Eötvös Loránd Tudományegyetem
Természettudományi Kar

Németh Ádám

A faktoranalízis alkalmazhatósága

Matematika BSc szakdolgozat
Alkalmazott matematikus szakirány

Témavezető: Pröhle Tamás

Matematikai Intézet
Valószínűségelméleti és Statisztika Tanszék

2015. Budapest

Köszönetnyilvánítás

Szeretnék köszönetet mondani témavezetőmnek, Pröhle Tamásnak, azért a rengeteg segítségért és hasznos tanácsért, amelyet a szakdolgozat elkészítése során nyújtott.

Köszönöm családomnak és barátaimnak, hogy folyamatosan támogattak, és nem hagyták, hogy bármi eltérítsen arról az útról, amely a dolgozat elkészüléséig vezetett.

Végül, de nem utolsó sorban szeretném megköszönni általános iskolai és gimnáziumi tanáraimnak, hogy végig segítettek, hogy eljuthassak idáig.

Tartalomjegyzék

Ábrák jegyzéke

1. fejezet

Bevezetés

A változók számának csökkentése során célunk, hogy a statisztikai mintaában rejlő információ lehetőleg kis csökkentésével, ugyanazt a jelenséget kevesebb változóval írjuk le. Szakdolgozatomban a faktoranalízisre térek ki részletesebben, ennek alkalmazhatóságát vizsgálom, és végül bemutatom a működését egy példán keresztül. A faktoranalízis során több megfigyelt változó kapcsolatát vizsgálhatjuk, ezeket szeretnénk közös faktorok segítségével jellemezni. Ez nyilván csak akkor lehetséges, ha az eredeti változók között szoros kapcsolatot feltételezünk, azaz magas a korrelációjuk. Ennek vizsgálatára többféle tesztet is ismerünk, én a Kaiser-Meyer-Olkin-tesztet fogom részletesebben bemutatni, amellyel könnyen el tudjuk dönteni, hogy érdemese faktorelemzést végrehajtani. Ehhez szükség lesz bizonyos alapdefiníciókat bevezetni, melyeket a dolgozat során felhasználok.

Végül egy lehetséges alkalmazást mutatok be az amerikai futballban szereplő statisztikák kapcsolatára vonatkozóan: ezekről szeretnék pontosabb képet adni, megvizsgálni, hogy melyek fontosak, melyek kevésbé, és milyen összefüggések vannak köztük.

2. fejezet

Felhasznált statisztikai eszközök

2.1. A változók közötti kapcsolatok

1. Definíció. Legyen X és Y valószínűségi változó. X és Y *kovarianciája*:

$$Cov(X,Y) = E[(X - EX)(Y - EY)] = E(XY) - EXEY.$$

A kovariancia tulajdonságai: Legyenek X és Y továbbra is valószínűségi változók, a és b pedig valós számok.

- $Cov(X,X) = D^2X$ – önmagával vett kovariancia a szórásnégyzettel egyenlő.
- $Cov(X,Y) = Cov(Y,X)$ – szimmetria.
- $Cov(X,a) = 0$ – konstanssal való kovariálatlanság.
- $Cov(X+a, Y+b) = Cov(X,Y)$ – eltolásinvariancia.
- $Cov(aX,bY) = abCov(X,Y)$ – linearitás.
- $D^2(X \pm Y) = D^2X + D^2Y \pm 2Cov(X,Y)$

A gyakorlatban gyakran egy normált értékre van szükség, hogy tudjuk mérni két teljesen eltérő értékű valószínűségi változópár kapcsolatának erősségét is. Ezt szolgálja a korreláció.

2. Definíció. Legyen X és Y valószínűségi változó. X és Y *korrelációja*:

$$R(X,Y) = \frac{Cov(X,Y)}{DX\,DY}.$$

A kovariancia tulajdonságai alapján nyilvánvaló, hogy a korreláció lineárisan invariáns. A korreláció a két valószínűségi változó lineáris kapcsolatát méri. Ha 1-hez közeli az R^2, akkor a két változó lineárisan erősen függ egymástól, míg ha 0-hoz közeli, akkor a két változó közötti kapcsolat gyenge.

Ha $Cov(X,Y) = 0$, akkor X és Y kovariálatlannak vagy korrelálatlannak mondjuk.

