Sajátértékek a statisztikában

Dr. Hajdu Ottó,
a Budapesti Corvinus Egyetem
Statisztikai Tanszékének
tanszékvezetője
E-mail: hajduotto@uni-corvinus.hu

A tanulmány a statisztikai kapcsolatok mérési skála által meghatározott típusai – variancia, korreláció, asszociáció, látencia – mérésének sokváltozós mérőszámait tekinti át az elemzendő mátrixok sajátértékeinek tükrében. Kiemelkedően fontos alkalmazási területekre koncentrál.

TÁRGYSZÓ:
Statisztikai módszertan.
Korrelációszámítás.
Mátrixelmélet.

A többváltozós statisztikai kapcsolatok mérése nevezetes mátrixok sajátértékeinek meghatározására vezet. A kapcsolat jellemzése – jellegétől függetlenül – alapvetően a szóródás egy-, illetve kétváltozós mérésén alapul. Kézenfekvő a több változót egybesűríteni, vagy a kapcsolatot minden párosításban vizsgálni. E célt szolgálja a *szóródási mátrix*, összekapcsolva a kétféle megközelítést. A kapcsolat jellegétől függően – korreláció, diszkriminancia, asszociáció – a szóródási mátrix nevezetes formákat ölt, melyek sajátértékei nyújtják a megfelelő szóródási, illetve kapcsolatvizsgálati mértékeket. A tanulmány áttekinti az egyes kapcsolatok vonatkozó szóródási mátrixait és azok sajátértékeinek statisztikai tartalmát.

Lévén a többváltozós elemzések alapvető eszköze az ún. *szinguláris érték* felbontás, kiindulásként e módszert ismertetjük. Ezt követően tárgyaljuk a *variancia* tömörítését, majd a *korreláció–diszkriminancia–asszociáció* hármas többdimenziós kiterjesztését, végül a kapcsolatok mögött húzódó *latens változók* kérdését. A sajátérték-feladat és az egyes kapcsolattípusok többváltozós módszertani alapjainak ismeretét feltételezzük.

1. Az SVD-eljárás

Statisztikai változók komponensekre bontásának alapvető módja az *Eckart–Young-féle szinguláris érték felbontás* (SVD-eljárás) mely szerint bármely valós (*n,p*) rendű **X** mátrix felírható az alábbi multiplikatív formában:[1]

$$\mathbf{X} = \mathbf{FDV}^T, \qquad /1/$$

ahol **X** a *p* változókra végzett *n* megfigyelés értékeit tartalmazza, az ugyancsak *(n,p)* rendű **F** oszlopai az **X** bal oldali, a (*p,p*) rendű **V** mátrix oszlopai pedig az **X** jobb oldali szinguláris vektorait adják. A $\mathbf{D} = \langle \mu_1, \mu_2, ..., \mu_p \rangle$ diagonális mátrix diagonális elemei **X** (megfelelő) ún. szinguláris értékei. Másképpen fogalmazva **V** oszlopai a *p* dimenziós tér főtengelyeinek a bázisát, **F** oszlopai pedig a főtengelyekre vonatkozó koordinátákat jelentik.

[1] Singular Value Decomposition. A képletben szereplő „*T*" felső index transzponálást jelent.

Részletesebben felírva a modellt:

$$\mathbf{X} = \begin{bmatrix} \mathbf{f}_1 & \mathbf{f}_2 & \dots & \mathbf{f}_p \end{bmatrix} \begin{bmatrix} \mu_1 & & & \\ & \mu_2 & & \\ & & \ddots & \\ & & & \mu_p \end{bmatrix} \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \dots & \mathbf{v}_p \end{bmatrix}^T.$$

Az SVD-feladat az $\mathbf{F}^T\mathbf{F} = \mathbf{I}$ és a $\mathbf{V}^T\mathbf{V}=\mathbf{I}$ ortonormáltsági feltételek mellett (ahol **I** a megfelelő rendű egységmátrixot jelöli) a (*p,p*) rendű $\Sigma = \mathbf{X}^T\mathbf{X}$ szóródási mátrix spektrális felbontásával oldandó meg, mivel a szóródási mátrix az SVD-szabály alkalmazásával az

$$\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{D}^2\mathbf{V}^T$$

sajátérték-sajátvektor feladatra vezet. Ekkor a szóródási mátrix $\mu_1^2 \geq \mu_2^2 \geq \dots \geq \mu_p^2$ sajátértékei a négyzetes szinguláris értékeket adják, miközben **V** oszlopai a megfelelő sajátvektorok. A szóródási mátrix főátló elemei a változónkénti szóródás, összegük pedig a totális szóródás mértéke. A saját értékek összege a spektrális felbontásból következően a totális szóródási mértékkel azonos:[2] $tr\left(\mathbf{X}^T\mathbf{X}\right) = tr\left(\mathbf{D}^2\right)$. Ezen összegen belül a rendre csökkenő sajátértékek feltételesen maximáltak. A szóródási mátrix pozitív (szemi-)definit, tehát minden sajátértéke nemnegatív, de empirikus adatokon alkalmazva gyakorlatilag szigorúan pozitív definit.

