változók páronként korrelálatlanok legyenek, akkor a feltétel egyrészt gátat szab a magyarázó változók száma minden határon túli növekedésének, másrészt a multikollinearitást is kiküszöböli. Ez az optimálitási kritérium igen hasonló az alapvető faktorok módszerénél alkalmazotthoz. Hangsúlyozni kívánom, hogy ez a kritérium csak a változók optimális számához való közelítés egyik lehetséges módja, s korántsem oldja meg teljesen a problémát. Mindenesetre a vázolt kritérium elég ésszerűnek látszik.

Ha elfogadjuk az előbbi kritériumot, akkor kézenfekvőnek látszik az a gondolat, hogy a (2.2.1) modellben szereplő magyarázó változókat azok páron-

ként korrelálatlan közös faktoraival helyettesítsük. Ez az alapgondolata az M. G. Kendalltól származó ún. mesterséges *ortogonalizálás módszerének.*

Induljunk ki a modellünk szempontjából szóbajöhető maximális számú magyarázó változóból, s jelöljük ezek halmazát X-szel. Tekintsük ezután az X változóhalmaz

$$\mathbf{x} = \mathbf{A}_k \mathbf{k} + \mathbf{A}_u \mathbf{u} = \mathbf{A}_k \mathbf{k} + \mathbf{v} \tag{2.2.3}$$

előállítását, ahol az alkalmazott jelölések pontosan megegyeznek az 1.1 alpontban használtakkal, $\mathbf{v}$ pedig az egyedi faktorok elhanyagolásából adódó hibát jelenti. Ha a (2.2.3)-ban szereplő közös faktorokat az 1.4 pontban ismertetett alapvető faktorok módszerével határoztuk meg, akkor ez azt jelenti, hogy a (2.2.1) modellben szereplő magyarázó változókat páronként korrelálatlan mesterséges változókkal helyettesíthetjük. Fogalmazzuk meg ezután a (2.2.1) modell

$$\mathbf{y} = \mathbf{K}^* \alpha + \delta \tag{2.2.4}$$

módosított változatát, amit az eredeti modell reparametrizált alakjának nevezünk.

A (2.2.4) modellt ezután az

$$\mathbf{y} = \mathbf{K}^* \mathbf{a} + \mathbf{d}$$

formában becsüljük, ahol

$$\mathbf{a} = (\mathbf{K}\,\mathbf{K}^*)^{-1}\,\mathbf{K}\,\mathbf{y} \text{ és } \mathbf{d} = \mathbf{y} - \mathbf{K}^* \mathbf{a}$$

és $\mathbf{a}$ m-elemű oszlopvektor.

Kihasználva azt a tényt, hogy a közös faktorok egyrészt páronként korrelálatlanok, másrészt a centrális határeloszlás tétele értelmében közelítőleg normális eloszlásúak, reparametrizált modellünk paraméterei — amelyek nem egyeznek meg az eredeti (2.2.1) modell paramétereivel, hanem azoknak lineáris függvényei — sokkal kisebb hibával becsülhetők, mint az eredeti modell paraméterei, sőt azoknál sokkal megalapozottabban vizsgálhatók a szokásos statisztikai próbákkal is.

Az ezután következő lépés a megoldásul kapott közös faktorok, s ezt felhasználva a reparametrizált modellben szereplő paraméterek értelmezése. Ez az előző alpontban elmondottaknak megfelelően történhet. Ha a közös faktorok, s így a reparametrizált modellben szereplő paraméterek végképp nem értelmezhetők, akkor kénytelenek vagyunk a

$$\mathbf{b}' = \mathbf{A}_k \mathbf{a} \tag{2.2.5}$$

transzformáció segítségével visszatérni az eredeti modell paramétereire. Ez azonban azt jelenti, hogy eredeti feladatunkat csak látszólag oldottuk meg. Itt ugyanis csak arról van szó, hogy az eredeti modell X magyarázó változóhalmaza által tartalmazott információ-mennyiség csökkentése után nyerjük a paraméterek becslését, ami esetleg még az eredeti adatok alapján kapható becsléseknél is bizonytalanabb lehet. Ennek ellenére a mesterséges ortogonalizálás módszere sok esetben hasznos eredményre vezethet.

