EÖTVÖS LORÁND TUDOMÁNYEGYETEM
TERMÉSZETTUDOMÁNYI KAR

A PARCIÁLIS LEGKISEBB NÉGYZETEK REGRESSZIÓ

Szakdolgozat

Horváth Vivien

Matematika BSc
Matematikai elemző szakirány

Témavezető:

Pröhle Tamás

Valószínűségelméleti és Statisztika Tanszék

Budapest,2012

Tartalomjegyzék

1. fejezet

Bevezetés

Az információ-technológia és a számítástudomány gyors fejlődésével nagy mértékben növekszik a piacgazdaság szereplőinek információigénye. Az adatok mennyiségének rohamos növekedése nem jár együtt a megfelelő mértékű információnövekedéssel. Az adatok felhasználóinak nem azok hiányával, hanem bőségével kell szembenézniük, mivel a legfinomabb becslések szerint is, az elektronikusan tárolt adatok mennyisége évente legalább megkétszereződik. A rendelkezésre álló adatok nagy mennyisége növeli ezek elemzésének összetettségét és az adatelemzőkkel szemben támasztott elvárásokat. Az ilyen esetekben muszáj matematikai statisztikához fordulni a minél pontosabb információ kinyerés és precíz becslések érdekében.
Az elemzésekhez többféle eljárást is lehet használni, melyek közül mindegyiknek megvan a maga előnye. A választásunk elsősorban az adatok tuladonságaitől függ, úgy mint paraméterek száma, megfigyelések száma és az egyes változók közötti kapcsolat.[11]

Szakdolgozatomban a parciális legkisebb négyzetek módszerét szeretném bemutatni és összehasonlítani négy másik népszerű eljárással úgy mint főkomponens-analízis, faktoranalízis, kanonikus korreláció és a SEM módszer. A dolgozat első felében ezen eljárások algoritmusát és elméleti hátterét kívánom ismertetni, míg a második részben a PLS eljárást szeretném bemutatni egy valós adathalmazon. A PLS eljárást elsősorban abban az esetben tekinthetjük az egyik leghitelesebb eljárásnak, amikor a paraméterek száma nagy a megfigyelések számához képest.

2. fejezet

Főkomponens-elemzés

A főkomponens-analízis a legegyszerűbb többváltozós statisztikai eljárások egyike. Elsődleges célja az összegzés és dimenziócsökkentés. A faktoranalízissel ellentétben, ahol az elsődleges cél a a mérési változók komplex, szerteágazó kapcsolatrendszerének egyszerűsítése közvetlenül nem mérhető, kevés számú tulajdonság úgynevezett latens változó vagy másképpen faktor alakulására visszavezetni, addíg a főkomponens-analízis elsődleges célja a teljes variancia nagy hányadát kevés változónak tulajdonítani. A fejezet logikai felépítésében leginkább az [1],[2] és [3] műveket vettem alapul.

Az egész eljárást érdemes úgy felfogni, hogy az $X_1, X_2 \dots X_p$ változók egy p dimenziós teret feszítenek ki, de nem ortogonálisan. Mi az eljárás során el szeretnénk készíteni a tér egy ortogonális bázisát. Az új bázis az $F_1, F_2, \dots F_p$ tehát ortogonális és F_t az $X_1, X_2 \dots X_p$ $(t = 1 \dots p)$ lineáris kombinációja. Nyilván 1-re normált lineáris kombinációk jöhetnek szóba, hogy az össz-szórás (covariancia mátrix normája) ne változzon. Szukcesszívan maximalizálni akarjuk a változók egy lineáris kombinációjának szórását. Lényegileg egy olyan irányt, dimenziót amely mentén a valószínűségi változók maximálisan szóródnak.

A főkomponensek a mérési változók olyan lineáris kombinációi, amelyek páronként ortogonális rendszert alkotnak.

A mérési változókat jelölje : $x_j \quad (j = 1, \dots, p)$
ezek lineáris kombinációit: $k_t \quad (t = 1, \dots, p)$

A főkomponensek az $i = 1 \dots N$ elemű sokaságban értelmezett, $\delta^2_{k_t}$ szóródási mértékeikkel a mérési változók szóródására vonatkozó, összesített δ^2_x információt maradék

nélkül visszaadják:

$$\delta_x^2 = \sum_{j=1}^{p} \delta_{x_j}^2 = \sum_{t=1}^{p} \delta_{k_t}^2$$

Ha a változónkénti szóródást az egyes változók λ másodrendű momentumával mérjük akkor:

$$\delta_x^2 = \sum_{j=1}^{p} \lambda_{x_j} = \sum_{t=1}^{p} \lambda_{k_t}$$

Amennyiben a mérési változók centrált, zérus átlagúak, akkor a főkomponensek is zérus átlagúak, s így varianciáik összegével a mérési változók varianciáinak összegét, vagyis szórásaik négyzetösszegét tudjuk teljes egészében magyarázni. A páronkénti ortogonalitásuk pedig páronkénti korrelálatlanságukat jelenti:

$$\delta_x^2 = \sum_{j=1}^{p} \sigma_{x_j}^2 = tr(C_{xx}) = \sum_{t=1}^{p} \sigma_{k_t}^2 = tr(C_{kk})$$

ahol C_{xx} a mérési változók kovariancia mátrixa, C_{kk} pedig a főkomponensek diagonális kovariancia mátrixa. Abban az esetben, ha a centrált mérési változók standardizáltak is, akkor kovariancia mátrixuk megegyezik korrelációs mátrixukkal, varianciáik összege pedig e változók számával:

$$\sum_{j=1}^{p} \sigma_{x_j}^2 = p \quad \sum_{j=1}^{p} \sigma_{x_j}^2 = tr(R_{xx}) = \sum_{t=1}^{p} \sigma_{k_t}^2 = tr(C_{kk})$$

A mérési változókat az X mátrix oszlopaiba helyezve, és képezve:

$$D_{xx} = \frac{1}{N} X^T X$$

szóródási mátrixot, ennek a mátrixnak a diagonális $\delta_{x_j}^2$ elemei megadják az egyes mérési változók szóródását, diagonális elemeinek összege pedig a mérési változók összesített δ_x^2 szóródását:

$$tr(D_{xx}) = \sum_{j=1}^{p} \delta_{x_j}^2$$

Az ilyen tulajdonságú főkomponensekkel szemben további követelmény, hogy szóródásuk mértéke monoton csökkenjen valamint, hogy az adott főkomponens az összes δ_x^2 információból a tőle nagyobb szóródásúak által nem reprodukált hányad maximális részét reprodukálja:

$$\delta_{k_1}^2 \geq \delta_{k_2}^2 \geq \cdots \geq \delta_{k_p}^2 \geq 0$$

A főkomponensek létrehozásával az a célunk, hogy ilyen tulajdonságú főkomponensek egy $k_1, k_2, \ldots k_{m<p}$ szűk csoportjával helyettesítsük a sokaság egyedeinek a

jellemzésére szolgáló változókat. A következő mutató a szóródásra vonatkozó összes információból megmagyarázott hányadot mutatja, melyet jelöljön IP

$$0 \leq IP = \frac{\sum_{t=1}^{m} \delta_{k_t}^2}{\delta_x^2} \leq 1$$

A változók számának csökkentésével további céljaink a következők:

1. A mérési változók össz-szóródásában rejlő információ nagy részét kevés számú változóba tömöríteni.

2. Az x mérési változókat a velük legszorosabban korreláló főkomponenshez rendelve olyan $X_{k_1}, X_{k_2}, \ldots X_{k_m}$ homogén csoportosulásait "kirajzolni", amely esetén az x változók csoporton belül egymással szorosan, más csoportok változóival viszont gyengén korrelálnak.

Fontos megjegyezni, hogy a szóródási mutatók δ_x^2 összesítésének csak akkor van értelme, ha a mérési változók azonos mértékegységűek, vagy ha normáltak, illetve standardizáltak. Ezenkívül, míg az egyes x változók magyarázatához való hozzájárulás mértékének megítélése egyenlő szóródású, azaz normált addíg az összesített szóródás reprodukálásához való hozzájárulás megítélése maximális szóródású főkomponensek alkalmazását igényli. Könnyen megállapítható, hogy a leginkább szóródó főkomponensek tartalmát adó mérési változókra vonatkozóan szóródik a leginkább, és a legkevésbé szóródó főkomponensekkel szorosan korreláló változók tekintetében szóródik a legkevésbé a sokaság.
A főkomponensek abban az esetben alkalmasak további elemzésre, ha kevés számú főkomponens minimális információvesztéssel képes helyettesíteni a mérési változókat. Az így nyert főkomponensek további felhasználási lehetősége az egyes főkomponensek megfigyelési egységekre számított értékének a hasznosítása. E területek az alábbiak:

1. A megfigyelési egységeket rangsorolhatjuk és csoportosíthatjuk az egyes főkomponensekben felvett értékeik alapján ezáltal a rangsor egyidejűleg több megfigyelési változó tekintetében jellemzi a megfigyelési egységek rangpozicióit.

2. Az extrém nagy vagy kicsiny főkomponens értékek lehetővé teszik az úgynevezett "outlierek" felismerését, kiszűrését.

3. A főkomponensváltozók létrehozása további statisztikai eljárások alapjául is szolgál.

2.1. A főkomponensek meghatározása

A mérési változók szóródási mátrixának vegyük a következő alakú spektrál felbontását:

$$D_{xx} = \frac{1}{N} X^T X = V \Lambda V^T$$

Ahol $V = [v_{jt}]$ jelöljön egy ortonormált mátrixot, oszlopaiban a $v_1, v_2, \ldots, v_p$ sajátvektorokkal.

Λ jelöljön egy diagonális mátrixot, diagonálisát képezzék a $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ sajátértékek.

A v_{jt} általános elem sorindexe mérési változóra, oszlopindexe pedig főkomponensre utal. Ezen információk birtokában a súlyozást az alábbi mátrix szemlélteti :

$$\begin{pmatrix} \text{változó} & k_1^{(\lambda_1)} \cdots & k_p^{(\lambda_p)} \\ x_1 & v_{11} \cdots & v_{1p} \\ \vdots & \vdots & \vdots \\ x_p & v_{p1} \cdots & v_{pp} \end{pmatrix}$$

Ezek alapján a t-edik maximalizált főkomponens meghatározása a megfelelő v_t sajátvektor segítségével:

$$k_t^{\lambda_t} = v_{1t}x_1 + v_{2t}x_2 + \cdots + v_{pt}x_p = \sum_{j=1} p v_{jt} x_j$$

Ekkor a t-edik normált főkomponenst úgy kapjuk meg, hogy főkomponens elemeit a $\sqrt{\lambda_t}$ négyzetes átlaggal normáljuk:

$$k_t^{(1)} = \frac{1}{\sqrt{\lambda_t}} k_t^{\lambda_t} = \sum_{j=1}^{p} \frac{v_{jt}}{\sqrt{\lambda_t}} x_j = \sum_{j=1}^{p} q_{jt} x_j$$

ahol a

$$q_{jt} = \frac{v_{jt}}{\lambda_t}$$

súlyokat a normált főkomponens-értékek előállítására szolgáló faktor együtthatójának hívjuk. Ha a mérési változók zérus átlagúak, akkor a lineáris kombinációjukként definiált főkomponensek is zérus átlagúak.

A súlyozási séma adott sorát tekintve, a j-edik mérési változó reprodukálása a maximalizált főkomponensekből:

$$x_j = v_{j1}k_1^{\lambda_1} + v_{j2}k_2^{\lambda_2} + \cdots + v_{jp}k_p^{\lambda_p} = \sum_{t=1}^{p} v_{jt} k_t^{\lambda_t}$$

Ez alapján, a mérési változó reprodukálása a normált főkomponensek felhasználásával az

$$a_{jt} = v_{jt}\sqrt{\lambda_t}$$

úgynevezett faktorsúlyok segítségével történik:

$$x_j = \sum_{t=1}^{p} a_{jt} k_t^{(1)}$$

Látható, hogy mind a főkomponensek, mind a mérési változók, kölcsönösen egymás lineáris kombinációi, és a súlyok megválasztásától függően:

1. A maximalizált főkomponenseknek a sajátvektorok megfelelő sor elemeivel súlyozott lineáris kombinációit mérési változónak hívjuk.

2. A mérési változóknak a megfelelő sajátvektor elemeivel súlyozott lineáris kombinációja a maximalizált főkomponens .

3. A mérési változóknak a faktor-együtthatókkal súlyozott lineáris kombinációja a normált főkomponens.

4. A normált főkomponenseknek a faktorsúlyokkal súlyozott lineáris kombinációja a mérési változó.

A főkomponensek további jellemzése miatt fontos, hogy a főkomponensek és a mérési változók kapcsolatrendszerét mátrix algebrai formában is megfogalmazzuk.
Alkossák a $K^{(1)}$ mátrix oszlopai a normált $k^{(1)}$, a $K^{(\lambda)}$ mátrix oszlopait pedig a $k^{(\lambda_t)}$ maximalizált főkomponensek. Ezen jelölésekkel a mérési változók X mátrixának SVD felbontása:

$$\begin{aligned} X &= K^{(1)} \Lambda^{\frac{1}{2}} V^T \\ &= K^{(\lambda)} V^T \\ &= K^{(1)} A^T \end{aligned} \tag{2.1}$$

az

$$\frac{1}{N} K^{(1)^T} K^{(1)} = V^T V = I \tag{2.2}$$

ortonormáltsági követelmények mellett, ahol A a faktorsúlyok mátrixa. A (2.1) feladat egyben a mérési változók szóródási mátrixának:

$$\frac{1}{N} X^T X = D_{xx} = V \Lambda V^T$$

spektrálfelbontása is.
A szóródási mátrix nyoma a szóródási mátrix sajátértékeinek az összege:

$$tr(D_{xx}) = tr(V^T V \Lambda) = tr(A) \tag{2.3}$$

A (2.1) modell első azonosságának átrendezésével (mivel $V^T = V^{-1}$) a normált főkomponensek mátrixa:

$$K^{(1)} = X(V \Lambda^{\frac{1}{2}}) = XQ \tag{2.4}$$

ahol Q mátrix elemeit az előzőekben bevezetett faktor-együtthatók alkotják. Ha a mérési változók zérus átlagúak, akkor a lineáris kombinációjukként definiált főkomponensek is zérus átlagúak, tehát normált voltuk egyben standardizált voltukat is jelenti.
A (2.3) második azonosságának átrendezésével — a maximalizált főkomponensek $K^{(\lambda)}$ mátrixa:

$$K^{(\lambda)} = XV$$

Mivel a fenti azonosságokból (2.4) ortonormáltsági követelmények miatt a maximalizált főkomponensek diagonális szóródási mátrixa:

$$\frac{1}{N}K^{(\lambda)^T}K^{(\lambda)} = D_{kk} = V^T D_{xx} = V^T D_{xx} V = \Lambda \tag{2.5}$$