2. A variancia tömörítése

Közvetlenül megfigyelhető, *manifest* jellegű x_j (j=1,2,...,p) változók helyettesítését, illetve tömörítését főkomponensek szolgálják, melyek magukból a változókból képzett k_t (t=1,2,...,p) lineáris kombinációk, páronként korrelálatlan rendszert alkotva, és a manifest változókat maradék nélkül reprodukálják:

$$k_t = v_{1t}x_1 + v_{2t}x_2 + \dots + v_{jt}x_j + \dots + v_{pt}x_p, \qquad /2/$$

ahol

$$x_j = v_{j1}k_1 + v_{j2}k_2 + \dots + v_{jt}k_t + \dots + v_{jp}k_p. \qquad /3/$$

[2] tr(.) a mátrix nyomát jelenti, mely a főátló elemek összege.

A súlyok dupla alsó indexében az első (*j*) index az *x* változóra, a második (*t*) pedig a *k* főkomponensre utal. A v_{jt} súlyokat a **V** mátrixba foglalva, annak *t*. oszlopa az *x* változók súlyozására szolgál a k_t főkomponens számítása érdekében, *j*. sora pedig a *k* főkomponensek súlyozására az x_j változó kalkulálása céljából.

A feladat a manifest változók olyan *k* lineáris kombinációit megadni, melyek az *x* változók totális szóródásához rendre maximált hányadban járulnak hozzá.

A megoldás az SVD-**F** főkomponensek meghatározásával kezdődően:

$$\mathbf{F} = \mathbf{XVD}^{-1}, \qquad /4/$$

melyből átskálázással

$$\mathbf{K} = \mathbf{FD}. \qquad /5/$$

A skálázott *k* főkomponensek szóródási mátrixa:

$$\mathbf{K}^T\mathbf{K} = \mathbf{D}^T\left(\underbrace{\mathbf{F}^T\mathbf{F}}_{\mathbf{I}}\right)\mathbf{D} = \mathbf{D}^2. \qquad /6/$$

Lévén a változók szóródását a szóródási mátrix főátló elemei mérik, valamely főkomponens szóródásának mértékét a manifest változók szóródási mátrixának megfelelő sajátértékei adják.

Ekkor, ha az **X** változók szóródási mátrixa:

1. a **C** kovarianciamátrix, a főkomponens varianciája a kovarianciamátrix megfelelő sajátértéke:

$$Var\left(k_t\right) = \mu_t^{2(\mathbf{C})} \qquad (t = 1,2,...,p), \qquad /7/$$

2. az **R** korrelációs mátrix, a főkomponens varianciája a korrelációs mátrix megfelelő sajátértéke:

$$Var\left(k_t\right) = \mu_t^{2(\mathbf{R})} \qquad (t = 1,2,...,p). \qquad /8/$$

Ha a főkomponenseket az SVD-modellben transzformáljuk (rotáljuk) a (*p*,*p*) rendű **T** transzformációs mátrix alapján ($\mathbf{TT}^{-1}=\mathbf{I}$) akkor elfordulnak a főkomponensek a $\mathbf{K}^*=\mathbf{KT}=\mathbf{FDT}$ módon, és így a szóródási mátrix:

$$\mathbf{K}^{*T}\mathbf{K}^* = \mathbf{T}^T\mathbf{D}^T\left(\underbrace{\mathbf{F}^T\mathbf{F}}_{\mathbf{I}}\right)\mathbf{DT} \neq \mathbf{D}^2, \qquad /9/$$

tehát a manifest szóródási mátrix sajátértékei többé nem varianciatartalmúak.

3. Kategóriák diszkriminálása

A szóródás mérésének egyik feladata a g=1,2,...,m számú csoportokra bontott sokaság szóródásának többdimenziós mérése, tekintettel a csoporttagságokra is. Ekkor a szóródás kétféle hatás eredője: a csoportközi különbségeket jellemző külső és a csoporton belüli eltérésekben jelentkező belső szóródásé.

Célunk elhatárolni a totális szóródásban a külső és a belső faktoroknak tulajdonított hányadot. A megoldás alapja a kovariancia (mátrix) csoportközi felbontása:

$$\mathbf{C} = \mathbf{C}_K + \mathbf{C}_B, \qquad /10/$$

ahol $\mathbf{C}_K$ a csoportátlagokkal helyettesített sokaság kovarianciamátrixa, $\mathbf{C}_B$ pedig a súlyozott, átlagos csoporton belüli kovarianciamátrix.

A csoporton belüli homogenitás, illetve a csoportközi heterogenitás jellemzésére a Wilks-féle lambda mutatót használjuk, mely a belső általánosított varianciának a teljes általánosított varianciához való arányát fejezi ki:[3]

$$\Lambda = \frac{\det(\mathbf{C}_B)}{\det(\mathbf{C})}. \qquad /11/$$

Minél alacsonyabb ez a hányad, annál homogénebbek a csoportok, és annál inkább a csoportközi szóródás dominál a sokaság totális szóródásában.