Az eddigiek során a faktoranalízisnek a multikollinearitás kiküszöbölésére vonatkozó alkalmasságát hangsúlyoztam. Mivel azonban a módszer ilyen esetekre történő alkalmazása szinte minden esetben együttjár a magyarázó

változók számának nagymértékű csökkentésével, a faktoranalízis nemcsak a multikollinearitás kiküszöbölésére szolgáló hatékony eszköznek tekinthető, hanem sikerrel alkalmazható igen bonyolult közgazdasági jelenségek viszonylag egyszerű, kevés számú változóval való közelítésére is. Ha ugyanis a faktorok változókkal való azonosítását, interpretálását a 2.1. pontban elmondottaknak megfelelően végezzük el, akkor a faktoranalízis alkalmas arra, hogy egy regressziós modell szempontjából szóbajövő nagyszámú magyarázó változó közül kiemelje a legfontosabb, közelítőleg páronként korrelálatlan változókat. Ilyen értelemben tehát a faktoranalízis a *modellalkotás igen hatékony segédeszközének* tekinthető. Egy ilyen tárgyú konkrét alkalmazásra még egy későbbi cikkben szeretnék visszatérni.

A mesterséges ortogonalizálás módszerét R. Stone alkalmazta először a gyakorlatban, aki az USA bruttó nemzeti termékére, illetve nemzeti jövedelmére ható 17 tényezőből indult ki, és arra az eredményre jutott, hogy a vizsgált függő változók alakulása 3 közös faktor segítségével gyakorlatilag teljes mértékben (97,5%-ban) megmagyarázható ([11]).

A faktoranalízis ezen túlmenően jól felhasználható egy már adott regressziós modell specifikációjának vizsgálatára is. Ha ugyanis a modell magyarázó változóinak faktorelemzését elvégezve arra az eredményre jutunk, hogy a magyarázó változóknak létezik egy *lényeges általános faktora*, akkor ez nagymértékben valószínűsíti a modell helyes specifikációját.

2.3 *Csoportosítási feladatok*

Igen sok esetben merül fel az az igény, hogy egy N elemből álló sokaság egységeit egy vagy egyidejűleg több ismérv szerint olyan kisebb csoportokba, részsokaságba soroljuk, hogy az egy csoportba tartozó egységek minél homogénebbek legyenek a csoportképző ismérv(ek) szempontjából.

A faktoranalízis ilyen területen történő alkalmazása két esetben válhat szükségessé. Az első eset az, amikor egy csoportképző ismérvet jelöltünk ugyan ki, de az közvetlenül nem mérhető. Ilyen csoportképző ismérv lehet például az ún. „városiassági fok", vagy „gazdasági fejlettség". Mindkét példaként említett csoportképző ismérvre az jellemző, hogy nem lehetséges egyetlen olyan mérhető változót kijelölni, amely teljesen azonosítható volna az adott csoportképző ismérvvel, sőt éppen ellenkezőleg, mindkét ismérvünkre az jellemző, hogy számtalan olyan mérhető tényező nevezhető meg, amely többékevésbé szoros kapcsolatban áll azokkal

Ilyen esetekben a következőképpen járhatunk el. Gyűjtsük össze mindazon mérhető változókat, melyekről feltételezhető, hogy valamilyen kapcsolatban állnak a kiválasztott csoportképző ismérvvel. Legyenek ezek az $X_1, X_2, \ldots, X_n$ változók, melyek között valamilyen többé-kevésbé önkényesen számszerűsített minőségi ismérvek is szerepelhetnek, s melyekről feltesszük most, hogy standardizált formában adottak. Határozzuk meg ezután a figyelembe vett változók **A** sémamatrixát az alapvető faktorok módszerével, s ennek alapján a (2.1.3) alapján állítsuk elő az első alapvető faktort. Tegyük fel, hogy ez a eredeti változók

$$K_1 = \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_n X_n \tag{2.3.1}$$

lineáris kombinációja.