Ezért a (2.2) és (2.4) egybevételéből látható, hogy:

$$tr(D_{xx}) = tr(\Lambda) = D_{kk}$$

tehát a páronkénti ortogonális főkomponensek négyzetes átlagainak négyzetösszege megegyezik a mérési változók négyzetes átlagainak négyzetösszegével, vagyis másodrendű momentumaik összegével. Nézzük most az:

$$A = V\Lambda^{\frac{1}{2}}$$

mátrixba foglalt faktorsúlyokat, e mátrix sorainak skaláris szorzatai maradék nélkül reprodukálják a mérési változók szóródási mátrixát:

$$AA^T = (V\Lambda^{\frac{1}{2}})(V\Lambda^{\frac{1}{2}})^T = V\Lambda V^T = D_{xx} \tag{2.6}$$

oszlopainak skaláris szorzatai pedig a főkomponensek diagonális szóródási mátrixát reprodukálják maradék nélkül:

$$A^T A = \Lambda^{\frac{1}{2}} V^T V \Lambda^{\frac{1}{2}} = \Lambda = D_{xx}$$

A (2.5) azonosságból következik a reprodukált szóródási mátrix invariancia tulajdonsága.
Nézzük a súlymátrix első m számú oszlopát tartalmazó A_m mátrixot és T legyen a transzformációs mátrix, melyre $TT^T = I$. Egyszerű átrendezéssel:

$$(A_m T)(A_m T)^T = A_m T T^T A_m^T = A_m A_m^T \tag{2.7}$$

vagyis a:

$$\hat{D}_{xx(m)} = A_m A_m^T$$

reprodukált szóródási mátrix (csak $m = p$ esetén egyezik meg az eredeti szóródási mátrixszal) invariáns az ortogonális transzformációkra. Innen kivonással megkaphatjuk a reziduális szóródási mátrixot is:

$$D_{e(m)} = D_{xx} - \hat{D}_{xx}(m)$$

Faktorstruktúrának nevezzük a mérési változók és a normált főkomponensek közötti páronkénti kapcsolatrendszerét leíró kovariancia (korrelációs) mátrixot, melyet a kovariancia (2.6) szerinti aszimmetrikus számítási lehetősége alapján a következőképpen írhatunk fel:

$$C_{xk} = \frac{1}{N} X_d^T K^{(1)}$$

és innen kifejezhetjük mind a faktorsúlyok, mind a faktor-együtthatók felhasználásával. Mivel a (2.1) harmadik egyenletét centrált változókra felírva, és behelyettesítve a fenti egyenletbe:

$$C_{xk} = A \frac{1}{N} K_d^{(1)^T} K(1) = A C_{kk}$$

adódik, vagyis a faktorstruktúra egyrészt a faktorsúlyok, másrészt a főkomponensközi kovarianciák függvénye. Ha a korrelálatlan főkomponensek egyben standardizáltak is, akkor C_{kk} egységmátrix, tehát ekkor a faktorstuktúra megegyezik a faktorsúlyok mátrixával: $C_{xk} = A$. A (2.5) azonosságot balról szorozva az $\frac{1}{N} X_d^T$ mátrixszal:

$$C_{xk} = \frac{1}{N} X_d^T X Q = C_{xx} Q$$

ahonnan a faktor-együtthatók mátrixa (az invertálhatóságot feltételezve):

$$Q = C_{xx}^{-1} C_{xk} \tag{2.8}$$

Ha a főkomponensek korrelálatlanok, akkor a faktor-együtthatók mátrixa a következő módon is számolható:

$$Q = C_{xx}^{-1} A$$

A (2.8) formulának az a gyakorlati jelentősége, hogy ha csak az első m főkomponenst tartjuk meg, akkor a normált változatuk előállításához szükséges súlyok mátrixa:

$$Q_m = C_{xx}^{-1} C_{xk_m} \tag{2.9}$$

ahol C_{xk_m} a struktúra mátrix első m oszlopát tartalmazza. Ha az első m orthogonális főkomponens A_m súlymátrixán olyan transzformációt hajtunk végre, amelynek eredményeképpen A'_m már korrelált főkomponensekre vonatkozik, akkor az első m korrelált főkomponens meghatározása a (2.9) alkalmazást igényli.

Az olyan statisztikai elemzések melyek variancia tömörítésre épülnek a mérési változók összesített varianciáját ismert arányban felosztó főkomponensekre épülnek. Ha a mérési változók átlagos értékei tetszőlegesek, akkor semmilyen általános következtetés nem vonható le a főkomponensek átlagaira, s így varianciáikra vonatkozóan sem. Viszont ha a mérési változók zérus átlagúak, akkor lineáris kombinációjukként létrehozott főkomponensek is zérus átlagúak, és ekkor másodrendű momentumuk a varianciájukkal egyezeik meg. Tehát a variancia tömörítésére szolgáló főkomponenseket — standardizált vagy nem standardizált — de zérus átlagú mérési változók főkomponenseinek a meghatározásával nyerjük.

3. fejezet

Faktoranalízis

A faktoranalízis során figyelmünket elsősorban nem a mérési változók összes varianciájára, hanem e változók kapcsolatrendszere felé fordítjuk. Szemben a főkomponens analízissel, mely a teljes variancia nagy hányadát kísérli meg kevés változónak tulajdonítani, a faktoranalízis a mérési változók komplex, szerteágazó kapcsolatrendszerét próbálja meg egyszerűsíteni közvetlenül nem mérhető, kevés számú tulajdonság úgynevezett latens változó vagy másképpen faktor alakulására visszavezetni, azok eredményének betudni. Ehhez a részhez az [2] és a [3] forrást vettem alapul.
Az eljárások során jelentősen csökkentjük a változók számát. Bizonyos változók a megfigyelések során alig változnak, szórásuk kicsi, ezeket tehát nem tekinthetjük jellemzőnek, ha tudjuk melyek ezek elhagyhatjuk őket. Gyakran nem egy változó kis szórású, hanem kettő összege, vagy valamelyik másik lineáris kombinációja, tehát ezeket keressük, illetve az eljárás szempontjából inkább azokat, amelyeknek nagy a szórása tehát nem hagyhatóak el. A faktoranalízis alapvetően nem variancia, hanem kovariancia, korreláció-orientált módszer.

Az eljárás során megkeressük az eredeti változók egymással szorosabb korrelációban lévő csoportjait, ezeket a változókat egy faktorhoz tartozónak tekintjük. Olyan faktorokat keresünk tehát amelyek közvetlenül nem figyelhetőek meg, de feltételezésünk szerint a vizsgált mérési változók alakulását befolyásolják, ily módon összekapcsolva azokat. Amennyiben sikerült ilyen csoportokat elkülönítenünk, a következő feladat a faktorok értelmezése. Így a nagyszámú eredeti változót néhány faktorban összesíthetjük, és ezekkel mint új változókkal dolgozhatunk tovább.

Annak eldöntésére, hogy mikor alkalmazzunk faktoranalízist a következő statisztikák segítenek:

1. Ha a korrelációs mátrix alapján a változók úgy csoportosíthatóak, hogy az egy csoporton belüli változók között viszonylag magas a korreláció, ezzel szemben a csoportok között pedig alacsony.

2. A parciális korrelációk kicsik.

3. A Kaiser-féle mutatószám, amelyet neveznek Kaiser-Meyer-Olkin statisztikának is, 0.8-nél nagyobb. Ha ez a mutatószám viszont 0.5-nél kisebb, akkor nem ajánlott faktoranalízis végrehajtása.

$$KMO = \frac{\sum_i \sum_j r_{ij}^2}{\sum_i \sum_j r_{ij}^2 + \sum_i \sum_j a_{ij}^2}$$

ahol r_{ij} az i -dik és j-dik megfigyelt változó közötti korrelációs együttható, a_{ij} az i-dik és j -dik megfigyelt változók közötti parciális korrelációs együttható.

A faktoranalízis egyaránt támaszkodhat a kovariancia illetve a korrelációs mátrix elemzésére. A választás azon múlik, hogy meg akarjuk-e őrizni az eredeti skálát vagy sem. A faktormodellben azt mondjuk meg, hogyan függnek az egyes változók a faktoroktól, mely lineáris kombinációval állíthatjuk elő őket. Fontos tudni, hogy a faktoranalízist többféle módszerrel hajthatjuk végre.

3.1. Az általános modell

Tegyük fel, hogy a faktorok korrelálatlanok és 0 várható értékűek.
Ennek megfelelően a faktoranalízis mátrixos modellje:

$$X = AF + D$$

X jelöli a megfigyelt változókból álló n dimenziós véletlen vektort,
F a faktorok k dimenziós véletlen vektorát ($k < n$),
$A = (a_{ij})$ egy $n \cdot k$ méretű rögzített mátrixot, ennek a neve átviteli (loading) mátrix, elemei a faktorsúlyok,
D pedig az X változóhoz tartozó egyedi hibafaktort.

Valamennyi faktor valamennyi mérési változó magyarázatához hozzájárul valamilyen mértékben, s az egyes változók alakulásának ily módon meg nem magyarázott

részét változóként külön-külön egy további véletlen változó, a D faktor képviseli. Az olyan faktort, amely valamenyi mérési változó alakulását befolyásolja, közös faktornak nevezzük. Látható, hogy a faktormodell egyenletei tulajdonképpen speciális regressziós egyenletek, a mérési változók az eredmény jellegű, a faktorok pedig a magyarázó jellegű változók. A feltételünkből, hogy az egyedi faktorok páronként korrelálatlanok, következik, hogy a kovariancia mátrixuk diagonális lesz:

$$C_{DD} = \Psi = \frac{1}{n} D^T D = < \psi_1^2, \dots \psi_p^2 >$$

ahol ψ_j^2, a j-dik egyedi faktor varianciája. Mivel az egyedi és a nem egyedi faktorok egymással korrelálatlanok, ezért:

$$Cov(F_k, D_j) = 0 \qquad (k = 1, \dots, m;) \qquad (j = 1, \dots, p)$$

Ezek után a mérési változók kovariancia mátrixa — a faktorokról rendelkezésre álló információk alapján — az alábbi módon reprodukálható :

$$\begin{aligned} C_{xx} = \frac{1}{n} X^T X = \frac{1}{n} (F\Phi^T + D)^T (F\Phi^T + D) = \\ = \Phi (\frac{1}{n} F^T F) \Phi^T + \frac{1}{n} D^T D = \\ = \Phi C_{FF} \Phi^T + C_{DD} \end{aligned} \tag{3.1}$$

ahol C_{FF} a latens faktorok kovariancia mátrixa.
A (3.1) egyenletet nevezzük a faktoranalízis általános alapegyenletének. Ha a mérési változók és a faktorok is mind standardizáltak, akkor a faktoranalízis alapegyenlete a következőképpen változik:

$$R_{xx} = \Phi R_{FF} \Phi^T + C_{DD} = R^* + C_{DD} \tag{3.2}$$

ahol R_{FF} a latens faktorok, R^* pedig a redukált korrelációs mátrix.
A faktorok korrelálatlanságán kívül tegyük fel, hogy a megfigyelt változók standardizáltak, ekkor kapjuk meg a faktoranalízis alapegyenletének azt a formáját, mely a legtöbb megszorítást tartalmazza a vizsgált változókra vonatkozóan:

$$R_{xx} = \Phi \Phi^T + C_{DD}$$

Ezek szerint a megfigyelt változók páronkénti korrelációs rendszere, vagyis az átlón kívül eső elemek teljesen reprodukálhatóak a faktorsúlyok segítségével, viszont az átlón elhelyezkedő egységnyi standardizált változók reprodukálásához az egyedi faktorok ψ^2 varianciáinak az ismerete is szükséges.

$$r_{x_j x_t} = \phi_{j1}\phi_{t1} + \dots + \phi_{jm}\phi_{tm} \quad (j \neq t)$$

és

$$r_{x_j x_j} = 1 = \phi_{j1}^2 + \cdots + \phi_{jm}^2 + \psi_j^2 \quad (j = 1, \ldots, p)$$

A mérési változók korrelációs rendszeréért a modell szerint tehát kizárólag a latens faktorok felelősek.

A közös variancia

A faktoranalízis általános modelljében az x_j mérési változó két korrelálatlan részre bontható:

$$x_j = c_j + D_j$$

ahol

$$c_j = \phi_{j1} F_1 + \cdots + \phi_{jm} F_m$$

a változóknak azon része, mely közös a többi megfigyelt változóval, míg D_j az egyes változók egyediségét jelenti. Mivel azt tettük fel, hogy a változók közös és egyedi része korrelálatlan, ezért az x_j változó varianciája is két részre bontható:

$$Var(x_j) = Var(c_j) + Var(D_j) \tag{3.3}$$

ahol $Var(c_j)$ a latens faktorok által közösen magyarázott, $Var(D_j)$ pedig e változó egyedi varianciáját képviseli. A közös varianciát az x_j változó kommunalitásának nevezzük. Valamely mérési változó kommunalitása az illető változó varianciájából a latens faktorok által adott megmagyarázott hányadot adja meg. Jelölje az x_i változó kommunalitását h_j^2, ahol

$$h_j^2 = \sum_{i=1}^{k} a_{ij}^2$$

Figyelembe véve, hogy $Var(D) = \psi_j^2$, a varianciát a következőképpen bonthatjuk fel:

$$Var(x_j) = h_j^2 + \psi_j^2$$

Mivel a faktoranalízis során a célunk nem a teljes variancia tömörítése, ezért az egyes változók varianciájának csak az a része érdekes amelyet a közös faktorok segítségével magyarázni tudunk. A mérési változók R_{xx} korrelációs mátrixában a diagonális elemeket a megfelelő változók h_j^2 kommunalitásaival helyettesítve az R^* úgynevezett redukált korrelációs mátrixot nyerjük:

$$\mathbf{R}^* = \begin{pmatrix} h_1^2 & r_{12} \cdots & r_{1p} \\ r_{12} & h_2^2 \cdots & r_{2p} \\ \vdots & \vdots & \ddots \\ r_{p1} & r_{p2} \cdots & h_p^2 \end{pmatrix}$$

Standardizált és korrelálatlan faktorok esetében a h_j^2 kommunalitás a faktorsúly mátrix j-dik sorában a súlyok öszege:

$$h_j^2 = \sum_{k=1}^{m} \phi_{jk}^2 = \phi_j^2$$

Mivel ϕ_{jk}^2 az adott x_j mérési változó és az F_k faktor korrelációs kapcsolatát jellemző determinációs együttható, ezért erre az x_j mérési változó F_k faktorra vonatkozó egyedi kommunalitásaként tekinthetünk.
Jelölje VP a faktorsúly mátrixot, k-dik oszlopában a súlyokat pedig:

$$VP_k = \sum_{j=1}^{p} \phi_{jk}^2 = \phi_{.k}^2$$

négyzetösszege adja meg, ami megmutatja, hogy az F_k faktor milyen mértékben járul hozzá a mérési változók összes varianciájának a magyarázatához.