A varianciahányados jellegű Wilks-lambda egyváltozós esetben a belső és a teljes variancia hányadosává egyszerűsödik. Többváltozós esetben kézenfekvő a külső és belső szóródás vizsgálatát visszavezetni egyváltozós esetre, a megfigyelt változók

$$z = b_1 x_1 + b_2 x_2 + ... + b_p x_p$$

lineáris kombinációját, a diszkriminanciaváltozót képezve, alkalmasan megválasztott b súlyok alkalmazásával. Ennek belső és külső varianciája:

$$Var(z) = Var_B(z) + Var_K(z),$$

mely kvadratikus formában (a b súlyokat a $\mathbf{b}$ vektorba foglalva):

$$Var(z) = \mathbf{b}^T \mathbf{C} \mathbf{b} = \mathbf{b}^T \left(\mathbf{C}_B + \mathbf{C}_K \right) \mathbf{b} = \mathbf{b}^T \mathbf{C}_B \mathbf{b} + \mathbf{b}^T \mathbf{C}_K \mathbf{b}. \qquad /12/$$

[3] A p-dimenziós tér általánosított varianciája a tér kovarianciamátrixának a determinánsa.

A diszkriminanciaváltozó egyváltozós Wilks-lambdája, illetve komplementere egységnyi belső varianciához normálva:

$$1-\Lambda(z)=\frac{Var_K(z)}{Var_B(z)+Var_K(z)}=\frac{Var_K(z)/Var_B(z)}{1+Var_K(z)/Var_B(z)}=\frac{\varphi}{1+\varphi}. \qquad /13/$$

Most a külső varianciát a belső varianciához viszonyító, értelemszerűen maximálandó diszkriminanciakritérium:

$$\varphi=\frac{Var_K(z)}{Var_B(z)}=\frac{\mathbf{b}^T\mathbf{C}_K\mathbf{b}}{\mathbf{b}^T\mathbf{C}_B\mathbf{b}}\rightarrow \max. \qquad /14/$$

A φ diszkriminanciakritérium **b** szerinti maximálása a

$$\frac{\partial\varphi}{\partial\mathbf{b}}=\frac{2\mathbf{C}_K\mathbf{b}\left(\mathbf{b}^T\mathbf{C}_B\mathbf{b}\right)-\left(\mathbf{b}^T\mathbf{C}_K\mathbf{b}\right)2\mathbf{C}_B\mathbf{b}}{\left(\mathbf{b}^T\mathbf{C}_B\mathbf{b}\right)^2}=\mathbf{0}$$

egyenlet megoldását igényli, mely a $\mathbf{b}^T\mathbf{C}_B\mathbf{b}$ skalárral való egyszerűsítés és keresztbeszorzás, majd φ /14/ definíciójának behelyettesítése után megfelelő átrendezéssel a

$$\left(\mathbf{C}_B^{-1}\mathbf{C}_K-\varphi\ \mathbf{I}\right)\mathbf{b}=\mathbf{0} \qquad /15/$$

sajátérték-sajátvektor feladatra vezet. Ez a

$$\left(\mathbf{C}_K-\varphi(\mathbf{C}-\mathbf{C}_K)\right)\mathbf{b}=\left((1+\varphi)\mathbf{C}_K-\varphi\mathbf{C}\right)\mathbf{b}=\mathbf{0}$$

átalakítással a

$$\left(\mathbf{C}^{-1}\mathbf{C}_K-\frac{\varphi}{1+\varphi}\mathbf{I}\right)\mathbf{b}=\mathbf{0}$$

sajátérték-sajátvektor feladat formában is megoldható. A súlyokat tartalmazó **b** sajátvektor mindkét feladatra közös.

A $\mathbf{C}^{-1}\mathbf{C}_K$ mátrixnak $min\{p,(m-1)\}=k$ számú pozitív sajátértéke van, melyek statisztikai tartalmuk szerint rendre egyváltozós Wilks-lambdák.

A $\mathbf{C}_B^{-1}\mathbf{C}_K$ nem szimmetrikus mátrix sajátértékei pedig statisztikai tartalmuk szerint rendre maximált diszkriminanciakritériumok.

Végül a több- és az egyváltozós Wilks-lambdák közötti kapcsolat:

$$\Lambda = \det(\mathbf{C}^{-1})\det(\mathbf{C}_B) = \det(\mathbf{C}^{-1}\mathbf{C}_B) = \det\left(\mathbf{C}^{-1}(\mathbf{C}-\mathbf{C}_K)\right) = \det(\mathbf{I}-\mathbf{C}^{-1}\mathbf{C}_K) = \quad /16/$$

$$= \prod_{j=1}^{k}\left(1-\frac{\varphi_j}{1+\varphi_j}\right) = \prod_{j=1}^{k}\left(\frac{1}{1+\varphi_j}\right). \quad /17/$$

4. Kanonikus korrelációk számítása

Többváltozós esetben a kétváltozós korreláció mérése kiterjeszthető két változócsoport közötti korreláció vizsgálatára, ha mindkét változócsoportot egy-egy lineáris kombinációval helyettesítjük. Tekintsük a standardizált változók $x_1,x_2,...,x_p$ magyarázó, és a velük oksági kapcsolatban lévő, eredmény jellegű, ugyancsak standardizált változók $y_1,y_2,...,y_q$ ($q \le p$) csoportját.