Ily módon eljárva a K_1 az adott csoportképző ismérv olyan *komplex* mutatójának tekinthető, ami magába sűríti az adott csoportképző ismérve ható tényezők által tartalmazott információ jelentős részét. A gyakorlati tapasztalatok ugyanis azt mutatják, hogy az első alapvető faktor a vizsgált változók szórásnégyzetének 60—80%-át „megmagyarázza", s a többi alapvető faktor csak jelentéktelen mértékben járul hozzá azokhoz. A (2.3.1)-ben szereplő α_j skalárok olyan „értékelési rendszernek", pontszámrendszernek tekinthetők, amelyek azt mutatják, hogy az egyes X_j változók milyen szerepet töltenek be, milyen súllyal vesznek részt a kiválasztott, közvetlenül nem mérhető csoportképző ismérvvel jellemezendő jelenség kialakításában.

Ilyen jellegű gyakorlati alkalmazás pl. R. S. Thorn [12] cikkében található.

Ezután a sokaság minden egyes egységére nézve meghatározzuk az első alapvető faktor K_{1i} értékét a

$$\mathbf{k}_1^{s} = [K_{11}, K_{12}, \ldots, K_{1N}]^* = \mathbf{X}\alpha \tag{2.3.2}$$

módon, ahol $\alpha = [\alpha_1, \alpha_1, \ldots, \alpha_n]^*$ a (2.3.1)-ben szereplő α_j skalárokból felépülő n-elemű oszlopvektor, $\mathbf{X}$ pedig a vizsgált változókra vonatkozó megfigyeléseket tartalmazó $N \times n$ méretű matrix. Ezen értékek alapján pedig elvégezhető a sokaság egységeinek rangsorolása, majd ennek felhasználásával viszonylag homogén csoportokba történő besorolása. Itt tehát a faktoranalízis annyiban segít, hogy — eltekintve magának a módszernek az önkényességétől — ami azonban a matematikai statisztika szinte minden eszközével kapcsolatban elmondható — kiküszöböli az ilyen problémák megoldására használt módszerek nagyfokú önkényességét.

A csoportosítási feladatok megoldására azonban akkor is szükség lehet a faktoranalízis alkalmazására, amikor azt tűzzük ki célul, hogy a csoportosítás elvégzésekor minden egyes csoportképző ismérvet egyenlő súllyal vegyünk figyelembe. Ilyen esetekben a kitűzött feladat a faktoranalízis alkalmazása nélkül a legtöbbször meg sem oldható. A csoportképző ismérvnek kijelölt változók ugyanis csak a legritkább esetben függetlenek egymástól. Ez a tény pedig lehetetlenné teszi a csoportképző ismérvek egyenlő súllyal történő figyelembevételét. Ha például négy olyan változó alapján kívánjuk elvégezni a csoportosítást, melyek közül kettő szoros sztochasztikus kapcsolatban áll egymással, akkor a négy változót a csoportosítás során egyenlő súllyal figyelembe véve kettőt majdnem dupla súllyal szerepeltetnénk. Ilyen esetekben ezért sokkal célszerűbb a változók helyett azok közös faktorait figyelembe venni a csoportosítás során.

2.4 *Az index-problémáról*

Ebban az alpontban a faktoranalízis ár- és volumenindexek meghatározására történő alkalmazását ismertetem röviden. Ezen alkalmazás elméleti kidolgozása *H. Theil* nevéhez fűződik, első gyakorlati alkalmazói pedig *T. Kloek* és *G. M. de Wit* voltak, akik cikkükben tovább is fejlesztik a H. Theil által kidolgozott módszert ([8] és [13]).

Az ár- és volumenindex-számítás feladata az, hogy több különnemű, s így közvetlenül nem összesíthető termék (árucikk) egységárának, illetve termelt (eladott stb.) mennyiségnek együttes átlagos időbeni változását vagy térbeni különbözőségét mutassa ki. Az egyszerűség kedvéért a továbbiakban csak az

időbeni összehasonlítással foglalkozom, de az ezután következő gondolatmenet — értelemszerű módosításokkal — térbeni összehasonlításra is alkalmazható.