A faktorok meghatározatlansága

A faktorok meghatározása nem egyértelmű, mivel a faktorok elforgathatóak, és a kommunalitások nem ismertek az eljárás elején.
Faktorok elforgatásán a következőt értjük: ha egy $n * n$-es T mátrix olyan, hogy $TT^T = I$ ahol I az egységmátrixot jelöli, akkor minden $B * T$ megoldása az alapegyenletnek.
Ha már tudjuk a faktormodell egy megoldását, akkor a faktorok bármilyen ortogonális forgatása az ortonormált T transzformációs mátrixszal az m dimenziós térben kielégíti a (3.2) alapegyenletet:

$$(\Phi T)(T^T R_{FF} T)(\Phi T)^T = \Phi T T^T R_{FF} T T^T \Phi^T = \Phi R_{FF} \Phi^T$$

Vagyis, a transzformációs mátrixszal a faktorsúlyoknak és a faktorkorrelációnak olyan új, elforgatott változatait tudjuk létrehozni, melyek a mérési változók korrelációs rendszerét változatlanul hagyják. Ilyen transzformációs mátrix viszont végtelen számú van, és mindegyikük más és más faktormodellt definiál. A faktoranalízis alkalmazásának kulcskérdése, hogy végülis melyik faktorstruktúrát válasszuk.
Viszont a redukált korrelációs mátrix összeállítása sem egyértelmű. A faktorok által megmagyarázott kommunalitások kiszámítása a faktorok ismereteit igényli, de a faktorok meghatározásának kiindulási pontja a redukált korrelációs mátrix ismerete. Ebből következik, hogy a kommunalitásnak mindíg induló becslést kell nyújtanunk, melyek az eljárás során egyre pontosabbakká válnak.

Ennek leggyakrabban használt módjai a következők:

– Az x_j változó kommunalitására azt a hányadot tekintjük induló becslésnek, amelyet e változóból az összes többi x változó megmagyaráz. Ez az arány megegyezik a kérdéses változónak az összes többi megfigyelt változóval vett többszörös determinációs együtthatójával:

$$\hat{h_j^2} = R^2_{j.1,\ldots,j-1,j+1,\ldots p.}$$

– A becsült kommunalitás lehet a megfigyelt változók korrelációs mátrixának j-dik sorának maximális abszolút értékű eleme.

– A kommunalitásokra úgy is adhatunk induló becslést, hogy x_j kommunalitásnak az összes többi megfigyelt változóval vett korrelációi abszolút értékeinek a számtani átlagát tekintjük:

$$\hat{h_j^2} = \frac{1}{p-1} \sum_{t=1} p|r_{jt}| \quad (T \neq j)$$

– Egy elöző lépésben végrehajtott főkomponens analízisből származó, az egyes főkomponensek által magyarázott többszörös determinációs arányokat induló becslésként szerepeltetjük.

A faktormodell felhasználása

Az általános faktormodell két alapvető felhasználási területe az exploratív és konfirmatív faktoranalízis.
Az exploratív faktoranalízisben nincs előzetes információnk a faktorok számáról valamint a faktor struktúráról. Ebben az esetben az adatállományt arra használjuk, hogy feltárjuk mindazon faktor jellemzőket, melyek a mérési változók korrelációs kapcsolatrendszerét a legjobban magyarázzák.

- meghatározzuk a faktorok minimális számát
- megkeressük a legmegfelelőbb faktorsúlyokat
- meghatározzuk a kommunalitásokat és az egyediségeket
- értelmezzük a faktorokat
- ha szükséges becsüljük a faktorokat megfigyelési egységekhez tartozó értékeit

Ezzel ellentétben a konfirmatív faktoranalízis során a priori hipotézissel élünk a faktorstruktúrára vonatkozóan, és azt vizsgáljuk, hogy a mintánk ellentmond-e ennek a hipotézisnek, vagy nem.

Exploratív faktoranalízis

A faktoranalízis alapvető célja előbb a faktorsúlyok, majd a közös faktorok előállítása. Mivel a faktorsúlyok számítása nem egzakt, ezért különbözőképpen indulhatunk ki a faktorsúlyok meghatározása során. Az alábbiakban két eljárást:

1. főfaktor analízis
2. maximum likelihood faktoranalízist

mutatjuk be, a faktoranalízis végrehajtására. Az exploratív faktor megoldások közös jellemzője, hogy a kommunalitásokra adott induló becsléseken alapulnak.

3.1.1. Főfaktorok módszere

A Faktoranalízisnek ez az eljárása a főkomponensek meghatározására vezeti visszza a faktorsúlyok becslését.
Tegyük fel, hogy a közös faktorok száma m, azonban sem m értékét sem az m számú latens faktorhoz tartozó kommunalitásokat nem ismerjük. Az ismeretlen redukált korrelációs mátrixot ekkor az alábbi módon írhatjuk fel:

$$R^* = A_m A_m^t$$

A faktorok számára vonatkozóan tehát hipotézissel, a kommunalitásokra vonatkozóan pedig becsléssel kell élnünk. Legyen a faktorok számára vonatkozó hipotetikus érték $q < m$. Ha téves a hipotézisünk, és kevesebb faktort szűrünk ki, mint valójában kellene, akkor a redukált korrelációs mátrix még akkor is csak maradékkal reprodukálható, ha egyébként ismernénk az egzakt kommunalitásokat. Természetesen a redukált korrelációs mátrix a latens faktorok egzakt számának az ismeretében is csak maradékkal reprodukálható, ha nem ismerjük a valódi kommunalitásokat. Ekkor a reziduális korrelációs mátrix:

$$R_{e(q)} = R_{xx} - \phi_q \phi_q^T$$

A főfaktoranalízis célja olyan közös főfaktorok előállítása, amelyek közül az első maximális arányban magyarázza a mérési változók összes varianciáját, majd a második a maradék varianciát magyarázza maximális arányban, miközben korrelálatlan az első faktorral, és így tovább, egészen F_q faktorig. A varianciából megmagyarázott hányad maximalizálásával egyidőben az is teljesül, hogy a reziduális korrelációs mátrix elemeinek a négyzetösszege minimális bármely másik q közös faktor esetéhez képest.

A főfaktoranalízis megoldása a következő lépésekből áll.

- Először induló becslést adunk az ismeretlen kommunalitásnak, s az így kapott értékekkel helyettesítve az R_{xx} mátrix átlóját, a redukált korrelációs mátrix egy induló becslését kapjuk, amit $\widehat{R^*_{(0)}}$ jelöl.
  Ezt követően — q számú faktort feltételezve — a becsült redukált korrelációs mátrixból kiindulva iteratív főkomponens-analízist hajtunk végre.

- Az első iterációs lépés során meghatározzuk az $\widehat{R^*_{(0)}}$ első q főkomponenséhez tartozó főkomponens súlyok A mátrixát. A megfelelő x változók kommunalitásaira nyerhetünk újabb becslést ha a főkomponens súlyok soronkénti négyzetösszegeit képezzük. Most ezen becsléseket helyettesítve az eredeti R_{xx} korrelációs mátrix átlójára kapjuk a redukált korrelációs mátrix újabb $\widehat{R^*_{(1)}}$ becslését.

- A második iterációs lépésben már ez utóbbi $\widehat{R^*_{(1)}}$ redukált korrelációs mátrix első q főkomponenséhez tartozó A súlymátrixot határoztuk meg, majd a súlyok segítségével újra becsüljük a kommunalitásokat. Ezeket az értékeket R_{xx} diagonálisában szerepeltetve, a redukálált korrelációs mátrix következő, $\widehat{R^*_{(2)}}$ becslésével rendelkezünk.

- Általánosságban az i-dik iterációs lépésben a megelőző lépésből származó becsült $\widehat{R^*_{(i-1)}}$ redukált korrelációs mátrix első q főkomponensének A súlyait keressük, majd ezek felhasználásával nyerjük a redukált korrelációs mátrix $\widehat{R^*_{(i)}}$ becslését.

Mivel valamennyi iterációs lépés egy-egy önálló főkomponens-analízis, ezért a sajátértékek, a sajátvektorok, s így a főkomponens súlyok és a kommunalitások iterációról iterációra változnak. A főfaktoranalízis során minden egyes iteráció után megvizsgáljuk, hogy jelentősen változnak-e a kommunalitások megfigyelt változóként külön:

$$|h^2_{j(i-1)} - h^2_{j(i)}| < \delta \quad (j = 1, \ldots, p)$$

ahol δ előre rögzített kicsi pozitív szám. Ha ez a leállási feltétel éppen az i-dik lépés után mindegyik x_j esetén teljesül akkor az $m = q$ hipotézis mellett végső megálláshoz jutottunk:

$$\Phi_q = A_{q(i)}$$

Az $m = q$ hipotézis nem biztos, hogy helytálló. Ezért célszerű a faktorokat szekvenciálisan, fokazatosan léptetni be a hipotézisünkbe, s mindig megvizsgálni a reziduális korrelációs mátrix elemeit. Ha a q-dik lépésben R^*_e elemei már nagyon kicsik, akkor

nincs értelme a $(q+1)$-dik faktort is meghatározni, s önkényesen elfogadjuk az $m = q$ számú latens faktor létezésének hipotézisét.
Fontos, hogy ha mégis ki akarjuk szűrni a $(q + 1)$-dik faktort akkor valamely x változók kommunalitása nagyobb lehet mint 1, ami értelmetlen. Ezért lehetséges, hogy a becsült redukált korrelációs mátrix nem pozitív szemidefinit, s így negatív sajátértékei is lehetnek. Mivel azonban $\hat{R}^*$ sajátértékeinek az összege egyenlő a kommunalitások összegével, ami az első m sajátérték $\sum_{k=1}^{m} \lambda_k$ összege is egyben, ezért a pozitív sajátértékek összege meghaladja a teljes kommulitás értékét. Ilyenkor maximum annyi faktort lehet és érdemes meghatározni, amelyek pozitív sajátértékeinek az összege felülről közelíti a teljes kommunalitást.

3.1.2. Maximum likelihood faktoranalízis

Tegyük fel, hogy m számú közös, standardizált faktor van, s ezen faktorok korrelálatlanok. Célunk, hogy a Φ és Ψ paramétereket a maximális likelihood elv alapján becsüljük. Mivel a ML módszer valamely paraméterekre rögzített eloszlás típus mellett nyújt becslőfüggvényt, ezért a ML faktoranalízis során további feltételezéseket is kell tennünk.
Tegyük fel, hogy az F közös és D egyedi faktorok független, többváltozós normális eloszlást követnek zérus átlagvektorral. Ebből következik, hogy mivel a mérési változók a faktorok lineáris kombinációi, feltételezésünk szerint a mérési változók is többváltozós normális eloszlást követnek zérus átlagvektorral, $\sum$ kovariancia mátrixszal.
Az n számú megfigyelésünket egy n elemű független mintának tekintve a mintából számított korrigált kovariancia mátrix:

$$S = \frac{1}{n-1} X^T X$$

ezért a maximum likelihood elvnek megfelelően az L likelihood függvény:

$$lnL = -\frac{n}{2}(ln|\sum| + tr(S\Sigma^{-1})) + konst. \rightarrow max.$$

szerinti logaritmusát maximalizáljuk.

Paraméterbecslés

Mivel a modell szerint $\Sigma = \Phi\Phi^T + \Psi$, ezért nyilván a likelihood is függvénye a Φ és Ψ paramétereknek, tehát változókként kezelve e paramétereket keressük azon $\hat{\Phi}$ és $\hat{\Psi}$ paraméterket, amelyekre lnL maximális.
Hipotézisvizsgálati szempontból azonban előnyösebb az

$$F_k(\Phi, \Psi) = ln|\Sigma| + tr(S\Sigma^{-1}) - ln|S| - p \tag{3.4}$$

célfüggvényt minimalizálni. A maximum likelihood normál egyenletekhez az F_k célfüggvény parciális deriválása útján jutunk:

$$\frac{\delta F_k}{\delta \Phi} = 2\Sigma^{-1}(\Sigma - S)\Sigma^{-1}\Phi$$

$$\frac{\delta F_k}{\delta \Phi} = diag.(\Sigma^{-1}(\Sigma - S)\Sigma^{-1})$$

ahol $diagonlis(Z)$ egy olyan diagonális mátrix, amelyet Z-ből úgy képezzük, hogy a nem diagonális elemeit 0-val helyettesítjük. Ezután az F függvény minimumát két lépés beiktatásával keressük:

1. Az első lépésben rögzített Ψ_0 egyediség mellett keressük F feltételes minimumát Φ-ben, amit jelöljön $F_k(\Phi_0, \Psi_0) = f_k(\Psi_0)$. Ekkor Φ_0 olyan értéke Φ-nek, amely eleget tesz az első egyenletnek.

2. Ezután pedig keressük a Ψ paramétert, amely már f_k-t feltétel nélkül minimalizálja. Ebből egy új Ψ_1 diagonális mátrixhoz jutunk, amelyből egy újabb Φ_1 mátrixot származtatunk, amelyből viszont egy újabb Ψ_2 számítható a nem lineáris minimalizálási eljárás révén.

A maximum likelihood faktoranalízis egy iterációs eljárást igényel, melynek során rögzítünk egy kiindulási Ψ_0 mátrixot, majd újabb és újabb $\Psi_1\Psi_2, \ldots \Psi_s, \Psi_{s+1}$ mátrixokhoz jutva vizsgáljuk, hogy az:

$$f(\Psi_{s+1}) < f(\Psi_s)$$

csökkenés jelentős-e? Ha a csökkenés már nem nagyobb egy előre rögzített pozitív értéknél, akkor a végső $\hat{\Psi}$ mátrixhoz jutottunk el, amiből $\hat{\Phi}$ már következik.
A maximum likelihood faktoranalízis fontos tulajdonsága, hogy skálainvariáns vagyis mindegy, hogy számításainkban az S az R korrelációs, vagy a C kovariancia mátrixot jelöli.

Hipotézisvizsgálat

A ML faktoranalízis előnye, hogy a maximum likelihood arány kritérium felhasználása lehetővé teszi azon H_m hipotézis tesztelését, miszerint éppen m számú közös faktor létezik.
Jelölje Ω a p-ed rendű szimmetrikus pozitív definit mátrixok halmazát, ω pedig azt a részhalmazt, amelyre:

$$\Sigma = \phi_m \phi_m^T + \Phi_m$$

teljesül, összhangban a H_m hipotézissel. Jelölje továbbá L_Ω a likelihood maximumát az Ω halmazon, L_ω pedig az ω részhalmazon. Ekkor a likelihood arány:

$$\lambda = \frac{L_\omega}{L_\Omega}$$

a likelihood arány teszt pedig:

$$-2ln\lambda = n(ln|\hat{\Sigma}| + tr(S\hat{\Sigma}^{-1}) - ln|S| - p)$$

Ez viszont nem más mint a (3.4) célfüggvény minimumhelyén vett értékének az n-szerese.

A H_m hipotézis helyessége és nagy mintaelemszám esetén a likelihood arány próbafüggvény közelítőleg χ^2 eloszlású:

$$\frac{p(p+1)}{2} + \Big(pm + p - \frac{m(m-1)}{2}\Big) = \frac{1}{2}\Big((p-m)^2 - p - m\Big)$$

szabadságfokkal.

4. fejezet

A kanonikus korreláció

Kanonikus korreláció esetén a változó halmaz természetes módon két részre van bontva, és a két változóhalmaz kapcsolatát vizsgáljuk. Különbséget teszünk a változók között aszerint, hogy függő vagy független, magyarázott vagy magyarázó, becsült vagy becslő változóról van szó. A kanonikus korreláció elemzést tekinthetjük a többszörös korreláció általánosításának is. Ezen fejezet alapjául a [4] és [6] irodalom szolgált.
A kanonikus korreláció esetén több x_i $(i = 1, \cdots m_1)$ magyarázó változó sztochasztikus kapcsolatát több y_j $(j = 1, \cdots m_2)$ magyarázott változóval vizsgáljuk.
A kanonikus korrelációs modellben az x_i és y_j változók olyan lineáris függvényeit keressük, amelyek közötti korreláció maximális. A kanonikus korrelációt úgy is felfoghatjuk, hogy az eljárás során faktorokat hozunk létre, ahol a két változóhalmaz azon faktorait keressük, amelyek közötti korreláció maximális.