Képezzük az x magyarázóváltozók lineáris kombinációjaként az u, és az y eredményváltozók csoportjából a z lineáris kombinációk t=1,2,...,q párosait:

$$u_t = v_{1t}x_1 + v_{2t}x_2 + ... + v_{pt}x_p$$

$$z_t = w_{1t}y_1 + w_{2t}y_2 + ... + w_{qt}y_q\,,$$

ahol valamennyi változó standardizált, és $q \le p$. A v és w súlyokat úgy határozzuk meg, hogy az u_t és z_t kanonikus változók közötti lineáris korreláció maximált legyen, miközben a kanonikus változók bármilyen más párosításban korrelálatlanok. E követelményeket fogalmazza meg a *kanonikus változók korrelációs mátrixa* az alábbi partícionált formában:

$$\mathbf{R}_{uz} = \begin{array}{c|cccc|cccc} & u_1 & \cdots & u_q & & z_1 & \cdots & z_q \\ \hline u_1 & 1 & & 0 & & r_1 & & 0 \\ \vdots & & & & & & & \\ u_q & 0 & & 1 & & 0 & & r_q \\ \hline z_1 & r_1 & & 0 & & 1 & & 0 \\ \vdots & & & & & & & \\ z_q & 0 & & r_q & & 0 & & 1 \end{array}.$$

E korrelálatlansági feltételek mellett maximált $Cov(u_t,z_t)=r_t$ lineáris korrelációt a t. kanonikus korrelációnak, az (u_t,z_t) változópárost pedig a t. kanonikus változópárnak nevezzük.

A kanonikus korrelációk meghatározása érdekében particionáljuk a manifest változók ($q+p,q+p$) rendű korrelációs mátrixát az alábbiak szerint:

$$\mathbf{R} = \left[\begin{array}{c|c} \mathbf{R}_{yy} & \mathbf{R}_{yx} \\ \hline \mathbf{R}_{xy} & \mathbf{R}_{xx} \end{array}\right],$$

ahol az egyes mátrixok méretét az indexben szereplő változók számossága adja: például $\mathbf{R}_{yx}$ (q,p) rendű, vagyis nem négyzetes. Feladatunk az

$$r_{u,z} = r = \mathbf{v}^T\mathbf{R}_{xy}\mathbf{w} \rightarrow \max$$

korreláció maximálása a **v** és **w** súlyvektorok tekintetében, a

$$Var(u) = \mathbf{v}^T\mathbf{R}_{xx}\mathbf{v} = 1, \qquad Var(z) = \mathbf{w}^T\mathbf{R}_{yy}\mathbf{w} = 1$$

standardizáltsági megszorítások mellett. A Lagrange-féle multiplikátor-módszert alkalmazva, a keresett kanonikus korrelációt és a megfelelő súlyokat az

$$\mathbf{R}_{xy}\mathbf{w} = r\mathbf{R}_{xx}\mathbf{v}, \qquad \mathbf{R}_{yx}\mathbf{v} = r\mathbf{R}_{yy}\mathbf{w} \qquad /18/$$

egyenletrendszer megoldása szolgáltatja. Az első egyenletből kifejezve a **v** vektort, majd ezt a második egyenletbe helyettesítve, és végül az utóbbit átrendezve, az

$$\left(\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy} - r^2\mathbf{I}\right)\ \mathbf{w} = \mathbf{0}$$

sajátérték-sajátvektor feladatra jutunk, ahol a (q,q) rendű $\mathbf{R}_{yy}^{-1}\mathbf{R}_{yy}\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}$ mátrix sajátértékei a kanonikus korrelációk négyzeteit, a megfelelő sajátvektorok pedig az y (szűkebb körű) változókhoz tartozó súlyrendszereket nyújtják. A **w** súlyok ismeretében /18/ bármely egyenletéből a **v** súlyok is következnek.

5. Korrespondenciák feltárása

Jellegét tekintve az asszociáció a kategóriaskálán mért változók kimenetei közötti kapcsolat. Exploratív elemzési eszközeinek általános kerete a korrespondencia-

analízis (CA), mely a nagyméretű kontingenciatábla adatait hivatott áttekinthetővé tenni. Mivel itt a kapcsolatrendszer struktúrája szempontjából az egyes kategóriák előfordulásának nem az abszolút, hanem a relatív gyakorisága érdekes, a CA induló adatállományát – valamennyi empirikus f_{ij} gyakoriságot a gyakoriságok n összegével (a megfigyelések számával) osztva – a kontingenciatábla normált változata, az ún. korrespondenciamátrix alkotja. Ennek általános eleme $p_{ij} = f_{ij} / n$, az i sor és a j oszlop együttes bekövetkezésének relatív gyakorisága.