Induljunk ki n számú termék (árucikk) t számú időszakra vonatkozó egységáraiból és termelt (eladott stb.) mennyiségeiből. E kiinduló adatokat a könnyebb áttekinthetőség érdekében célszerű egy-egy $t \times n$ típusú ár-, illetve volumenmatrixba foglalni a

$$\mathbf{P} = [p_{ij}] \text{ és } \mathbf{Q} = [q_{ij}]$$

módon, ahol p_{ij} a j-edik termék (árucikk) i-edik időszakra vonatkozó egységára, q_{ij} pedig a j-edik termék (árucikk) i-edik időszakban termelt (eladott stb.) mennyisége. A $\mathbf{P}$ és $\mathbf{Q}$ matrixok segítségével igen egyszerűen előállítható az indexszámítás alapját képező aggregátumok

$$\mathbf{V} = \mathbf{P}\mathbf{Q}^* = [v_{ij}] \tag{2.4.1}$$

t-edrendű matrixa, amelynek v_{ij} eleme a j-edik időszak i-edik időszaki egységárak alapján meghatározott termelési értéke (eladási forgalma stb.). A $\mathbf{V}$ matrix elemeinek felírásakor az egyszerűség kedvéért eltekintettünk az egyes termékekre (árucikkekre) utaló futóindex kiírásától.

Tekintsük ezután feladatunknak egy olyan $\mathbf{p}$ árvektor és egy olyan $\mathbf{q}$ volumenvektor meghatározását, amelyek segítségével az agregátumok (2.4.1) matrixa a legkisebb négyzetek módszere értelmében a lehető legjobban megközelíthető, reprodukálható, azaz amelyekre nézve a

$$\mathbf{D} = \mathbf{V} - \mathbf{p}\,\mathbf{q}^* \tag{2.4.2}$$

ún. eltérésmatrix elemeinek négyzetösszege minimális. Ez másképpen megfogalmazva annyit jelent, hogy minden $\Sigma p_i q_j$ alakú agregátumot egy az i-edik időszakra jellemző „átlagár" ($\bar{p}_i$) és egy a j-edik időszakra jellemző „átlagos mennyiség"($\bar{q}_j$) szorzatával kívánunk közelíteni. Az eltérésmatrix elemeinek négyzetösszege a legegyszerűbben a

$$tr\,(\mathbf{D}\,\mathbf{D}^*) = tr\,(\mathbf{V}\,\mathbf{V}^*) - 2\,\mathbf{p}^*\,\mathbf{V}\,\mathbf{q} + (\mathbf{p}^*\,\mathbf{p})\,(\mathbf{q}^*\,\mathbf{q}) \tag{2.4.3}$$

módon írható fel. A szélsőértékszámítás szokásos módszerét alkalmazva, s az így adódó egyenleteket kissé átalakítva arra az eredményre jutunk, hogy a keresett

$$\mathbf{p} = [\bar{p}_1, \bar{p}_2, \ldots \bar{p}_t]^* \text{ és } \mathbf{q} = [\bar{q}_1, \bar{q}_2, \ldots \bar{q}_t]^*$$

vektorok a

$$\begin{cases} [\mathbf{V}\,\mathbf{V}^* - (\mathbf{p}^*\,\mathbf{p})\,(\mathbf{q}^*\,\mathbf{q})\,\mathbf{E}]\,\mathbf{p} = \mathbf{0} \\ [\mathbf{V}^*\,\mathbf{V} - (\mathbf{p}^*\,\mathbf{p})\,(\mathbf{q}^*\,\mathbf{q})\,\mathbf{E}]\,\mathbf{q} = \mathbf{0} \end{cases} \tag{2.4.4}$$

egyenletek megoldásaként adódnak. A (2.4.4) alatti egyenletek alapján belátható, hogy a (2.4.3) eltérés négyzetösszeget minimalizáló $\mathbf{p}$ árvektor, illetve $\mathbf{q}$ volumenvektor a $\mathbf{V}\mathbf{V}^*$, illetve $\mathbf{V}^*\mathbf{V}$ matrix maximális,

$$\lambda^2 = (\mathbf{p}^*\,\mathbf{p})\,(\mathbf{q}^*\,\mathbf{q})$$

sajátértékéhez tartozó sajátvektora. Ha még azt is kikötjük, hogy mind a $\mathbf{p}$, mind a $\mathbf{q}$ vektor $\sqrt{\lambda}$ hosszúságú legyen, a megoldás egyértelművé válik.