4.1. A módszer leírása

Jelölje X az x_i változók n-szeri megfigyelésének mátrixát.
Y pedig az y_j változók n-szeri megfigyelésének mátrixát.
Tegyük fel az általános eset korlátozása nélkül, hogy $m_2 < m_1$. Standardizáljuk a változókat, mivel ebben az esetben a kovarianciamátrix megegyezik a korrelációs mátrixszal. Tehát tegyük fel, hogy x_i és y_j változók standardizáltak.

A kanonikus korreláció elemzés során az

$$x = [x_1, x_2, \cdots x_{m1}]' \quad \text{és az} \quad y = [y_1, y_2, \cdots y_{m2}]'$$

vektorváltozók sztochasztikus kapcsolatát vizsgáljuk az x és y vektorváltozók komponenseinek lineáris függvényein keresztül.
Legyenek a vektor elemei az x_i változók lineáris függvényei:

$$v = X \cdot c$$

ahol a c vektor $(m_1 \times l)$-es méretű elemeit súlyoknak nevezzük. Hasonlóan definiáljuk az y_i változók lineáris kombinációját:

$$w = Y \cdot d$$

ahol d egy $(m_2 \times l)$-es vektor és az úgynevezett következmény súlyokat tartalmazza. Kanonikus korrelációelemzés esetén a cél megtalálni a c és d súlyokat, amelyek mellett a v és w kanonikus változók közötti korreláció maximális. A v és w közötti korrelációt nevezzük kanonikus korrelációs együtthatónak. Az eljárás során két változó halmazból állítunk elő nem megfigyelt változópárokat, melyek maximálisan korrelálnak. Tegyük fel, hogy x_i és y_j változók standardizáltak, ekkor a korreláció mátrix megegyezik a kovarincia mátrixszal, és a következőképpen írhatjuk fel:

1. Az x_i valószínűségi változók között $R_{xx} = \frac{1}{n}X'X$ $(m_1 \times m_1)$,
2. az y_i valószínűségi változók között: $R_{yy} = \frac{1}{n}Y'Y$ $(m_2 \times m_2)$,
3. és az x_i és y_i változók között $R_{xy} = \frac{1}{n}X'Y$ $(m_1 \times m_2)$

A fenti mátrixokat egy általános korrelációmátrix $R = \frac{1}{n}(XY)'(XY)$ partícióinak nevezhetjük. Tegyük fel, hogy a v és w kanonikus változók standardizáltak, vagyis

$$\frac{1}{n}v'v = \frac{1}{n}c'X'Xc = c'R_{xx}c = 1 \tag{4.1}$$

$$\frac{1}{n}w'w = \frac{1}{n}d'Y'Yd = d'R_{yy}d = 1.$$

Ha a kanonikus változók standardizáltak akkor a v és w közötti korreláció:

$$\frac{1}{n}v'w = \frac{1}{n}c'X'Yd = \lambda \tag{4.2}$$

Olyan c és d súlyokat szeretnénk találni, figyelembe véve az (4.1) egyenlet szerinti feltételeket, hogy λ értéke maximális legyen. Vagyis egy feltételes szélsőértékfeladatot kell megoldanunk, ahol a feltételek egyenlőség formájában adottak. Erre a Lagrange-féle multiplikátor-módszer talál megoldást. A Lagrange-függvény:

$$L = c'R_{xy}d - \frac{1}{2}\mu[c'R_{xx}c - 1] - \frac{1}{2}\rho[d'R_{yy}d - 1] \tag{4.3}$$

Az L függvény maximuma ott keresendő, ahol a parciális deriváltak egyenlők nullával. A μ és a ρ multiplikátor előtti $\frac{1}{2}$ szorzó egyszerűsíti a deriválás után kapott egyenleteket. A parciális deriváltak:

$$\frac{\delta L}{\delta c'} = R_{xy}d - \mu R_{xx}c = 0 \tag{4.4}$$

$$\frac{\delta L}{\delta d'} = R_{yx}c - \rho R_{yy}d = 0$$

innen:

$$R_{xy}d = \mu R_{xx}c \quad \text{és} \quad R_{yx}c = \rho R_{yy}d \tag{4.5}$$

Ha az (4.5) egyenletek közül az elsőt szorozzuk balról c-vel, a másodikat balról d'-vel akkor kapjuk, hogy

$$c'R_{xy}d = \mu c'R_{xx}c = \mu \tag{4.6}$$

$$d'R_{yx}d = \rho d'R_{yy}d = \rho$$

A (4.2) egyenlet szerint $c'R_{xy}d = d'R_{yx}c = \lambda$ amiből következik, hogy $\mu = \rho = \lambda$. Így a (4.4) egyenlet helyett a következőket írhatjuk:

$$-\lambda R_{xx}c + R_{xy}d = 0 \tag{4.7}$$

$$R_{xy}c - \lambda R_{yy}d = 0$$

Tehát definiáltunk egy $m_1 + m_2 = m$ homogén egyenletből álló rendszert m ismeretlennel (c és d) és egy ismeretlen λ együtthatóval . Ekkor a (4.7) egyenletrendszert a következőképpen írhatjuk :

$$\begin{bmatrix} -\lambda R_{xx} & R_{yx} \\ Ryx & -\lambda R_{yy} \end{bmatrix} \cdot \begin{bmatrix} c \\ d \end{bmatrix}$$

Látható, hogy az egyenletrendszernek akkor és csak akkor van a triviálistól ($c \neq 0$ és $d \neq 0$) különböző megoldása, ha a determinánsa egyenlő nullával, tehát:

$$\begin{bmatrix} -\lambda R_{xx} & R_{yx} \\ Ryx & -\lambda R_{yy} \end{bmatrix} = 0 \tag{4.8}$$

Felhasználva a determinánsokra vonatkozó megfelelő tulajdonságokat: ha egy oszlopot vagy egy sort szorzunk (vagy osztunk) egy konstanssal, a determináns értéke is szorzódik (vagy osztódik) a konstanssal. Ha a (4.8) determináns első m_1 sorát megszorozzuk $(-\lambda)$-val, és az utolsó m_2 oszlopát elosztjuk $(-\lambda)$-val, a determináns értéke 0 marad.
Így kapjuk, hogy:

$$\begin{bmatrix} -\lambda^2 R_{xx} & R_{yx} \\ Ryx & R_{yy} \end{bmatrix} = 0 \tag{4.9}$$

Ha a determinánst kifejtjük λ^2-nek egy m_1-ed fokú polinomjához jutunk, így m_1 különböző megoldást kapunk. A legnagyobb érték érdekel bennünket először, ugyanis ez adja a maximális korrelációt: jelöljük λ_1-gyel. Ha ezt a λ_1 becslést beírjuk a (4.7) egyenletbe, és az így adódó homogén egyenleteket megoldjuk, megkapjuk a c és a

d ismeretlenek becsléseit, amelyeket szintén indexszel látunk el c_1 és d_1. A (4.7) egyenletet röviden a következő alakban írhatjuk le:

$$c_1 = \frac{1}{\hat{\lambda}_1} R_{xx}^{-1} R_{xy} d_1 \quad \textbf{és} \quad d_1 = \frac{1}{\hat{\lambda}_1} R_{yy}^{-1} R_{yx} c_1$$

Viszont nem biztos, hogy kielégítik a (4.1) egyenlet szerinti feltételeket ezek a megoldások, mivel egy önkényesen megválasztott skalár (λ_1) is befolyásolja őket. Jelöljük a (4.7) egyenletek megoldásait $\tilde{c_1}$-gyel és $\tilde{d_1}$-gyel. Végezzük el a következő korrekciót :

$$c_1 = \frac{\tilde{c_1}}{(\tilde{c_1} R_{xx} \tilde{c_1})^{\frac{1}{2}}}$$

és

$$d_1 = \frac{\tilde{d_1}}{(\tilde{d_1} R_{xx} \tilde{d_1})^{\frac{1}{2}}}$$

akkor az így kapott c_1 és d_1 már kielégíti a (4.1) egyenletek szerinti feltételeket. A továbbiakban a transzformáció után kapott c_1 és d_1 értékekkel dolgozzunk.
Ezekután a maradék gyökök közül választjuk a legnagyobbat λ_2-t, majd behelyettesítjük a (4.7) egyenletbe és a korrekciót elvégezve kapjuk a hozzá tartozó c_2 és d_2 értékeket, és így általánosságban megkapjuk a λ_i-hez tartozó c_i-t és d_i-t. Jelölje a c_i vektorokat tartakmazó mátrixot C, és D tartalmazza a d_i vektorokat. A kanonikus korreláció-együtthatókat (λ_i) helyezzük el az L diagonális mátrix diagonál elemeibe. Tartalmazzák a nem megfigyelt v és w kanonikus változókat pedig a V és W mátrixok.
Általánosságban a c és d vektorokban lévő súlyokat a következő egyenletekből számíthatjuk ki a (4.7) egyenlet átalakított változatával:

$$C = R_{xx}^{-1} R_{xy} D L^{-1}$$

és

$$D = R_{yy}^{-1} R_{yx} C L^{-1}$$

Ha nagy a változók száma, az előző eljárást nehéz elvégeznünk. Különösen a (4.9) egyenlet determináns kifejtése és megoldása okoz gondot. Ezt a problémát megkerülve nézzünk egy másik eljárást!
Induljunk ki az (4.5) egyenletekből. Szorozzuk be a második egyenlet mindkét oldalát balról $\lambda^{-1} R_{yy}^{-1}$-zel (ha az inverz létezik):

$$d = \lambda^{-1} R_{yy}^{-1} R_{yx} c \tag{4.10}$$

Ezt az elsőbe behelyettesítve:

$$R_{xy} R_{yy}^{-1} R_{yx} c = \lambda^2 R_{xx} c$$

Balról mindkét oldalt szorozva R_{xx}^{-1}-zel (ha az inverz létezik):

$$R_{xx}^{-1}R_{xy}R_{yy}^{-1}R_{yx}c = \lambda^2 c \tag{4.11}$$

Ez a következő alakban írható fel:

$$(R_{xx}^{-1}R_{xy}R_{yy}^{-1}R_{yx} - (-\lambda^2)E)c = 0$$

Látszik, hogy λ^2 megegyezik az $R_{xx}^{-1}R_{xy}R_{yy}^{-1}R_{yx}$ mátrix sajátértékével, c pedig a hozzátartozó sajátvektorral. A maximális kanonikus korrelációt a legnagyobb λ_1 sajátérték adja. Az ehhez tartozó c_1 ismeretében (4.10) egyenletből kiszámolhatjuk d_1-et. Ebben az esetben is ugyanaz a probléma mint elöző esetben. Ahhoz, hogy az (4.1) egyenlet szerinti feltételt teljesíteni tudjuk, normalizálni kell a c_1 és a d_1 vektorokat (hogy $c'R_{xx}c = 1$ legyen). Világos, hogy ha először c-t fejezzük ki az (4.5) egyenletből, nem d-t, az eredmény ugyanaz lesz. Mivel az $R_{xx}^{-1}R_{xy}R_{yy}^{-1}R_{yx}$ mátrix nem szimmetrikus, a sajátérték és a sajátvektor számolásnál ez problémát okozhat. Mivel a legtöbb sajátértéket meghatározó algorimtus feltételezi, hogy a mátrix szimmetrikus, ezért meg kell határoznunk a (4.11) egyenlet egy szimmetrikus alternatíváját. Definiáljuk a következő kisegítő vektort:

$$g = R_{xx}^{\frac{1}{2}}c \quad \text{amelyből} \quad c = R_{xx}^{\frac{1}{2}}q \tag{4.12}$$

Ezt beírva a (4.11) egyenletbe, és balról megszorozva az egyenlet mindkét oldalát $R_{xx}^{\frac{1}{2}}$-vel, olyan egyenlethez jutunk, amelyben a bal oldali mátrix már bizonyíthatóan szimmetrikus:

$$R_{xx}^{\frac{1}{2}}R_{xy}R_{yy}^{-1}R_{yx}R_{xx}^{\frac{1}{2}}q = \lambda_q^2 \tag{4.13}$$

A (4.13) egyenletben található mátrix $(m_1 \times m_1)$-es, m_1 különböző sajátértéke és m_1 különböző sajátvektora van. A Q mátrix ezeket a sajátvektorokat tartalmazza. Mivel a $Q'Q$ szorzatmátrix diagonális lesz, és feltételezve hogy a sajátvektorok egységnyi hosszúságúak, akkor $Q'Q = E$. A (4.12) egyenlet alapján $Q'Q = C'R_{xx}C = E$ amely éppen az (4.1) egyenlet szerinti feltétel kielégítését jelenti. Mivel $V = XC$, a kanonikus változók szórásnégyzete a következőképpen írható fel:

$$\frac{V'V}{n} = C'R_{xx}C = E \tag{4.14}$$

Ebből következik, hogy a V oszlopvektorai ortonormáltak. Tehát a lineáris kombinációval képzett v_i kanonikus változók páronként korrelálatlanok, valamint standardizáltak. Ugyanez belátható W-re is (azaz az y változókból képzett w_i kanonikus változók egymással korrelálatlanok és standardizáltak). A V és W közötti korrelációról:

$$\frac{V'W}{m} = C'R_{xy}D$$

pedig tudjuk látni, hogy diagonális mátrix:(L)
A (4.5) egyenletből:

$$R_{xy}D = R_{xx}CL$$

amit balról beszorozva C'-vel:

$$C'R_{xy}D = C'R_{xx}CL = L \tag{4.15}$$

Ez azt jelenti, hogy minden v_i-hez létezik egy w_j, amellyel maximálisan korrelál, míg a többi w_j-vel korrelálatlan,

$$v_i w_j = \begin{cases} 0 & \text{ha } i \neq j \\ \lambda_i & \text{ha } i = j \end{cases}$$