1. táblázat

Korrespondenciatábla

Kategória	Oszlop					*Sorösszesen*
	1.	...	*j.*	...	*J.*	
Sor *1.*	p_{11}		p_{1j}		p_{1J}	s_1
Sor *i.*	p_{i1}		$p_{ij} = f_{ij} / n$		p_{iJ}	s_i
Sor *I.*	p_{I1}		p_{Ij}		p_{IJ}	s_I
Oszlopösszesen	o_1		o_j		o_J	1

A sorok s_i és az oszlopok o_j összesen adatai peremgyakoriságként értelmezendők. A tábla sorainak, illetve oszlopainak belső szerkezeteit összehasonlítva a peremmel hozzuk egymással kapcsolatba azon (*i,j*) kategóriapárosításokat, melyek a sorok és az oszlopok szóródásához, illetve a közöttük lévő asszociációhoz a leginkább hozzájárulnak. Az egymást vonzó, illetve taszító (*i,j*) kategóriapárosítást a peremszerkezet alapján vártnál kiugróan magasabb vagy alacsonyabb p_{ij} gyakoriság jelzi.[4]

Matematikailag a korrespondenciaanalízis az asszociáció Pearson-féle χ^2 mértékét bontja komponensekre hasonló módon, mint azt a főkomponens-analízis a varianciával teszi. Az eljárás a sorokat (oszlopokat) a megoszlásaikból képzett, redukált dimenziójú, mesterséges térbe helyezi. Itt a tengelyeket úgy definiáljuk, hogy rendre csökkenő százalékos mértékben (sorrendben) járuljanak hozzá a χ^2 statisztikához.

A korrespondenciatábla kategóriái közötti asszociáció mértékét jellemző, egységnyi megfigyelésre jutó Pearson-féle χ^2 érték definíció szerint:[5]

[4] Az 1. táblázat „összesen" sorában és oszlopában foglalt relatív peremgyakoriságok szerkezete alapján várható gyakoriság: $p^*_{ij} = s_i \cdot o_j$.

[5] E tanulmányban *Pearson*-χ^2 alatt mindig az egységnyi megfigyelésre normált χ^2 értéket értjük.

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(p_{ij} - s_i o_j)^2}{s_i o_j} = \sum_{i=1}^{I}\sum_{j=1}^{J} g_{ij}^2,$$

ahol $s_i o_j$ az (i,j) cellának a peremmegoszlások alapján várt relatív gyakorisága az asszociáció teljes hiánya esetén. Ebből következően a

$$g_{ij} = \frac{p_{ij} - s_i o_j}{\sqrt{s_i o_j}}$$

standardizált korrespondenciagyakoriság zéró értéke az asszociáció hiányát, pozitív értéke pozitív, negatív értéke pedig negatív asszociációt jelez az i sor és a j oszlop között. Pozitív asszociáció esetén az i és j kategóriák gyakran következnek be együtt, vagyis vonzzák egymást, negatív asszociáció esetén pedig ritkán járnak közösen, tehát taszítják egymást. Az előzők alapján g_{ij}^2 az (i,j) cellának, $\Sigma_j g_{ij}^2$ az i sornak, $\Sigma_i g_{ij}^2$ pedig a j oszlopnak a hozzájárulását adja a χ^2 mértékhez.

Az oszlop- és sorprofilok ábrázolása nemcsak két, hanem kettőnél több szempont (változó) szerint kategorizáló táblák esetén is lehetséges. Az i sor és a j oszlop közötti kapcsolat vizsgálatát egyszerű korrespondenciaanalízisnek nevezzük. Ebből a szempontból érdektelen, hogy adott sor (oszlop) esetleg több változó kategóriáinak valamely együttes kombinációját definiálja. Többszörös korrespondenciaanalízist végzünk viszont akkor, ha a vizsgált változók számát kettőnél többre bővítve, az asszociáció vizsgálatát az előforduló kategóriák valamennyi párosítására kiterjesztjük.

5.1. Egyszerű korrespondenciaanalízis

Az egyszerű korrespondenciaanalízis a gyakorisági tábla sorait egy pontfelhő pontjaiként tekinti az oszlopok terében, oszlopait pedig egy másik pontfelhő pontjaiként a sorok terében. A pontfelhőket egy redukált, alacsony dimenziójú térben ábrázoljuk, és a pontok helyzetéből következtetünk arra, hogy a vizsgált változók mely kategóriái vonzzák, illetve taszítják egymást. A redukált tér dimenziója $K \leq \min\{I-1, J-1\}$, a sorok CA-koordinátáit az **X**, az oszlopokét pedig az **Y** mátrixok tartalmazzák.

Az asszociáció feltárása érdekében vegyük a sorok (majd az oszlopok) origóperemhez centrált szerkezeteit – profiljait –, melyeket általános jelölésekkel a 2. és 3. táblázatokba foglaltunk, ahol s_{ij} a j oszlop centrált részesedése az i sorban, míg o_{ij} az i sor centrált részesedése a j oszlopban.