Az itt követett módszer technikailag igen hasonlít az alapvető faktorok módszeréhez, bár nem teljesen azonos azzal. A legszembetűnőbb eltérés a két módszer között az, hogy itt nem a korrelációs matrixból kiindulva végezzük el a számításokat.[15] Ennek ellenére Kloek és de Wit már idézett cikkükben a **p** árvektort a (2.4.1) matrix oszlopai első alapvető faktorának, a **q** volumenvektort pedig a (2.4.1) matrix sorai első alapvető faktorának nevezik.

Az így kapott **p** és **q** vektorok elemei az egyes időszakokra jellemző „átlagárak", illetve „átlagos" mennyiségek konstansszorosai, amelyek még nem értelmezhetők közvetlenül indexekként. Ha az i-edik időszakot tekintjük bázisidőszaknak, akkor maguk az indexek az

$$\text{(2.4.5)} \qquad \mathbf{i}_p = \frac{1}{\mathbf{e}_i^* \mathbf{p}} \mathbf{p} \quad \text{illetve} \quad \mathbf{i}_q = \frac{1}{\mathbf{e}_i^* \mathbf{q}} \mathbf{q}$$

módon állnak elő, ahol $\mathbf{e}_i^*$ az i-edik t-elemű egységvektort jelenti sorvektorként felírva.

Két időszak esetén megmutatható, hogy az ár (P)- és volumenindex (Q) az alábbi közelítő formulák segítségével határozható meg:

$$\text{(2.4.6)} \qquad P \simeq P_0 \left(1 + \eta \frac{Q_0^2}{1 + Q_0^2}\right), \text{ ill. } Q \simeq Q_0 \left(1 + \eta \frac{P_0^2}{1 + P_0^2}\right),$$

ahol P_o, illetve Q_o a Laspeyres-féle ár-, illetve volumenindex, P_1 illetve Q_1 pedig a Paasche-féle ár-, illetve volumenindex, és

$$\eta = \frac{P_1}{P_0} - 1 = \frac{Q_1}{Q_0} - 1$$

Ami az ily módon meghatározott indexek közgazdasági tartalmát illeti, meg kell jegyeznünk, hogy kettőnél több időszak esetén igen nehéz annak megítélése. Ez részben matematikai korlátokba ütközik, részben pedig a kapott eredmények nehéz értelmezhetőségén alapszik. Általánosságban csak annyit szögezhetünk le, hogy az a tény, hogy a fenti indexek egyes időszakokra vonatkozó „átlagárakon" illetve „átlagos mennyiségeken" alapulnak nagymértékben korlátozza azok alkalmazhatóságát. Egy adott időszakra vonatkozó átlagár, illetve átlagos mennyiség meghatározása ugyanis csak olyan esetekben indokolt közgazdaságilag, amikor a **P** illetve **Q** matrix egyes soraiban álló elemek összegezhetők. (Ez a helyzet például akkor, ha azonos fajta termék (árucikk) különböző minőségeiről van szó.) Azt is megállapíthatjuk, hogy a fenti indexek állandó súlyozásúak, azaz az említett átlagárakra, illetve átlagos mennyiségekre nézve

$$\mathbf{p} = \mathbf{P}\,\alpha \quad \text{illetve} \quad \mathbf{q} = \mathbf{Q}\,\beta$$

érvényes. Ehhez mindjárt hozzá kell tennünk azt is, hogy ez az „állandó" súlyozás bizonyos értelemben változó is. Ugyanis a fenti indexek legfőbb pozitívuma az, hogy mind az ár-, mind a volumenindex-számítás e módszere az összes vizsgált időszak mennyiségi- és áradatát figyelembe veszi. Ez pedig azt

[15] Itt jegyezzük meg, hogy a faktoranalízis nem szükségképpen a korrelációs matrixból indul ki. Egyes esetekben a korrelációs matrix helyett az ún. variancia-kovariancia matrix képezi a faktoranalízis alapját.

jelenti, hogy a vizsgált időszakot akár egy időszakkal kibővítve megváltozik az egész indexsor, megváltozik az indexsor súlyrendszere. Ez pedig bizonyos szempontból változó súlyozást is jelent.