A kanonikus korrelációs együttható nagyságának vizsgálata mellett fontos a kanonikus változók értelmezése is. A c és d együtthatók ismeretében tudjuk, hogy az eredeti változók milyen súlyú lineáris kombinációi állítják elő a kanonikus faktorokat. Ha kiszámítjuk a faktorelemzésnél jól bevált faktorsúlyok mátrixához hasonló, kanonikus faktorsúlyok mátrixát, akkor a kanonikus faktoroknak könnyen értelmezhető módjához jutunk. A két változóhalmaz között maximális korrelációt adó első kanonikus faktorpárt és az azokat előállító változók közötti korrelációkat tartalmazó kanonikus faktorstruktúrát a következőképpen számíthatjuk ki:
Az első bal oldali kanonikus faktor ($v = Xc$) és a bal oldali változók x közötti korrelációk:

$$s_1 = \frac{1}{n}X'v = \frac{1}{n}X'Xc = R_{xx}c$$

Ily módon az első jobb oldali kanonikus faktor struktúrája:

$$s_2 = \frac{1}{n}Y'w = \frac{1}{n}Y'Yd = R_{yy}d$$

A többi kanonikus faktor esetén is ugyanígy kapjuk meg, hogy az adott kanonikus faktor előállításában melyek a legjelentősebb változók. A kanonikus faktorsúlyok segítségével kiszámíthatjuk, hogy a kanonikus faktorok a változók varianciájának milyen arányát magyarázzák.
Az első bal oldali kanonikus faktor (v) a bal oldali változók varianciájának:

$$\frac{s_1's_1}{m_1}$$

arányát magyarázza.
Ha ezt az arányt megszorozzuk az első kanonikus korreláció négyzetével, akkor a bal oldali változók varianciájának a jobb oldali változók első kanonikus faktora által magyarázott arányát kapjuk:

$$r_x = \frac{s_1's_1}{m_1}\lambda_1^2$$

Ezt nevezzük a bal oldali változóhalmaz redundanciájának az adott jobb oldali változóhalmaz kanonikus faktora esetén. Az első jobb oldali kanonikus faktor a jobb oldali változóhalmaz varianciájának:

$$\left(\frac{s_2' s_2}{m_2}\right)$$

100 %-át magyarázza.
A jobb oldali teljes variancia:

$$r_y = \frac{s_2' s_2}{m_2}\lambda_1^2$$

arányát magyarázza a bal oldali változóhalmaz első kanonikus faktora, és az lesz a jobb oldali változóhalmaz redundanciája az első bal oldali kanonikus faktorra. A két arány, r_x és r_y nem kell hogy megegyezzen. Ha a bal oldali első kanonikus faktor az első főkomponenshez hasonló, a jobb oldali kanonikus faktor pedig a jobb oldali változóhalmaz egy kis varianciájú főkomponenséhez hasonlít, akkor a bal oldali változóhalmazhoz tartozó redundancia nagyobb lesz ($r_x > r_x$)
Mivel több kanonikus faktorpár is számítható, ezért egy bal oldali változóhalmaz teljes redundanciáját, adott jobb oldali változóhalmaz esetén az egy-egy kanonikus faktorpárra számított redundanciák összege adja:

$$r_{d1} = \sum_{i=1}^{m_1} r_{x_i}$$

Ez alapján a jobb oldali változóhalmaz teljes redundanciáját a kanonikus modell szerint az alábbi adja:

$$r_{d2} = \sum_{i=1}^{m_2} r_{y_i}$$

A Bartlett-féle χ^2 próba alapján fogjuk a kanonikus korrelációs együtthatók szignifikancia-próbáját elvégezni. E próba elvégzéséhez fel kell tennünk, hogy x és y többváltozós normális eloszlású valószínűségi-változók.
A próbához definiálnunk kell a Wilks-féle Λ:

$$\Lambda_1 = \prod_{i=1}^{m_2}(1 - \lambda_i^2) \tag{4.16}$$

Nézzük a következő változót (amely Λ függvénye):

$$\chi^2 = -[n - 1 - 0,5(m_1 + m_2 + 1)]ln\Lambda_1 \tag{4.17}$$

amely közelítően χ^2 eloszlású, $(m_1 \times m_2)$ szabadságfokkal.
Az a nullhipotézisünk, hogy x vektorváltozó korrelálatlan y vektorváltozóval. A próbát a szokásos módon végezzük el. Ha elvetjük a hipotézist, akkor az első (maxi-

mális) kanonikus korrelációt hagyjuk el a Λ_1-ből és a maradék (m_2-1) kanonikus korrelációs együttható szignifikanciáját vizsgáljuk. Ekkor az új Λ_2 és χ^2 a következő

$$\Lambda_2 = \prod_{i=2}^{m_2}(1-\lambda_i^2)$$

és

$$\chi^2 = -[n-1-0,5(m_1+m_2+1)]ln\Lambda_2$$

$(m_1-1)(m_2-1)$ szabadságfokkal. Ha Λ-ból $(r-1)$ kanonikus korrelációs együtthatót hagyunk el akkor:

$$\Lambda_r = \prod_{i=r}^{m_2}(1-\lambda_i^2) \tag{4.18}$$

és

$$\chi^2 = -[n-1-0,5(m_1+m_2+1)]ln\Lambda_r \tag{4.19}$$

$(m_1-r+1)(m_2-r+1)$ szabadságfokkal. Összességében megállapíthatjuk, hogy azok a kanonikus korrelációs együtthatók fognak szignifikánsan különbözni 0-tól, amelyeknél elvetettük a nullhipotézist

4.1.1. A függő változók regressziós becslése a kanonikus változók segítségével

Kanonikus változók segítségével két változóhalmaz közötti sztochasztikus kapcsolatot vizsgáltuk. A kapcsolat szorosságán kívül az is érdekel bennünket, hogy a két halmaz közül a függőnek tekintett változóhalmaz hogyan becsülhető a magyarázó változók segítségével. Ezt a fajta regressziós problémát a kanonikus változók felhasználásával oldjuk meg. Tegyük fel, hogy az x változók függetlenek, azaz $R_{xx} = I$ valamint, hogy az összes független változó hatást gyakorol az összes függő változóra. Ebből az következik, hogy a függő változók között érvényesülnek kölcsönhatások. Tehát a kanonikus korrelációs együtthatók és a kanonikus változók becsléseit az előzöek alapján kapjuk meg. A kanonikus változók az előzőek alapján:

$$V = XC \quad \text{és} \quad Y = WD$$

ezenkívül a variancia-kovarianciamátrixok:

$$\frac{V'V}{n} = I, \quad \frac{W'W}{n} = I, \quad V'W = L \tag{4.20}$$

Először a "kanonikus függő" változókat becsüljük a "kanonikus független" változók segítésével.
$W = VB$ ahol B becslését a legkisebb négyzetek módszerével kaphatjuk meg.

$$\hat{B} = (V'V)^{-1}V'W = L \tag{4.21}$$

$$\hat{W} = VL$$

Ezután V kanonikus változókkal figyeljük az Y megfigyelt függő változókat:

$$\hat{Y} = VA$$

Az A becslését a legkisebb négyzetek módszerével határozzuk meg, és felhasználjuk a (4.20) egyenletben megadott feltételeket:

$$\hat{Y} = XC(C'X'XC)^{-1}C'X'Y$$

$$\hat{Y} = XC(nE)^{-1}C'R_{xy}n \tag{4.22}$$

$$\hat{Y} = XCC'R_{xy}$$

A (4.7) egyenletből:

$$R_{yx}C = R_{yy}DL$$

Ha transzponáljuk, behelyettesíthetünk vele a (4.22) egyenletbe:

$$C'R'_{yx} = L'D'R_{yy}$$

Így a kanonikus korrelációs együttható, a kanonikus változók, a kanonikus súlyok segítségével megadtuk az y függő változók regressziós becslését. Megmutatható, hogy ez megegyezik a legkisebb négyzetek módszerével közvetlenül kapható becsléssel.

5. fejezet

SEM módszer

A fejezet az [5] cikk alapján íródott. A SEM modell egy olyan statisztikai technika mely a path analízisből ered, amelynek lényege, hogy a különböző változók között ok-okozati viszonyt tételezünk fel, és ez alapján írunk fel regressziós egyenleteket, amelyek összekapcsolják őket. A változók közötti kapcsolatok szemléltetésére egy irányított gráfot rajzolunk fel, melyben a csúcsok a változók és a köztük futó irányított élek a regressziós együttható. Tehát a SEM ennek a modellnek az egyenletekkel felírt rejtett változókat is tartalmazó továbbfejlesztése. Azonban ez a módszer sokkal jobb mint a szokásos latens változós eljárások, mivel itt a rejtett változók közötti strukturális viszonyt is felírhatjuk, és a modell illesztésekor ezt is figyelembe tudjuk venni. Ennek következtében az egyenleteket és a változókat is két csoportba sorolhatjuk. A változók lehetnek exogén (vagy külső) és endogén (vagy belső) változók. Az exogén változók alatt olyan változókat értünk, melyekre nincs másik olyan latens változó, amely rájuk közvetlen hatással lenne, más szóval ezek a magyarázó változók. Endogén változók alatt olyan változókat értünk melyeket más latens változók magyaráznak. Az egyenletek azon csoportját fogjuk strukturális egyenleteknek nevezni amelyek a látens változók közötti viszonyt írják le, míg mérési egyenleteknek azokat amelyek a mért és a latens változók kapcsolatát írják le.

5.1. A latens változós modell

A modelltől a következőket követeljük meg:

- minden változó 0 várható értékű

- a hibák egymástól és a latens változóktól is függetlenek

- $diagonlis(B) = 0$

Legyen η egy oszlopvektor amely m exogén latens változóból áll és ξ egy olyan oszlopvektor amely k endogén centrált latens változóból áll.
Ekkor a struktúrált latens változós modell a következőképpen írható fel:

$$\eta = B\eta + \Gamma\xi + \zeta$$

ahol B egy $m * m$-es regressziós együtthatókból álló null-diagonálisú mátrix.
Γ egy $m * k$ mátrixa a regressziós együthatóknak,
ζ egy m dimenziós vektor.
Mind az exogén, mind az endogén változókhoz tartozik egy egy mérési modell:

$$y = \Lambda_y\eta + \epsilon$$

$$x = \Lambda_x\xi + \Sigma$$

Ezért ez a három egyenlet együttesen a SEM modell. Ha nem nulla várható értékű esetet szeretnénk, akkor egyszerűen minden egyenletben hozzá kell adnunk a jobb oldalhoz a bal oldal várható értékét. A modellünk ezen felírása még elég általános, hiszen itt csak az x és y, és az ő tapasztalati kovariancia mátrixukat ismerjük, és nagyon sok ismeretlen változónk van. Ellenben a $\Lambda_x, \Lambda_y, \Gamma$ és B együttható mátrixoknak speciális alakja van, amit mindig az adott modell felírás határoz meg. Ezeknek a mátrixoknak az elemei a legtöbb esetben zérusok, mivel azt előre meghatároztuk, hogy mely mért változókra mely latens változók hatnak. Ez a modell általános esetként magában foglal más eljárást is például a többváltozós regressziót.

A megfigyelt változók kovariancia mátrixának faktorizációja

Legyen

$$\begin{aligned}
\Phi &= Cov(\xi) = E(\xi\xi') \\
\Psi &= Cov(\zeta) = E(\zeta\zeta') \\
\Theta_\epsilon &= Cov(\epsilon) = E(\epsilon\epsilon') \\
\Theta_\delta &= Cov(\delta) = E(\delta\delta')
\end{aligned}$$

A faktoranalízishez hasonlóan, a SEM-et is a a tapasztalati kovarianciamátrix aproximációjával számoljuk. Az a célunk, hogy a modell által számolt kovarianciamátrix minél jobban közelítse a minta alapján számolt tapasztalati kovarianciamátrixot. A tapasztalati kovarianciamátrixot definiálljuk a következőképpen :

$$\mathbf{\Sigma} = \begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix}$$

A fenti egyenletből az egyes részeket már könnyen ki tudjuk számolni. I legyen egységmátrix és $(I - B)$ invertálható.

$$\Sigma_{xx} = \Lambda_x \phi \Lambda_x' + \Theta_\delta$$
$$\Sigma_{yy} = \Lambda_y[(I-B)^{-1}(\Gamma\phi\Gamma' + \Psi)][\Lambda_y[(I-B)']^{-1}\Lambda_y + \Theta_\epsilon$$
$$\Sigma_{xy} = \Lambda_x \phi \Gamma'[(I-B)']^{-1}\Lambda_y$$

Amiből megkaphatjuk, hogy:

$$\mathbf{\Sigma} = \begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix} =$$

$$= \begin{bmatrix} \Lambda_y[(I-B)^{-1}(\Gamma\phi\Gamma' + \Psi)][\Lambda_y[(I-B)']^{-1}\Lambda_y + \Theta_\epsilon & \Lambda_y[(I-B)']^{-1}\Gamma\phi\Lambda_y' \\ \Lambda_x\phi\Gamma'[(I-B)']^{-1}\Lambda_y & \Lambda_x\phi\Lambda_x' + \Theta_\delta \end{bmatrix}$$

Az egyenletek megoldására explicit képlet csak nagyon ritkán adódik mivel annak az a feltétele, hogy pontosan ugyanannyi egyenlet legyen, mint paraméter. Két fő módszer létezik a SEM modellek illesztésére, az iteratív és a Bayesi eljárás.
Az iteratív eljárások számítják például a legkisebb négyzetes becslést.

5.2. Iteratív eljárás

Többféle program létezik amely SEM illesztésre alkalmas (AMOS, LISREL). Ezek a programok iteratív módszerrel számolnak, de a konkrét számolási módok eltérőek. Röviden ismertetve a módszert, első lépésként be kell állítanunk egy kezdőértéket minden ismeretlennek, ezek után ki kell számolni a modellből a kovarianciamátrixot. Ennek a mátrixnak és a tapasztalati kovarianciamátrixnak valamely függvénye fogja megadni a mátrixok közötti távolságot. Ezután ki tudjuk számolni a kezdeti értékben a távolságfüggvény parciális deriváltjait minden ismeretlen szerint, és az értékeiket ez alapján megváltoztatjuk.
Az illeszkedés jóságának mérésére leggyakrabban három mérőszámot használunk. Ezek a legkisebb négyzetes eltérés, az általánosított legkisebb négyzetes eltérés és a maximum likelihood mérőszám. Jelöljük S-sel a tapasztalati kovarianciamátrixot és Σ-val a modellből számolt kovarianciamátrixot, melyek mérete $m * m$.

$$\begin{aligned} OLS &= tr(S-\Sigma)^2 \\ GLS &= \frac{1}{2}[(S-\Sigma)S^{-1}]^2 \\ ML &= log[det(\Sigma)] - log[det(S)] + trS(\Sigma)^{-1} + m \end{aligned}$$

az egyenleteinkben *tr* jelöli a nyomoperátort, *det* a deteminánst és *log* a természetes alapú logaritmust.
Ezekenek a mérőszámoknak két fő célja van, egyrészt a keresést segítsék az algoritmus minden lépésében másrészt a kapott eredményt értékeljék.

Illeszkedés-vizsgálat

Az eljárás végén szükségünk van valamilyen mutatószámra, amely meghatározza, a modell illeszkedés-jóságát. Ilyen szempontból többféle mutatószám áll a rendelkezésünkre, amely mind támpontot adnak arra, hogy mennyire elfogadható a modellünk. Fontos azonban figyelembe venni, hogy alternatív modellek illeszkedését is meg tudjuk vizsgálni. Ha egy adott modellt elfogadtunk, attól még nem zárhatjuk ki, hogy létezik egy másik modell, ami az előzőnél lényegesen jobban írja le az adatainkat. Ebben az esetben mindkét modellre kiszámoljuk az illeszkedési mutatókat. Létezhetnek olyan esetek amikor két modell összehasonlítására van lehetőségünk. Akkor beszélünk ilyen modellekről, ha az egyik modell megkapható a másikból úgy, hogy ez utóbbiban egyes paraméterek értékeit rögzítjük. Ennek segítségével meg tudjuk vizsgálni, hogy a regressziós együttható fontos része e a modellünknek. Az illeszkedés jóságának egyik lehetősége, hogy a négyzetes eltérések összegét vizsgáljuk, ennek a neve Goodness-of-Fit Index, röviden GFI. Különösképpen akkor hasznos ez az eljárás, ha a mátrixok elemei nem azonos skálán mozognak, vagy eltérő nagyságrendűek.

$$GFI = 1 - \frac{||S - \sum(\hat{\Lambda_x}, \hat{\Lambda_y}, \hat{B}, \hat{\phi}, \hat{\Psi}, \hat{\Gamma}, \hat{\Theta_\epsilon}, \hat{\Theta_\delta}||^2}{||S||^2}$$

A modell akkor fogadható el, ha a GFI nagyobb mint 0.9.