2. táblázat

Centrált sorprofilok és helyettesítő korrespondenciakoordinátáik

Sorprofil	Centrált profil: **S** mátrix					Sor CA-koordináta: **X**				
1.	s_{11}	...	s_{1j}	...	s_{1J}	x_{11}	...	x_{1k}	...	x_{1K}
i.	s_{i1}		s_{ij}		s_{iJ}	x_{i1}		x_{ik}		x_{iK}
I.	s_{I1}		s_{Ij}		s_{IJ}	x_{I1}		x_{Ik}		x_{IK}
Centroid*	0		0		0	0		0		0

* A sorok az origó körül szóródnak.
Megjegyzés. $s_{ij} = p_{ij} / s_i - o_j$.

3. táblázat

Centrált oszlopprofilok és helyettesítő korrespondenciakoordinátáik

Oszlopprofil	Centrált profil: **O** mátrix					Oszlop CA-koordináta: **Y**				
1.	o_{1I}	...	o_{1i}	...	o_{1I}	y_{1I}	...	y_{1k}	...	y_{1K}
j.	o_{j1}		o_{ji}		o_{jI}	y_{j1}		y_{jk}		y_{jK}
J.	o_{J1}		o_{Ji}		o_{JI}	y_{J1}		y_{Jk}		y_{JK}
Centroid*	0		0		0	0		0		0

* Az oszlopok az origó körül szóródnak.
Megjegyzés. $o_{ji} = p_{ij} / o_j - s_i$.

A CA-koordináták súlyozott centroidja az origó:

$$\sum_{i=1}^{I} s_i x_{ik} = 0, \quad \sum_{j=1}^{J} o_j y_{jk} = 0.$$

Most a χ^2 mérőszám az előző jelölésekkel a következő formában is megfogalmazható:

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{s_i}{o_j}(s_{ij})^2 = \sum_{j=1}^{J}\sum_{i=1}^{I} \frac{o_j}{s_i}(o_{ij})^2 = INR. \qquad /19/$$

Ebben a formában a χ^2 mutatót *inerciamértéknek* nevezzük, mely láthatóan *a pontfelhő súlyozott, többdimenziós varianciája* egyidejűleg mind a sorok, mind az oszlopok azonos mértékű szóródását jellemezve saját peremeik körül. A centrált CA-koordinátákat (**X,Y**) úgy definiáljuk, hogy adott pontnak a saját centroidtól vett távolsága, és így a teljes inercia értéke változatlan maradjon:

$$INR = \sum_{i=1}^{I} s_i \sum_{k=1}^{K} x_{ik}^2 = \sum_{j=1}^{J} o_j \sum_{k=1}^{K} y_{jk}^2. \qquad /20/$$

A CA-koordináták meghatározása érdekében definiáljuk a $\mathbf{D}_s=<s_1,...,s_I>$, $\mathbf{D}_o=<o_1,...,o_J>$, $\mathbf{D}_\mu=<\mu_1,...,\mu_K>$ diagonális mátrixokat és a g_{ij} standardizált korrespondenciagyakoriságokat tartalmazó $\mathbf{G}_{(I,J)}$ mátrixot. Ekkor a **G** mátrix SVD-felbontása az alapja a teljes inercia CA-tengelyek közötti szétosztásának:

$$\mathbf{G} = \mathbf{D}_s^{1/2}\mathbf{S}\mathbf{D}_o^{-1/2} = \mathbf{D}_s^{-1/2}\mathbf{O}\mathbf{D}_o^{1/2} = \mathbf{U}\mathbf{D}_\mu\mathbf{V}^T. \qquad /21/$$

Az **U** mátrix oszlopai adják **G** oszlopfelhőjének főtengelyeit, míg a **V** oszlopai **G** sorfelhőjének főtengelyeit. A keresett **X** és **Y** CA-koordináták a főtengelyekre vonatkozó megfelelő főkoordinátákból származnak.

Látható, hogy a $\mu_1,\mu_2,...,\mu_K$ szinguláris értékek négyzetei a $\mathbf{G}^T\mathbf{G}$ és a $\mathbf{G}\mathbf{G}^T$ szóródási mátrixok közös sajátértékei, és egyben a CA-tengelyek maximált varianciái. Ekkor a teljes inercia:

$$INR = \mathrm{tr}(\mathbf{G}^T\mathbf{G}) = \mathrm{tr}(\mathbf{G}\mathbf{G}^T) = \sum_{k=1}^{K} \mu_k^2. \qquad /22/$$

5.2. Többszörös korrespondenciaanalízis

Kettőnél több kategóriaváltozót elemezve, célszerű a korrespondenciaanalízis többszörös változatát alkalmazni. Ez ekvivalens az indikátormátrix egyszerű analízisével. A $\mathbf{Z}_{(n,J)}$ indikátormátrix sorait az i=1,2,...,n megfigyelések, míg oszlopait a Q számú Z_q (q=1,2,...,Q) kategóriaváltozók kategóriái képezik, ahol a Z_q változónak J_q számú lehetséges kategóriája van. Így a mátrix oszlopainak száma $J=J_1+J_2+...+J_Q$, és az oszlopok a Q számú csoport valamelyikének a tagjai. Az indikátormátrix mindegyik sora Q számú „1” elemet tartalmaz attól függően, hogy az illető megfigyelés adott változó melyik kategóriájához tartozik. Egyébként a mátrix elemei zérók.