Két időszak esetére az eddig elmondottakon kívül még az is igaz, hogy mind az ár, mind a volumenindex a megfelelő Laspeyres-féle indexek körül ingadozik, ami egyúttal arra is rávilágít, hogy a Laspeyres-féle ár-, illetve volumenindexek bizonyos aszimptotikusan optimális tulajdonsággal is rendelkeznek.

Ezzel a faktoranalízis közgazdasági alkalmazásának legfőbb lehetőségeit — ha nagy vonalakban is — áttekintettük. Az eddig elmondottakból látható, hogy a faktoranalízis igen érdekes, hasznos, bár korántsem problémamentes módszer. Véleményem szerint a felsorolt alkalmazási lehetőségek közül a 2.2 és 2.3 alpontban ismertetettek a legérdekesebbek és legizgalmasabbak, mert azok a közgazdasági kutatómunka igen hatékony segédeszközeivé válhatnak.

IRODALOM

[1] Bellman, R.: Introduction to Matrix Analysis. London—New York—Toronto, 1960. Mc Graw-Hill Book Company, 328 p.

[2] Berry, J. L. B.: An Inductive approach to the Regionalization of Economic Development. Geography and Economic Development (Edited by Ginsburg, N) 78—107. p.

[3] Farrar, D. E.—Glauber, R. R.: Multicollinearity in Regression Analysis. The Review of Economics and Statistics, Vol. 49. (1967). No. 1, 92—107. p.

[4] Graybill, F. A.: An Introduction to Linear Statistical Models, Vol, I. New York, 1961. Mc Graw-Hill Book Company, Inc. 463 p.

[5] Hajós Gy.: Bevezetés a geometriába. Budapest, 1960. Tankönyvkiadó, 594 p.

[6] Harman, H. H.: Modern Factor Analysis. The University of Chicago Press, 1960. 469 p.

[7] Jánossy, F.: A gazdasági fejlettség mérhetősége és új mérési módszere. Budapest, 1963. Közgazdasági és Jogi Könyvkiadó, 323 p.

[8] Kloek, T.—Wit, G. M. de: Best Linear and Best Linear Unbiased Index Numbers. Econometrica, Vol, 29 (1961). No. 4, 602—616. p.

[9] Köves, P.—Párniczky, G.: Általános statisztika (egyetemi jegyzet). Budapest, 1967. Tankönyvkiadó, 571 p.

[10] Krekó, B.: Matrixszámítás. Budapest, 1964. Közgazdasági és Jogi Könyvkiadó, 374 p.

[11] Richter, P.: Anwendungen der Faktorenanalyse auf ökonomische Daten, Allgemeines Statistisches Archiv. Bd. 52 (1968), Nr. 2., 125—152. p.

[12] Schäffer, K.-A.: Faktorenanalyse und ihre Anwendungsmöglichkeiten, Allgemeines Statistisches Archiv. Bd 53 (1969). Nr. 1., 51—72. p.

[13] Theil, H.: Best Linear Index Numbers of Prices and Quantities, Econometrica, Vol. 28 (1960), No. 4., 464—480. p.

[14] Thorn, R. S.: Per Capita Income as a Measure of Economic Development. Zeitschrift für Nationalökonomie, Bd. 28 (1968), Heft 2., 206—216. p.

[15] Yule, G. U.—Kendall, M. G.: Bevezetés a statisztika elméletébe. Budapest, 1964. Közgazdasági és Jogi Könyvkiadó, 697. p.

[16] Lawley, D, N. — Maxwell, A. E.: Factor Analysis as a Statistical Method. London 1963. Butterworth and Co. Ltd., 117 p.