6. fejezet

PLS regresszió

Az eljárás pontos és precíz megértését valamint leírását a [7], [8], [9] és [10] cikkek segítették. Regressziószámítás során két vagy több véletlen változó között fennálló kapcsolatot modellezzük, valamint több jellemző által az eredményváltozóra gyakorolt hatását vizsgáljuk. A regressziós egyenletben a magyarázandó vagy más néven célváltozót (Y)-t a magyarázó változók vagy regresszorok (X) mint független változók segítségével magyarázzuk. A regressziós modellek szerkesztésekor a legelső feladat, hogy megkeressük azokat a változókat, amelyek az eredményváltozóval szignifikáns kapcsolatban vannak. Az eljárás során arra próbálunk választ adni, hogy a független változók egységnyi változása, a függő változó milyen mérvű megváltozását vonhatja maga után. A regressziós egyenlet fontos része a maradék vagy más néven rezidum vagyis a modellünk által nem magyarázott rész.
A parciális legkisebb négyzetek elve egy módszer, amellyel regressziós egyenlet együtthatóit becsüljük, melyek mellett a megfigyelésből származó és a regressziófüggvény alapján becsült Y értékek különbségének eltérésnégyzet-összege a legkisebb. Következésképpen a kapott együtthatók nem valódi mért adatok együtthatói azokat ugyanis nem ismerjük. Az együtthatókból a regressziós egyenlet segítségével kiszámolhatjuk az eredeti adatokat, vagyis megnézhetjük, hogy az ismert független változókhoz az egyenlet alapján milyen függő változóbeli értékek tartoznak. A PLS főként akkor hasznos paraméteres egyenletek konstruálására amikor sok magyarázó változó és viszonylag kevés mintaadat van. Ahhoz, hogy meg tudjuk határozni a kapcsolatot az Y változó, és az $X_1, X_2, \ldots X_m$ magyarázó változók között a PLS során latens változót konstruálunk, úgy hogy mindegyik latens változó az $X_1, X_2, \ldots X_m$ változók kombinációja. A módszer hasonlít a főkomponens analízishez azzal a különbséggel, hogy míg ott a főkomponenseket kizárólag az X változó adatértékei határozzák meg, addig a PLS-ben X és Y adat értékei is befolyásolják azt. A fő célja a PLS-nek olyan komponensek létrehozása, amelyekkel a lehető legtöbb információt nyerjük ki az X

változokból ez pedig az Y változók minél pontosabb becslését eredményezi.

6.1. Az általános modell

Legyenek:

- $Y_1, Y_2, \ldots, Y_l$ a magyarázott változók
- $X_1, X_2, \ldots, X_m$ a magyarázó változók

A parciális regressziós modellt az alábbi matematikai egyenlettel írjuk fel:

$$Y = \beta_0 + \beta_1 T_1 + \beta_2 T_2 + \cdots + \beta_p T_p + \epsilon$$

ahol:

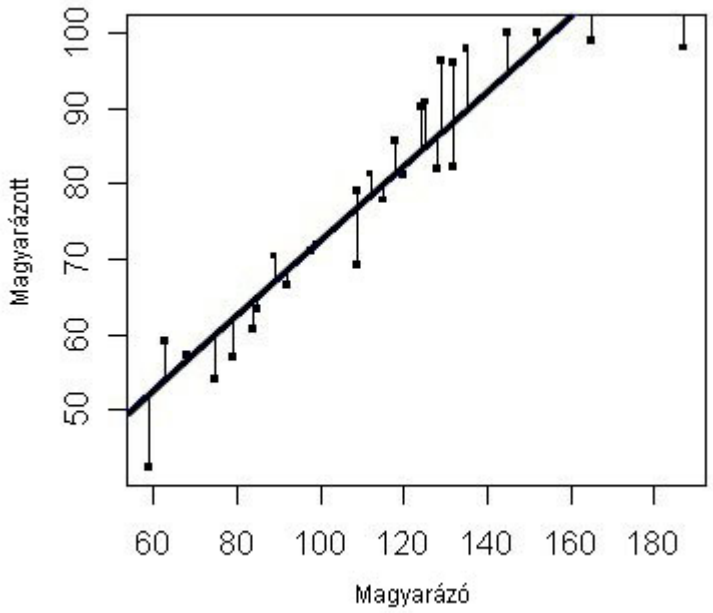
- $\beta_1, \beta_2, \ldots \beta_p$ az együtthatók
- T_k az X_j lineáris kombinációja ahol $k = 1 \ldots p \quad$ és $\quad p \leq m$
- β_0 a függvény konstans tagja
- ϵ a regressziós hibatag

Egy egyenlet minél több paramétert foglal magában, annál könnyebben befolyásolható véletlen hibákkal. Ezért a regressziós módszereknek az a célja, hogy csökkentsék a tagok számát a regressziós modellben.

A $\beta_1, \beta_2, \ldots \beta_p$ parciális regressziós együtthatókat a következőképpen értelmezzük: ha T_i értékét egy egységgel növeljük miközben a többi értékét változatlan hagyjuk akkor az eredményváltozó becsült értéke éppen B_i egységgel változik. A regressziós együttható tehát kifejezi, hogy egy adott latens változó egységnyi növekedése mekkora növekedést vagy csökkenést okoz az eredményváltozó becsült értékében, miközben a többi tényező változó értéke változatlan. Fő feladatunk az ϵ hibatag minimalizálása, amit akkor érünk el, ha a becslő függvény értékei minimálisan térnek el az eredeti tapasztalati értéktől.
A legkisebb négyzetek módszere szerint minimalizálnunk kell:

$$\sum_{i=1}^{m} e^2 = \sum [Y - (\beta_0 + \beta_1 T_1 + \beta_2 T_2 + \cdots + \beta_p T_p)]^2 \to min$$

A többváltozós függvények illesztésének pontosságát a regressziós függvény hibájának nagysága alapján ítélhetjük meg.



Az illeszkedés hibája S_e

$$S_e = \sqrt{\frac{\sum e^2}{m-p}}$$

Az illeszkedés relatív hibája

$$V_{S_e} = \frac{S_e}{\hat{Y}} * 100$$

A relatív hiba azt fejezi ki, hogy a számított y_i értékek azaz a regressziós becslések, átlagosan hány százalékkal térnek el az eredeti eredményváltozó mért Y_i értékeitől.

6.2. Az egyváltozós PLS

Adott n darab megfigyelés, ekkor:
Y jelölje a megfigyelésünkből származó egy darab magyarázott változót,
$X_1, X_2 \dots X_m$ jelölje az m darab magyarázó változót.
A komponensek között a korreláció 0.

Az i.-dik adatot jelölje az $[X_1(i), X_2(i), \dots X_m(i); Y_i]$
x_j és y pedig jelölje az X_j és Y megfigyelt vektorok értékét.

$$y = [y(1), y(2) \dots y(n)]$$

$$x_j = [x_j(1), x_j(2), \dots x_j(n)] \quad j = 1 \dots m$$

A minta átlaga $\bar{Y} = \sum_i^n \frac{y(i)}{n}$ és $\bar{X_j} = \sum_i^n \frac{x_j(i)}{n}$.
A minta alapján vett regressziós egyenletet a következőképpen írhatjuk fel:

$$\widehat{Y} = \beta_0 + \beta_1 T_1 + \beta_2 T_2 + \dots + \beta_p T_p \tag{6.1}$$

A könnyebb jelölés érdekében vezessük be a következő két változót:

$$U_1 = Y - \bar{Y}$$

$$V_{1j} = X_j - \bar{X_j} \quad j = 1 \dots m \tag{6.2}$$

A minta átlaga U_1-re és $V_1(j)$ -re 0. A komponenseket egymás után határozzuk meg, ezen konstrukció alatt a korreláció a $V_1(j)$-k között elhanyagolható.
Az első komponens amit meghatározunk T_1 a $V_1 j$ vektor lineáris kombinációja. Mivel a minta átlaga 0, ezért a regressziós egyenletet a következő alakban írhatjuk fel:

$$\widehat{U_1(j)} = b_{1j} V_{1j} \quad \text{ahol} \quad b_{1j} = \frac{v'_{1j} u_1}{v'_{1j} v_{1j}} \tag{6.3}$$

V_{1j} adott értékei az m egyenlet mindegyikében biztosítják U_1 becslését, tehát ezeket a becsléseket összegeznünk kell.
Első lépésként ehhez vegyük a súlyozott átlagot:

$$\sum_j w_{ij} = 1.$$

Ekkor:

$$T_1 = \sum_j w_{ij} b_{1j} V_{1j} = \sum_j w_{ij} \widehat{U_1(j)}. \tag{6.4}$$

Az X változó potenciálisan további információkat hodoz magában az Y-nal kapcsolatban, viszont az X_j-ben lévő információk nincsenek benne T_1-ben, ezeket az információkat az X_j-nek a T_1-en végzett regressziós maradékával becsülhetjük meg, ami azonos a V_{1j}-nek a T_1-en végzett regressziós maradékával.

Ezen maradékok segítségel ki tudjuk számolni V_{2j}-t V_{1j} által és $U_{2(j)}$-t $U_{1(j)}$ által. A következő komponens, amit megtudunk határozni az T_2, amely a lineáris kombinációja V_{2j}-nek, ugyanakkor U_2 meghatározásában is segít.
Tehát a fenti eljárás segítségével találhatunk egy természetes módot a $T_1, T_2, \dots T_p$ komponensek meghatározására ahol mindegyik komponens meghatározható a regressziós maradékokból és a korábbi komponensek segítségével.
Ahhoz, hogy megtudjuk határozni $T_{(i+1)j}$ ahogy azt az előbb már láthattuk először $V_{(i+1)j}$-t és U_{i+1}-t kell meghatároznunk. $V_{(i+1)j}$ meghatározásához T_i-szer kell regresszállnunk V_{ij}-t az új regressziós együtthatóval. Vagyis:

$$V_{(i+1)j} = V_{ij} - \Big(\frac{v_{ij} t'_i}{t'_i t_i}\Big) T_i \tag{6.5}$$

Hasonlóan definiálhatjuk U_{i+1}-et is:

$$U_{i+1} = U_i - \Big(\frac{u_i t'_i}{t'_i t_i}\Big) T_i$$

tehát U_{i+1}-et úgy kapjuk meg ha T_i-szer regresszálljuk U_i-t az új regressziós együtthatóval.
Az X_j-ben lévő "információs maradék"-ot $V_{(i+1)j}$, a j-dik regressziós hozamot pedig $V_{(i+1)j}b_{(i+1)j}$ jelöli ahol:

$$b_{(i+1)j} = \frac{v'_{(i+1)j}u_{i+1}}{v'_{(i+1)j}v_{(i+1)j}} \tag{6.6}$$

Ezeknek a paramétereknek a lineáris kombinációját véve kapjuk meg a következő komponenst:

$$T_{i+1} = \sum_j w_{(i+1)j}b_{(i+1)j}V_{(i+1)j} = \sum_j w_{(i+1)j}\hat{U}_{i+1}(j) \tag{6.7}$$

Ha ezt a módszert megismételjük akkor már meg tudjuk határozni $T_{i+2}, T_{i+3}, \ldots T_p$.
Miután megkaptuk az összes komponenst és beírtuk őket a (6.1) egyenletbe, megkapjuk a regressziós modell egy becslését.
Egy ismert tulajdonsága a PLS-nek, hogy a komponensek között a korreláció 0. Ennek oka, hogy:

- (a) $T_{i+1} \ldots T_p$ komponensek a lineáris kombinációi a $V_{(i+1)j}$-nek,
- (b) a regressziós maradékok korrelálatlanok a regresszorral
  és ezekből következik, hogy korrelálatlanok a T_i-vel.

A komponensek korrelálatlanságának következménye, hogy a (6.1) egyenletben a regressziós együttható egy egyszerű egyváltozós regresszorral becsülhető. További következménye még, hogy u_{i+1} és $v_{(i+1)j}$ a megfelelő vektora az X_j-nek a $T_1, T_2, \ldots T_i$-n végzett regressziós maradékához, ami egyben le is egyszerűsíti az értelmezését U_{i+1}-nek és $V_{(i+1)j}$-nek. Miután meghatároztuk a regressziós modell egy becslését a (6.2),(6.5),(6.7) egyenletet felhasználva ki tudjuk fejezni az eredeti változókkal X_j-t. Így tehát egy sokkal alkalmasabb egyenletet kapunk Y becslésére, és ez további mintákat eredményez X értékei alapján.

6.3. A többváltozós PLS

Többváltozós PLS esetén adott a megfigyelésünkből származó l db magyarázott változó $Y_1, Y_2, \ldots Y_l$ és m db magyarázó változó $X_1, X_2 \ldots X_m$.
A többváltozós eljárás célja, megkeresni azokat a magyarázó változókat, melyek a magyarázott Y változók minél pontosabb becslését eredményezik.
A modellt a következőképpen írhatjuk fel:

$$\hat{Y_k} = \beta_{k0} + \beta_{k1}T_1 + \beta_{k2}T_2 + \cdots + \beta_{kp}T_p \quad k = 1 \ldots l \tag{6.8}$$

ahol mindegyik $T_1, T_2, \ldots T_p$ komponens az X változók lineáris kombinációja. Fontos megjegyezni, hogy azonos komponensek előfordulhatnak a modellben minden Y változóra, csak a regressziós együtthatók változnak.
A célunk az, hogy konstruáljunk egy olyan algoritmust, ami világossá teszi a hasonlóságot és a különbséget az egyváltozós és a többváltozós PLS között.
Az X változóra használjuk ugyanazt a jelölést, mint korábban, az Y változóra pedig vezessük be a következőt:

$$\begin{aligned} Y_k &= [Y_k(1), \ldots Y_k(n)] \quad k = 1 \ldots l \\ X_j &= [X_j(1), X_j(2), \ldots X_j(n)] \quad j = 1 \ldots m \end{aligned}$$

A minta átlaga: $\bar{Y} = \sum_i^n \frac{y_k(i)}{n}$ és $\bar{X_j} = \sum_i^n \frac{x(i)}{n}$.
A könnyebb számolás érdekében, legyen:

$$R_{1k} = Y_k - \hat{y_k} \tag{6.9}$$

$$V_{1j} = X_j - \bar{X_j} \quad j = 1 \ldots m.$$

Az első komponens meghatározásához, definiálnunk kell a következő két mátrixot:

$$\begin{aligned} V_1 &= (v_{11}, \ldots v_{1m}) \\ R_1 &= (r_{11}, \ldots r_{1l}) \end{aligned}$$