1. Megjegyzés. *A változók függetlenségéből következik a korrelálatlanság, azonban a korrelálatlanságból nem következik a függetlenség!*

3. Definíció. Legyen X m-dimenziós és Y n-dimenziós valószínűségi változó. X és Y *kovarianciamátrixa* az egydimenziós kovarianciához hasonlóan a következő módon definiálható:

$$\Sigma = E[(X - EX)(Y - EY)^T] = E(XY^T) - EX(EY)^T.$$

4. Definíció. Legyen X és Y n-dimenziós valószínűségi változó. X és Y *korrelációmátrixa*

$$R = \left(\frac{s_{ij}}{\sigma_i \sigma_j}\right),$$

ahol σ_i és σ_j a megfelelő változók szórásait jelölik, s_{ij} pedig a kovarianciamátrix (i,j)-edik elemét.

Hogyan tudunk adott kovarianciamátrixból egyszerűen, az egyes változók ismerete nélkül korrelációmátrixot csinálni? Legyen a kovarianciamátrix a következő:

$$\mathbf{\Sigma} = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1n} \\ s_{21} & s_{22} & \dots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \dots & s_{nn} \end{pmatrix}$$

Mivel s_{ij} az i-edik és j-edik változó kovarianciájának értékét mutatja, s_{ii} illetve s_{jj} pedig az i-edik és j-edik változó szórásnégyzetét, így az i-edik és j-edik változó közötti korreláció kiszámítható a következő módon:

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}$$

Ezáltal a korrelációmátrix a következőképpen néz ki:

$$\mathbf{\Sigma} = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ r_{21} & r_{22} & \dots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \dots & r_{nn} \end{pmatrix}$$

Itt a főátlóban 1-esek szerepelnek.

Ez az algoritmus az R programcsomagban egyszerűen implementálható a *cov2cor* paranccsal, amely paraméterként egy kovarianciamátrixot vár, majd azt korrelációs mátrixszá alakítja. Ehhez először a *corpcor* csomagot kell betöltenünk. Lássunk erre az alábbiakban egy konkrét példát. Ehhez használom a [5]-ös forrást. Legyen Σ az alábbi kovarianciamátrix:

$$\mathbf{\Sigma} = \begin{pmatrix} 36 & 45 & 27 \\ 45 & 81 & 45 \\ 27 & 45 & 64 \end{pmatrix}$$

Ezt a *cov2cor* függvénnyel átalakítjuk az alábbi R korrelációs mátrixszá:

$$\mathbf{R} = \begin{pmatrix} 1.0000 & 0.8333 & 0.5625 \\ 0.8333 & 1.0000 & 0.6250 \\ 0.5625 & 0.6250 & 1.0000 \end{pmatrix}$$

A korrelációs mátrixból le tudjuk olvasni az egyes változók lineáris függésének erősségét: például látható, hogy az 1-es és 3-mas változó függése viszonylag alacsony, míg az 1-es és 2-es változó között igencsak erős kapcsolatot tapasztalunk.

A statisztikában azonban nem állnak rendelkezésre az elméleti értékek, ezért a megfigyelésekből számolt tapasztalati értékekkel dolgozunk.

5. Definíció. Legyen $X_1, X_2, ..., X_n$ az X eloszlásából, $Y_1, Y_2, ..., Y_n$ az Y eloszlásából származó minta. Ekkor a tapasztalati korreláció értékét a következő képlet adja meg:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}, \tag{2.1}$$

ahol $\bar{x}$ és $\bar{y}$ a mintákból számolt átlagok, s_x és s_y pedig a tapasztalati szórások.

Több változó esetén gyakran találkozunk meglepő korrelációs adatokkal. Például az Amazonas vízszintje, és a kávé ára között magas korrelációt tapasztalhatunk. Ez nyilvánvalóan nem függ közvetlenül egymástól, azonban találhatunk olyan közös faktort, amellyel mindkét változót magyarázni tudjuk: ez a csapadékmennyiség. Minél több eső esik, annál magasabb lesz a vízszint, és annál több kávé terem, ezért alacsonyabb lesz annak a világpiaci ára. Azonban ha kevés csapadék hull, akkor alacsonyabb a vízszint, valamint a termés is kevesebb lesz, ezért drágábban fogják árulni a kávét.