4. táblázat

Indikátormátrix

Megfigyelés	A **Z** indikátor mátrix oszlopai *(j=1,2,…,J)*														*Összesen*
	Z_1 kategóriái: $\mathbf{Z}_1$				...	Z_q kategóriái: $\mathbf{Z}_q$				...	Z_Q kategóriái: $\mathbf{Z}_Q$				
	1	2	…	J_1	…	1	2	…	J_q	…	1	2	…	J_Q	
1	1						1							1	Q
2		1					1					1			Q
⋮															
i				1		1						1			Q
⋮															
n		1							1		1				Q
Összesen (f_j)	f_1^1	f_2^1	…	$f_{J_1}^1$	…	f_1^q	f_2^q	…	$f_{J_q}^q$	…	f_1^Q	f_2^Q	…	$f_{J_Q}^Q$	nQ

A **Z** mátrix tehát nQ egyest tartalmaz, n darabot minden egyes $\mathbf{Z}_q$ almátrixban, $\mathbf{Z}_q$ bármely sorának összege 1, és **Z** bármely sorának összege Q. A többszörös CA eredményeinek értelmezése az indikátormátrix alábbi tulajdonságain alapul:

1. A $\mathbf{Z}_q$ mátrix $o_j = f_j / (nQ)$ peremprofiljainak az összege bármely q=1,2,…,Q esetén: 1/Q. Így bármely változó egyforma relatív súlyt kap, melyet szétoszt az 1,2,…,J_q kategóriái között, az f^q gyakoriságoknak megfelelően.

2. Az $O_{ij} = (1/f_j) = 1/(n \cdot Q \cdot o_j)$ oszlopmegoszlások centroidja bármely $\mathbf{Z}_q$ blokkon belül egybeesik az oszlopprofilok globális centroidjával. Adott sor relatív gyakorisága $s_i = Q/(n \cdot Q) = 1/n$ és megoszlása: 1/Q.

3. A $\mathbf{Z}_q$ változó valamennyi oszlopához tartozó teljes inercia:

$$INR(q) = \sum_{j_q=1}^{J_q} INR(j_q) = \frac{J_q}{Q} - \frac{1}{Q}.$$

4. Az oszlopok (sorok) totális inerciája:

$$INR = \sum_{q=1}^{Q} INR(q) = \frac{J}{Q} - 1.$$

5. A pozitív inerciával bíró, nem triviális dimenziók száma legfeljebb J–Q.

6. Az n számú sorprofil mindegyike $J_1,J_2,\ldots,J_Q$ számú egymástól különböző pont valamelyikével esik egybe.

7. A $\mathbf{B}_{(J,J)}=\mathbf{Z}^T\,\mathbf{Z}$ *Burt*-mátrix analízisének standardizált korrespondenciakoordinátái azonosak a $\mathbf{Z}$ indikátormátrix analízisében az oszlopok standardizált korrespondenciakoordinátáival. A *Burt*-mátrix az alábbi blokkstruktúrában is írható:

$$\mathbf{Z}^T\mathbf{Z}=\mathbf{B}=\begin{bmatrix}\mathbf{Z}_1^T\mathbf{Z}_1 & \mathbf{Z}_1^T\mathbf{Z}_2 & \cdots & \mathbf{Z}_1^T\mathbf{Z}_Q\\ \mathbf{Z}_2^T\mathbf{Z}_1 & \mathbf{Z}_2^T\mathbf{Z}_2 & & \mathbf{Z}_2^T\mathbf{Z}_Q\\ \vdots & & \ddots & \\ \mathbf{Z}_Q^T\mathbf{Z}_1 & \mathbf{Z}_Q^T\mathbf{Z}_2 & & \mathbf{Z}_Q^T\mathbf{Z}_Q\end{bmatrix}.$$

Mindegyik $\mathbf{Z}_q^T\mathbf{Z}_{q^*}$ ($q\neq q^*$) mátrix, mely $\mathbf{B}$ diagonálisán kívül esik, egyben egy kétváltozós kontingenciatábla, mely a q és q^* változók közötti asszociációt sűríti az n számú megfigyelés alapján. Ugyanakkor a $\mathbf{B}$ diagonálisán mindegyik $\mathbf{Z}_q^T\mathbf{Z}_q$ mátrix diagonális, és diagonálisán $\mathbf{Z}_q$ oszlopösszesen értékei szerepelnek.