Az $R_1' V_1 V_1' R_1$ mátrix legnagyobb sajátértékéhez c_1 legyen a megfelelő sajátvektor, és u_1-et definiáljuk úgy mint:

$$u_1 = R_1 c_1 \tag{6.10}$$

Ezután T_1-et könnyen megtudjuk konsturálni $[u_1, v_{11}, v_{12} \ldots v_{1m}]$-ből ugyanúgy, mint az egyváltozós esetben.
U_1 konstruálásának ezen módját Hoskuldsson mutatta meg [7], felhasználva, hogy ha veszünk két egységhosszú vektort f és g, és maximalizáljuk a $[c\hat{o}v(V_1 f, R_1 g)]^2$, akkor $R_1 g$ megfelel u_1-nek.
Ahhoz, hogy meg tudjuk határozni T_i-t, V_{ij}-t és R_{ij}-t meg kell adnunk az algoritmus általános lépését. Az általános lépés megadása után már meg tudjuk mutatni, hogy hogyan kapjuk meg ezeket a komponenseket az i-dik lépés után, vagyis hogyan tudjuk meghatározni $i \to i+1$-et.
Először is, $V_{(i+1)j}$ azon regressziós maradék, amit akkor kapunk ha V_{ij}-t regresszáljuk T_i-n tehát $V_{(i+1)j}$-t a (6.5) egyenlet alapján lehet kiszámolni. Hasonlóan $R_{(i+1)j}$ az a maradék amit akkor kapunk ha R_{ij}-t regresszáljuk T_i-n ez alapján:

$$R_{(i+1)k} = R_{ik} - \Big(\frac{t_i' r_{ik}}{t_i' t_i}\Big) T_l \quad \text{ahol } r_{ik} \text{ a minta értéke}$$

Világos, hogy amikor Y_k-t regresszálljuk a $T_1, T_2, \ldots T_i$-n akkor $r_{(i+1)k}$ egyben a maradék is. Legyen:

$$R_{i+1} = (r_{(i+1)1}, r_{(i+1)2} \ldots r_{(i+1)l})$$

$$V_{i+1} = (v_{(i+1)1}, v_{(i+1)2} \ldots v_{(i+1)m})$$

c_{i+1} legyen az $R'_{i+1}V_{i+1}V'_{i+1}R_{i+1}$ mátrix legnagyobb sajátértékéhez a megfelelő sajátvektor, és u_{i+1}-et definiáljuk úgy mint:

$$u_{i+1} = R_{i+1}c_{i+1}$$

T_{i+1}-et ugyanúgy tudjuk meghatározni, mint az egyváltozós esetben, felhasználva a (6.6) és (6.7) egyenleteket. Miután meghatároztuk a $T_1, T_2, \ldots T_p$ komponenseket, mindegyik Y változót regresszálljuk újra külön-külön, és ezek a komponensek fogják a β együtthatókat megbecsülni a (6.8) egyenletben.
Az, hogy pontosan hány latens változót érdemes létrehoznunk azt a cross validation eljárás segítségével tudjuk meghatározni.

A következő lépés, hogy megmutatjuk, a korábbi algoritmus megegyezik a többváltozós PLS standard verziójával.
Jelölje V_1 és R_1 az Ω_1 és Φ_1-el centrált adatmátrixokat, azon feltétel mellett, hogy Ω_i és Φ_i már meg van határozva.

1. Legyen ϕ a Φ_i első oszlopa
2. Legyen $\psi = \frac{\Omega'_i\phi}{\phi'\phi}$ ahol ψ arányos az egységhosszal
3. $\tau = \Omega_i\psi$
4. Legyen $\zeta = \frac{\Phi'_i\tau}{\tau'\tau}$ és ζ arányos az egységhosszal
5. Legyen $\phi = \Phi_i\zeta$ ha ez konvergens, akkor következzen a 6. lépés különben pedig újra a 2. lépés.
6. $\theta = \frac{\Omega'_i\tau}{\tau'\tau}$
7. $\lambda = \frac{\tau'\phi}{\tau\tau'}$
8. maradék mátrix: $\Omega_{i+1} = \Omega_i - \tau\theta'$ és $\Phi_{i+1} = \Phi_i - \lambda\tau\zeta'$

Tegyük fel, hogy $\Omega_i = V_i$ és $\Psi_i = R_i$, és mivel w_{ij} a (6.7) egyenletben választott súly, ezért $w_{ij} = v'_{ij}v_{ij}$.
Be kell látnunk, hogy:

- a) $\tau \propto t_i$

– b) $\Omega_{i+1} = V_{i+1}$

– c) $\Phi_i = R_{i+1}$

a) bizonyítása: Hoskuldsson mutatta meg, hogy ha konvergencia van az 5. lépésben, akkor ζ egy megfelelő sajátvektor a $\Phi_i'\Omega_i\Omega_i'\Phi_i$ mátrix legnagyobb sajátértékéhez. A feltétel szerint $\Phi_i'\Omega_i\Omega_i'\Phi_i = R_i'V_iV_i'R_i$, ezért ζ arányos lesz c_i-vel, vagyis az 5. lépésből és a (6.10) egyenletből következik, hogy $\theta \propto u_i$.
A konvergencia után mindenféle következmény nélkül, megismételve a 2-5 lépést, arra a következtetésre jutunk, hogy:

- 2. lépésből következik, az $\psi \propto \Omega_i'\phi \propto V_i'u_i$,
- a 3. lépésből következik, hogy $\tau \propto V_iV_i'u_i$.
- a (6.6) egyenlet alapján a j-dik komponense a $V_i'u_i$-nek $w_{ij}b_{ij}$,
- tehát a (6.7)-ből $t_i = V_iV_i'u_i$.

Ezek alapján kapjuk, hogy: $\tau \propto t_i$.

b) bizonyítása:

- A 6. és 8. lépésből következik, hogy $\Omega_i - \Omega_{i+1} = \tau\theta' = \frac{\tau\tau'\Omega_i}{\tau'\tau} = \frac{t_it_i'V_i}{t_i't_i}$ mivel $\tau \propto t_i$ és $\Omega_i = V_i$
- A j.-dik oszlopa a $\frac{t_it_i'V_i}{t_i't_i}$ mátrixnak $\frac{t_it_i'v_{ij}}{t_i't_i} = \frac{t_i(t_i'v_{ij})}{t_i't_i}$
- A (6.5) egyenletből és a legutóbbi kifejezésből ez egyenlő lesz $v_{ij} - v_{(i+1)j}$

Ezért $\Omega_i - \Omega_{i+1} = V_i - V_{i+1}$

c) bizonyítása:

- Legyen $\zeta = \frac{K\phi_i'\tau}{\tau'\tau}$ ahol K egy konstans melyre $\zeta'\zeta = 1$.
- Ezek után az 5. és a 7. lépésből $\lambda = \frac{\tau'\phi}{\tau'\tau} = \frac{\tau'\Phi_i\zeta}{\tau'\tau} = \frac{\zeta'\zeta}{K} = \frac{1}{K}$
- Tehát $\lambda\zeta'Z\frac{\zeta'}{K} = \frac{\tau'\Phi_i'}{\tau'\tau}$.
- A 8. lépésből $\Phi_i - \phi_{i+1} = \lambda\tau\zeta'$, tehát $\Phi_i - \phi_{i+1} = \frac{\tau\tau'\Phi_i}{\tau'\tau} = \frac{t_it_i'R_i}{t_i't_i}$.

A (6.9) egyenletből az utóbbi kifejezés egyenlő $R_i - R_{i+1}$.

Abban az esetben ha több Y változónk van és csak többváltozós PLS-t tudunk használni, egy alternatíva az, hogy többször megismételjük az egyváltozós PLS-t. Mindegyik Y változóra egymás után kerül rá a sor, és a mintaértékek valamint a magyarázó változók alapján határozzuk meg a regressziós egyenletet. Összehasonlítva az egyváltozós és a többváltozós PLS-t, azt reméljük, hogy a regressziós egyenletet az egyik magyarázott változó $\widetilde{Y}$ határozza meg, és ez meghatároz egy olyan analógiát, ami alapján meg tudjuk konstruálni T_{i+1}-et miután a $T_1, T_2, \dots T_p$-et már meghatároztuk.
Mindegyik módszer T_{i+1}-et u_{i+1}-ből és $v_{(i+1)j}$-ből határozza meg, ahol $v_{(i+1)j}$ az X_j-nek a $T_1, T_2, \dots T_l$-en végzett regressziós maradéka. Az egyetlen különbség a két módszernél az u_{i+1} konstruálásában van. Az egyváltozós PLS esetén az u_{i+1} az a maradék, amit akkor kapunk ha $\widetilde{Y}$ -ot regresszáljuk a $T_1, T_2, \dots T_l$, amíg a többváltozós esetben mindegyik Y_k-t külön külön regresszáljuk $T_1, T_2, \dots T_i$-n és az u_{i+1} ezen maradékok lineáris kombinációja, ahogy azt a (6.10) egyenlet is mutatja.
Az egyváltozós és a többváltozós PLS közötti választás ekvivalens azzal, hogy meghatározzuk hogy az u_{i+1}-et milyen úton szeretnénk megkapni, bár azt gondolnánk, hogy a többváltozós PLS során több információt használunk fel, mint az egyváltozós esetben, de valójában azonos mértékű információt használunk, csak az algoritmus más más lépéseiben.

7. fejezet

A módszerek bemutatása R program segítségével

A szakdolgozatomban tárgyalt eljárásokat az R program segítségével szeretném bemutatni. A program nagy előnye, hogy bárki számára elérhető és letölthető, valamint az alkalmazáshoz szükséges függvényeket megtaláljuk benne. A módszer interpretációjára azokat az adatokat használjuk fel, amiket a Kiskörei tározó területén az Óhalászi-Holt-Tisza vízminőségével kapcsolatban az ezeredforduló környékén mértek. Ezekre az adatokra a figyelmet Márialigeti Károly [13] hívta fel. Az adatok Teszárné Nagy Mariann [12] dolgozatában érhetőek el.
Az Óhalászi-Holt-Tisza egy, a Tisza szabályozásakor levágott folyó kanyarulatból keletkezett. A folyó vízével közvetlenül csak magasabb vízállás mellett érintkező állóvíz. Vizsgálata a Tisza élővilágának megértése szempontjából igen fontos eszköz. Emiatt is került ismételten a kutatók (Dévai György [14]) érdeklődésének középpontjába.
Az élővíz nem homogén. Hőfoka, fizikai, biológiai, kémiai állapota rétegesen változik. A változás módja évszakonként is különböző. Az idézett kutatók — igen nagy gondossággal, sok év kitartó munkájval — egy olyan referencia méréssort állítottak elő ami 25-50 centis rétegenként különböző időpontokban, pontosan ugyanazon a helyen vett minták alapján, a holtág vízminőségét mutatják. Jelen feldolgozás csak interpretációs céllal készült, a nyomtatásban megjelent adatok kézzel való beirása alapján. Vagyis az adatok ebben a formában nem tekinthetőek hitelesnek. Emiatt az eredmények közvetlen hirológiai vonatkoztatása hibás volna. De az eredmények interpretációs jellegének több további fontos oka is van.
A rendelkezésre álló adatsorok hiányosak a rendkívül gondos gyüjtés ellenére. A kutatók alkalmanként és rétegenként 32 változót vizsgáltak.
Ezt és az adatsorok hiányosságát az alábbi táblázat szemlélteti:

11.	2000.06.14															
1	vizmelyseg	20	100	150	200	225	250	275	300	325	350	375	400	430	470	500
2	vizhomerseklet	27,8	26,5	24,5	22,3	-	19,8	-	18,3	-	17,3	-	16,4	15,5	-	14,9
3	viz_szine	sb	sb	sb	sb	-	sb	-	zzs	-	zzs	-	zzs	zzs	-	zzs
4	viz_szaga	sztlan	sztlan	sztlan	sztlan	-	sztlan	-	ztsz	-	ztsz	-	ztsz	ztsz	-	ztsz
5	fenytranszmisszio	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
6	oldott_oxigen	6,2	3,4	1,4	0,6	-	0,1	-	<0,02	-	<0,02	-	<0,02	<0,02	-	<0,02
7	oxigentelitettseg	80	43	177	<2	-	<2	-	<2	-	<2	-	<2-	<2	-	<22
8	KOI	5,7	6,2	6,6	5,6	-	5,9	-	6,2	-	6,9	-	7,1	8,1	-	14
9	elektr_potencial	381	383	389	391	-	405	-	411	-	160	-	130	69	-	77
10	pH	7,3	7,3	7,1	7	-	7	-	7,1	-	7,1	-	7,1	7,1	-	7,1
11	fajlagos_vezkep	355	354	347	344	-	352	-	362	-	357	-	365	382	-	432
12	szabad_szendioxid	3,7	8,4	12	8,4	-	15	-	12	-	15	-	12	14	-	22
13	kalcium_ion	42	43	44	44	-	44	-	43	-	43	-	44	47	-	51
14	magnezium_ion	12	10	8,7	11	-	10	-	13	~	20	-	18	19,9	-	19
15	ossz_kemenyseg	86	84	82	S8	-	86	-	90	-	106	-	104	112	-	116
16	szulfat_ion	29	28	24	33	-	30	-	29	-	24	-	23	19	-	7
17	oldott_szulfidok	0,002	0,002	0,002	0,002	-	0,177	-	0,161	-	0,318	-	0,353	0,499	-	0,751
18	ammoium_ion	0,03	0,04	0,03	0,04	-	<0,03	-	<0,03	-	<0,03	-	0,5	1,6	-	2,9
19	ammoium_N	0,02	0,03	0,02	0,03	-	<0,02	-	0,02	-	0,02	-	0,4	1,2	-	2,2
20	nitrit_ion	0,02	0,02	0,02	0,01	-	0,01	-	0,01	-	0,01	-	0,01	0,01	-	0,01
21	nitrit_N	0,005	0,005	0,006	<0,002	-	<0,002	-	<0,002	-	<0,002	-	<0,002	<0,002	-	<0,002
22	nitrat_ion	<0,5	<0,5	<0,5	<0,5	-	<0,5	-	<0,5	-	0,66	-	<0,5	<0,5	-	0,7
23	nitrat_N	<0,1	<0,1	<0,1	<0,1	-	<0,1	-	<0,1	-	0,15	-	0,1	<0,1	-	0,2
24	szervetlen_N	0,07	0,09	0,1	<0,07	-	<0,07	-	0,1	-	0,17	-	0,5	1,3	-	1,4
25	oldott_ortofoszfat	<0,02	<0,02	<0,0z	0,02	-	<0,02	-	0,03	-	0,03	-	0	0	-	0,1
26	oldott_ortofoszfat_P	0	0,01	0,01	0,01	-	0,01	-	0,01	-	0,01	-	0	0	-	0
27	ossz_vas	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
28	klorofill	9,2	25	53	30	-	37	-	25	-	13	-	9,7	11	-	13
29	telepszam	1200	1200	200	310	-	240	-	120	-	150	-	170	140	-	300
30	clostr_szam	80	150	240	86	-	280	-	460	-	540	-	400	290	-	240
31	peptonbol_ztsz	250	2500	25000	1100	-	4500	-	25000	-	25000	-3E+05	4500	-		40000
32	plank_bakt_szam	2	1,4	1,5	1,7	-	2,7	-	1,4	-	1,3	-	1,5	1,6	-	1,4
33	plank_biomassza	0,231	0,158	0,198	0,276	-	0,358	-	0,164	-	0,146	-	0,169	0,162	-	0,205

Az adatsorok hiányosságának oka tipikusan szerkezeti: a vizállás korlátozza, hogy maximálisan milyen mélységü adatok létezhetnek egyáltalán. De feltehetően technonógia zavar is előfordult egy-két esetben. Megjegyzendő, hogy az eredményeket elemezgetve látható néhány olyan kiugró érték aminek értelmezése az adatokhoz értő hidrológus segítsége nélkül aligha lehetséges.