Általánosabban mi arra lennénk kíváncsiak, hogy két változó között milyen összefüggés tapasztalható, ha a harmadik változó – esetünkben a csapadék – hatásától eltekintünk. Erre a kérdésre a parciális korreláció adhatja meg nekünk a választ. Ebben a részben az [1]-es cikket használom.

6. Definíció. Legyen X, Y és Z valószínűségi változó. Érdekelne minket az X és Y közötti azon korreláció, ami Z-vel nem magyarázható. Ennek a mérésére szolgál a *parciális korreláció*, melyet a következő módon számíthatunk ki:

Először a későbbi 2.2-es részben leírtak szerint határozzuk meg a Z változó X-re vonatkozó lineáris regressziós becsléseként az X változónak azt a részét, amely lineárisan függ Z-től (Z_X). Ekkor X-ből kivonva ezt a részt, a maradék már nem függ Z-től, azaz az X változó Z-től független része:

$$X_{mar} = X - X_Z.$$

Haonlóan az Y változó Z-től független része:

$$Y_{mar} = Y - Y_Z.$$

A Z-től lineárisan nem függő X_{mar} és Y_{mar} közötti korrelációt nevezzük parciális korrelációnak:

$$R_{XY.Z} = R(X_{mar}, Y_{mar}).$$

A parciális korrelációs együttható kiszámításának egyszerű módja az alábbi képlettel történhet:

$$R_{XY.Z} = \frac{R_{XY} - R_{XZ}R_{YZ}}{\sqrt{1 - R_{XZ}^2}\sqrt{1 - R_{YZ}^2}}, \tag{2.2}$$

ahol R_{ij} az i és j változók közötti korrelációt jelöli.

Ezen képlet ismételt alkalmazásával további változók hatását is ki lehet szűrni X és Y korrelációjából, ha a jobb oldalon szereplő korrelációknál Z helyére az újonnan figyelembe vett magyarázó változót kell írni, és a korrelációk helyett azokat a parciális korrelációkat írjuk, amelyet a már előzőleg figyelembe vett változók szerint kaptunk. Fontos megjegyezni, hogy a végső parciális korreláció értéke nem függ a kiszűrések sorrendjétől, tehát például

$$R_{XY.ZUV} = R_{XY.UVZ} = R_{XY.ZVU},$$

amely az X és Y változók korrelációját jelöli a Z, V és U változók kiszűrése után.

2. Megjegyzés. *A parciális korreláció nem azonos a feltételes korrelációval. Ez abból is látszik, hogy a feltételes korreláció egy véletlen érték, míg a parciális korreláció egy -1 és $+1$ közti érték, ami valójában a fent leírt két maradék korrelációja.*

3. Megjegyzés. *Elterjedt hiba azt gondolni, hogy a parciális korreláció minden esetben abszolútértékben kisebb, mint a korreláció. Ez a képzet abból származik, hogy a két lineáris regresszióval a két változóban levő információ közös részének egy hányadat eltüntetjük. Ez ugyan igaz, de ebből nem következik a korreláció csökkenése. Ez a következő geometriai interpretációból is könnyen látszik. A korreláció értelmezhető úgy, mint a valószínűségi változóknak megfelelő vektorok szögének koszinusza, tehát az, hogy két változó korrelációja nagyobb vagy kisebb, azon múlik, hogy a változópár szöge kisebb vagy nagyobb. A*

parciális korreláció geometriai értelmezéséhez pedig vegyük a következő modellt: Legyen Z egy 3-dimenziós vektor, legyen e és f két a Z-t tartalmazó sík, amelyek rendre tartalmazzák az X és Y vektorokat is. Ekkor az X és Y Z-szerinti parciális korrelációja a két sík szögének felel meg, míg az X és Y korrelációjának a két vektor szöge.

Mint az nyilvánvaló, az X és Y szöge lehet akár kisebb, akár nagyobb, sőt egyenlő is az őket tartalmazó két sík szögével. Ez mutatja, a fenti kijelentés tévességét.

A változópárok parciális korrelációja többféle módon is kiszámítható. Egyik lehetséges út a fenti definíció szerinti, másik lehetséges út az inverzmátrix (-1)-szereséből képzett korrelációmátrixból (ezt az eljárást követi az általunk alkalmazott corpcor csomag is).

A parciális korreláció normális eloszlás esetén megegyezik a feltételes korrelációval. A parciális korreláció szokványos bevezetése az is, hogy a normális esetbeli feltételes korreláció képletét általánosítjuk, ami egyébként a regressziós módszerrel azonos képletet eredményez.