A Burt-mátrix oszlopainak és sorainak analízise azonos CA-koordinátákat eredményez. Tehát az egyetlen különbség $\mathbf{B}$ és $\mathbf{Z}$ oszlopainak korrespondencia-analízise között a főinerciák értéke, mely érinti a főkoordináták skáláját. Ezért az indikátormátrix oszlopainak az analízise inkább tekinthető *páronkénti kétváltozós*, mint *tömörített többváltozós* elemzésnek.

A Burt-mátrix partícionált formában Q számú változó kovarianciamátrixának analógiája, ahol minden egyes $\mathbf{Z}_q^T\mathbf{Z}_{q^*}$ mátrix egy-egy kovarianciának felel meg.

6. Latens dimenziók feltevése

A latens modell szerint adott x_j manifest változó indikátorjellegű abban az értelemben, hogy értékei megfigyelésenként valamely latens – létező, de nem megfigyelhető – f_t faktorok mozgásainak megfelelően alakulnak, és az indikátort végül egy, csak hozzá tartozó *egyedi hibafaktor egészíti ki* teljessé:[6]

$$x_j=\lambda_{j1}f_1+\lambda_{j2}f_2+\ldots+\lambda_{jt}f_t+\ldots+\lambda_{jm}f_m+u_j\,. \qquad /23/$$

[6] A következőkben a mátrix zárójelben szereplő alsó indexe a mátrix rendjére utal.

Valamennyi (j=1,2,…,m) indikátor változót közös vektorba foglalva, mátrix formában írva:

$$\mathbf{x}_{(p,1)} = \Lambda_{(p,m)}\mathbf{f}_{(m,1)} + \mathbf{u}_{(p,1)}, \qquad /24/$$

ahol $\mathbf{x}=[x_1,x_2,...,x_p]^T$ tartalmazza a p indikátort, $\mathbf{f}=[f_1,f_2,...,f_m]^T$ az $m<p$ latens faktort és $\mathbf{u}=[u_1,u_2,...,u_p]^T$ a unique (egyedi) faktorokat.

A $\mathbf{\Lambda}$ súlymátrix elemei a λ_{jk} értékek. Minél magasabbak abszolút értelemben, annál fontosabb a faktor. Megfigyeléseket végezve, valamennyi indikátorra az SVD-modellel analóg, de lényegileg eltérő formula adódik:

$$\mathbf{X}_{(n,p)} = \mathbf{F}_{(n,m)}\Lambda^T{}_{(m,p)} + \mathbf{U}_{(n,p)}. \qquad /25/$$

A faktoranalízis hipotézise szerint az indikátorok körének korrelációs rendszerét mögöttes, latens változók okozati köre generálja.

A /24/ kifejezés alapján az indikátorok $\Sigma_{xx} = \mathbf{X}^T\mathbf{X}$ *szóródási* mátrixa:

$$\Sigma_{xx} = \Lambda\Sigma_{ff}\Lambda^T + \Sigma_{uu} + \Lambda\Sigma_{fu} + \Sigma_{uf}\Lambda^T, \qquad /26/$$

ahol $\Sigma_{fu} = \Sigma_{uf} = \mathbf{0}$. Korrelálatlansági megszorításokat téve az egyedi faktoroknak közös faktorokkal való kapcsolatára

$$\Sigma_{xx} = \Lambda\Sigma_{ff}\Lambda^T + \Sigma_{uu} \qquad /27/$$

adódik. Ha Σ_{uu} és Σ_{ff} *diagonálisak*, akkor a modellhez az

$$\Sigma_{xx} - \Sigma_{uu} = \Lambda\Sigma_{ff}\Lambda^T \qquad /28/$$

megoldására van szükség, mely csak akkor sajátérték-feladat, ha Σ_{ff} diagonális, és csak akkor végrehajtható, ha létezik az $\Sigma - \Sigma_{uu}$ redukált szóródási mátrix (vagy becslésének) spektrális felbontása. A megoldásra iteratív algoritmusok állnak rendelkezésre, figyelembe véve, hogy a redukált szóródási mátrix már nem pozitív definit.

Irodalom

HAJDU, O. [2002]: Category Selection and Classification Based on Correspondence Coordinates. *Hungarian Statistical Review*. 80. évf. 7. sz. 103–126. old.

HAJDU O. [2003]: *Többváltozós statisztikai számítások*. Központi Statisztikai Hivatal. Budapest.

HAJDU, O. [2004]: Diagnostics of the Error Factor Covariances. *Hungarian Statistical Review*. 82. évf. 9. sz. 68–94. old.

HUNYADI L. – VITA L. [2002]: *Statisztika közgazdászoknak*. Központi Statisztikai Hivatal. Budapest.

KERÉKGYÁRTÓ GY.-NÉ ET AL. [2008]: *Statisztikai módszerek és alkalmazásuk a gazdasági és társadalmi elemzésekben.* Aula Kiadó. Budapest.

Summary

The paper deals with the basic statistical relations – correlation, discrimination, association – in a multivariate approach with regard to the eigenvalues of the corresponding matricies to be analysed. The focus is mainly on the statistical meaning of the eigenvalues. A brief overview is presented.