A PLS eljárás — és annak az általunk alkalmazott implementációja is — alkalmas arra, hogy hiányzó és kiugró adatokat kezeljen. De az efajta feldolgozás olyan részletességet igényel ami meghaladja ennek a dolgozatnak a keretét. Emiatt a rendelkezésre álló adatoknak csak azt a részét használhattuk ami teljes, nem tartalmaz hiányzó adatokat.

Öt olyan méréssor áll rendelkezésre ami "teljes vízmélység", azaz 500 cm mellett keletkezett. Az azonos vízmélység (a számunkra jelentős technikai könnyebbség mellett) azért jelentős, mert ahogyan van a rétegek egy a rétegnek a vízfelszintől mért távolsága alapján leírható jellegzetessége, ugyanúgy van egy rétegnek a fenéktől mért távolsága alapján leírható jellegzetessége is. Ezt, tárgyszerü modell esetén, több vízjellemző modelljében feltétlen figyelembe kellene venni. Azaz gondos modellezés esetén egy-egy réteget két réteg paraméterrel — a felszín alatti mélység, és a talaj feletti magasság — kell leírni. Az öt 500 centiméteres mintából egy speciális: jégalól vett minta. A másik négy két egymást követő évben egy-egy nyári illetve őszi minta, 2001-2002-ből. Ezeket az adatokat elemeztük.

A kutatók alkalmanként és rétegenként 32 változót vizsgáltak. A mért adatok változékonysága erősen eltérő. Vannak változók, amiknek az értékei egyes mérésekkor konstansnak bizonyultak. Azaz vannak olyan változók, amik látszólag 0 szórásuak. Ez az alkalmazott módszerek mellett problémát jelent. További nehézséget jelent, hogy a PLS becslések megbízhatóságának vizsgálatára tipikusan a jackknife illetve a bootsrap eljárást alkalmazzák. Ezen eljárások a vizsgált modellt egy-egy pszeudo minta sorozatra alkalmazzák. Olyan pszeudómintákat véve, amik az eredeti minta részmintái, illetve amik annak a mintavételezésével keletkeznek. Márpedig ha egy-egy változó tipikusan konstans, azaz ha az értéke csak egy-egy esetben tér el a konstans értéktől, akkor a mintavételezéskor könnyen adódhat, hogy a pszeudo minta egy-egy esetben nulla szórású.

A mondottak miatt a kiválasztott adatokból (elhagyva az olyan diszkrét értékü változókat is mint a "víz szagának minősítése", vagy például a "víz szine") a magyarázandó változókból 28, a magyarázó változókból pedig 37 marad. A PLS modell illesztésére az R-project "pls" nevü kiegészítését alkalmaztuk.
Ezt a kiegészítést Bjo/n-Helge Mevik, Ron Wehrens és Kristian Hovde Liland készítették "Partial Least Squares and Principal Component regression" címmel. Felhasználható és elérhető a [15] helyen irtak figyelembevételével. A "pls" kiegészítés (package) a klasszikus Sijmen de Jong féle SIMPLS és ortogonális eljárások mellett a Dayal-MacGregor féle "kernelpls" és a Ra:nner féle "widekernelpls" eljárások alkalmazását is lehetővé teszi.

Számtalan kisérletet folytattunk. De ennek csak igen kis részét dokumentáljuk. Ugyanis a kisérletek javarészt technikai jellegüek voltak. A programrendszer működési módjának felderítésére irányultak.

Az első programrészlet az adatok előkészítését mutatja be. Az adatokat előzőleg a "T0233.csv" nevü állományban 15 sorban, 480 oszlopba, a mezőket pontosvesszővel elválasztva rögzítettük. A rendelkezésre álló adatokat az idézett forrás mellékletének 5.-19. táblázatából vettük. A vízmintavételi mélységeket a program "mely" nevü változójába irtuk. A "minta" nevü változó a forrástábla sorszámán kívül tartalmazza azt is, hogy az adott minta "Jeges", "Tavaszi", "Nyári" vagy "Őszi".

Az második programrészlet a modellillesztést és a későbbiekben bemutatott kontrol ábrák elkészítésének módját illusztrálja.

```
setwd("D:/-12e/-oeu/Tisza")
T<-read.table("T0233.csv",sep=";",stringsAsFactors=FALSE)
nev<-read.table("nevek.csv",stringsAsFactors=FALSE)
mely<-c(20,100,150,200,225,250,275,300,325,350,375,400,430,470,500)
minta<-paste(" ",5:19,c("J","J","T","T","T","T","N","N","N","N","O","O","O","O"),sep="")
minta<-substr(minta,nchar(minta)-2,nchar(minta))
rownames(T)<-mely
colnames(T)<-paste(minta,rep(1:32,each=15),sep="-")
save(T,file='T.Rdata')
```

7.1. ábra. Adatok előkészítése

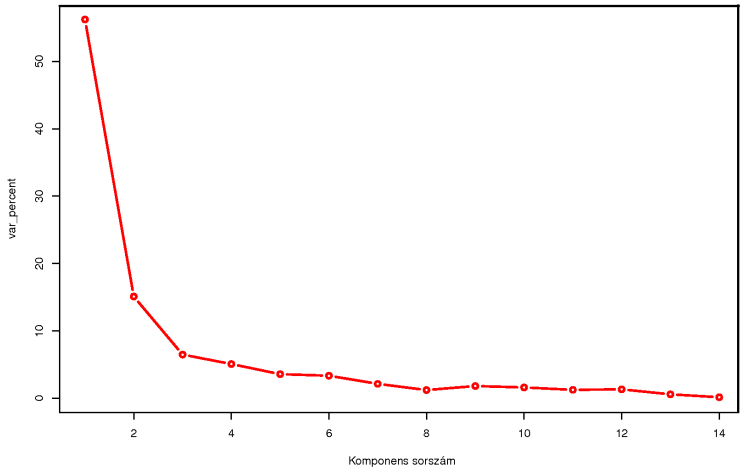
```
# a két oszi (O) mérést a két nyári (N) méréssel magyarazuk

O<-as.matrix(D[,colnames(D)[rep(((1:28)-1)*5,each=2)+4:5]])
N<-as.matrix(D[,colnames(D)[rep(((1:28)-1)*5,each=2)+2:3]])
dim(O);dim(N)# 15x56  15 melyseg, 2*28 valtozo
O<-O[,!diag(cov(O))<.1]
N<-N[,!diag(cov(N))<.1]
dim(O);dim(N)# kivettuk a kis szórásuakat: 28, 37 maradt
M<-mvr(O~N,scale = TRUE,ncomp=3)# Y:28, X:37, T:3
Mc<-mvr(O~N,scale=TRUE,ncomp=3,validation = "LOO")
str(M,m=1,give.a=FALSE)
str(Mc,m=1,give.a=FALSE)
var_percent<-explvar(mvr(O~N,scale=TRUE))
postscript("var_percent.ps")
plot(var_percent,t='b',lwd=3,col='red',xlab='Komponens sorszám')
dev.off()
r<-1/sqrt(c(1,1.5,2.25,4,8))
postscript("corr_plot.ps")
corrplot(M,radii=r,col="red",pch=20,labels="num")
dev.off()
postscript("pre_plot.ps")
predplot(Mc,nCols=1,nRows=1,col='red',labels="names")
dev.off()
```

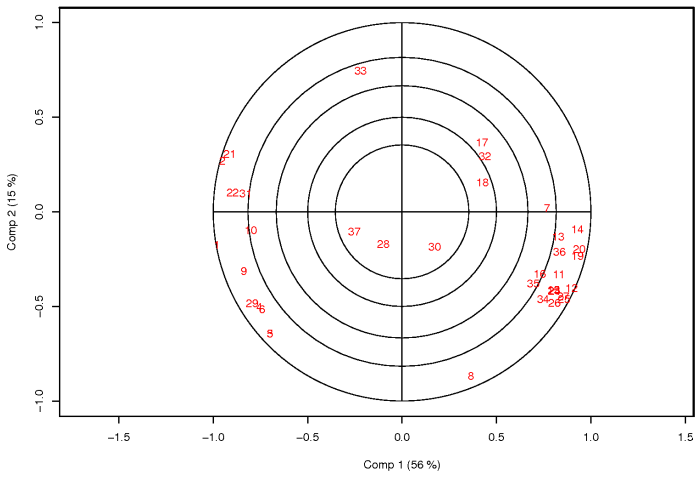
7.2. ábra. Modellillesztés, és kontrol ábrák elkészítése

A [7.3] ábrán azt mutatjuk be, hogyan változnak a szekvenciálisan meghatározható T háttérváltozók szórásai. Az ábráról leolvasható, hogy az első komponens az infor-

máció kb 56%-át, a második az információ 15%-át tartalmazza. A további komponensek súlya egyenként nem haladja meg a 3-5%-ot. Tehát a modellezéskor (külön ok híján) elégséges az első két komponenst figyelembe venni.



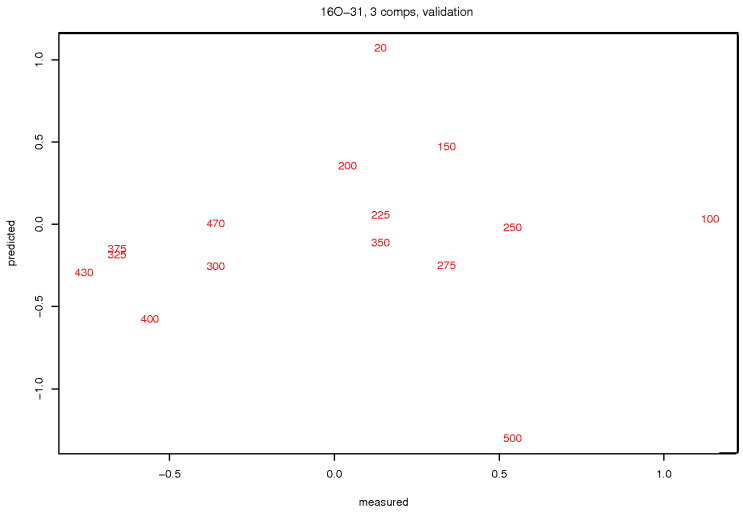
7.3. ábra.



7.4. ábra.

A [7.4] ábrán a 37 magyarázó változónak a komponensekhez való viszonyát mutatjuk be. A pontok koordinátái az egyes változók és a komponensek közti korreláció. Jól látható, hogy mely változók azok, amik a többiekkel azonos, melyek azok amik pedig a többiektől eltérő módon viselkednek. Csak az érdekesség kedvéért jegyezük meg, hogy a három legkevébé megmagyarázott változó az ami a kör középpontjához

legközelebb esik, azaz 28, 30, 37 vagyis a "klorofill", "telepszam" és "plank bakt szam". Az ábrán jól látszik, hogy a másik önálló csoport a 17, 18, a két "magnezium ion" adat és a "clostr szam", valamint a két különálló változó a 33 és a 8 a "clostr szam" illetve a "KOI". Ennek akár jelentőséget is tanusíthatunk. Ám komoly következtetések levonása elött mégegyszer idézzük a bevezetésben fejtegetetteket. Miszerint az itt közölt 'eredmények' bizonytalanok. Az ábrák tartalmi értékelésére irányuló szándék nélkül, technikai szempontok alapján kiválogatott adatok alapján készültek.



7.5. ábra.

A [7.5] ábra megfelel annak a klasszikus regresszió diagnosztikai rajznak, amikor azt vizsgáljuk, hogy a célváltozó modell szerinti és mért értéke közti eltérésben van-e valamilyen tendenciaszerü. Az alkalmilag kiválasztott 31. változó — ami a "planktonikus baktérium szám" — esetében azt láthatjuk, hogy modellünkből nyilvánvalóan hiányzik a réteg szerinti függés modellezése. Ugyanis a felszintől vett mélység értékével megcimzett 15 adat — mint ahogyan látható — kis hibával egy görbevonal mentén helyezkedik el.

Köszönetnyilvánítás

Ezúton szeretném kiemelten megköszönni Pröhle Tamásnak, konzulensemnek, a kitartó segítségét, minden jó tanácsát, és az időt, amit rám szánt, amivel jelentősen hozzájárult szakdolgozatom elkészítéséhez. Továbbá köszönöm családomnak, akik mindvégíg hittek bennem és támogatták az egyetemi pályafutásomat. Nem utolsó sorban pedig szobatársaimnak a biztatásukért, valamint szaktársamnak Péter Zsófiának, akivel egymást segítve küzdöttük le az elmúlt pár év nehézségeit.

Irodalomjegyzék

[1] Hajdú Ottó: *Statisztikai elemzések*, AULA kiadó, 2003.

[2] Márkus László: *Idősorok és többdimenziós statisztikai módszerek jegyzet* ELTE.

[3] Móri Tamás: *Főkomponens- és Faktoranalízis jegyzet*,ELTE, 1999.

[4] Füstös László, Meszéna György, Simonné Mosolygó Nóra: *A sokváltozós adatelemzés statisztikai módszerei*, Akadémiai Kiadó, Budapest, 1986.

[5] Michel Tenenhaus: *Component-based Structural Equation Modelling*

[6] Petres Tibor Tóth László: *Statisztika*, Szegedi Tudományegyetem, 2001.

[7] Paul H. Garthwaite: *An interpretation of Partial Least Squares*, 1994.

[8] Hervé Abdi: *Partial Least Square Regression PLS-Regression*, 2007.

[9] Roman Rospial and Nicole Krämer: *Overview and Recent Advances in Partial Least Squares*, 2006.

[10] Domán Csaba: *Többváltozós korreláció- és regressziószámítás jegyzet*, Miskolci Egyetem, 2005.

[11] Hunyadi László: *Statisztika közgazdászoknak*, AULA kiadó, 2003.

[12] Teszárné Nagy Mariann: *Az élővíz rétegzetségének hidroökológiai jelentősége a Kiskörei-tározó területén*, Debreceni Egyetem, 2006.

[13] Márialigeti Károly: *személyes közlés*.

[14] Dévai György: *A Tisza-tó két védett morotvája, Debreceni Egyetem*, kézirat, 1997.

[15] Bjo/rn-Helge Mevik: http://mevik.net/work/software/pls.html, Publication: 2011-11-28 (Ellenőrizve: 2012-05-28 12:00:06)

NYILATKOZAT

Név: Horváth Vivien
ELTE Természettudományi Kar, szak: Matematika BSc.
ETR azonosító: HOVQABT.ELTE
Szakdolgozat címe: Pls regresszió

A szakdolgozat szerzőjeként fegyelmi felelősségem tudatában kijelentem, hogy a dolgozatom önálló munkám eredménye, saját szellemi termékem, abban a hivatkozások és idézések standard szabályait következetesen alkalmaztam, mások által írt részeket a megfelelő idézés nélkül nem használtam fel.

Budapest, 2012. május 29.

a hallgató aláírása