R-ben egy korrelációs mátrixot a *corpcor* csomagból a *cor2pcor* paranccsal azonnal parciális korrelációs mátrixszá tudjuk alakítani. Nézzünk erre is egy példát! Vegyük az alábbi R korrelációs mátrixot:

$$\mathbf{R} = \begin{pmatrix} 1.0000 & 0.8333 & 0.5625 \\ 0.8333 & 1.0000 & 0.6250 \\ 0.5625 & 0.6250 & 1.0000 \end{pmatrix}$$

Ezt a *cor2pcor* függvénnyel átalakítjuk az alábbi P parciális korrelációs mátrixszá:

$$\mathbf{P} = \begin{pmatrix} 1.0000 & 0.7464 & 0.0966 \\ 0.7464 & 1.0000 & 0.3419 \\ 0.0966 & 0.3419 & 1.0000 \end{pmatrix}$$

Ebből a mátrixból például leolvshatjuk, hogy az első és második változók közötti korrelációt a harmadik változó szinte nem is befolyásolja, míg az első

és harmadik változók közötti korreláció főként a második változó hatásának az eredménye. Ha kivesszük a második változó hatását a rendszerből, akkor a korreláció értéke mindössze 0.0966.

2.2. Lineáris regresszió

Gyakori eset, hogy nem ismerjük a számunkra érdekes mennyiség (Y) pontos értékét (pl. holnapi vízállás, csapadékmennyiség). Van viszont információnk hozzá kapcsolódó mennyiségről (X, mai adatok). Célunk X segítségével minél jobban közelíteni Y-t. Matematikailag célunk olyan f_{opt} függvény keresése, amely megoldása a

$$\min_f E(Y - f(X))^2 \tag{2.3}$$

szélsőérték-feladatnak. Ezt nevezzük legkisebb négyzetes becslésnek. Ezzel azonban az a probléma, hogy csak akkor oldható meg, ha ismert X és Y együttes eloszlása, ami ugyan közelíthető a megfigyelések alapján, de általában a valóságban nem ismert.

1. Állítás. *Ha a 2.3-mas feladatban szereplő f helyére konstanst írunk, akkor a*

$$\min_a E(Y - a)^2 \tag{2.4}$$

feladat megoldása $a = EY$.

Bizonyítás.

$$E(Y - a)^2 = EY^2 - 2aEY + a^2$$

Ezt a szerint deriválva kapjuk, hogy a minimumhely valóban $a = EY$, a minimum értéke pedig D^2Y. □

4. Megjegyzés. *Hasonlóan belátható, hogy X tetszőleges függvénye esetén az $E(Y|X)$ feltételes várható érték adja a megoldást.*

Tekintsünk egy példát: Két érmével dobunk. Másodszor annyi érmével dobtunk újra, amennyi fejet kaptunk az első dobásnál. Csak azt tudjuk, hogy a második dobásnál 1 fejet dobtunk. Közelítsük ennek segítségével az első dobás eredményét! Jelölje Y az első dobás során dobott fejek számát, F pedig a második dobás során dobott fejek számát, ami esetünkben 1. Ekkor a megoldás

$$E(Y|F=1) = \frac{P(X=1,F=1)+2P(X=2,F=1)}{P(F=1)} =$$

$$= \frac{1\cdot 0.5\cdot 0.5+2\cdot 0.25\cdot 0.5}{0.5\cdot 0.5+0.25\cdot 0.5} = \frac{4}{3},$$

tehát az első dobással dobott fejek várható értéke $\frac{4}{3}$.

Keressük most a 2.3-mas pont szerinti legkisebb négyzetes becslésben szereplő f függvényt $aX+b$ alakban. Ebben az esetben a megoldáshoz nem szükséges ismernünk az együttes eloszlást. Végezzük el ebben az esetben is a deriválást.

$$E[Y-(aX+b)^2] = EY^2 + a^2EX^2 + 2abEX + b^2 - 2aEXY - 2bEY$$

Deriválva a szerint:

$$2aEX^2 + 2bEX - 2EXY = 0.$$

Deriválva b szerint:

$$2aEX + 2b - 2EY = 0.$$

Utóbbiból kapjuk, hogy

$$\hat{b} = EY - aEX,$$

ezáltal

$$\hat{a} = \frac{EXY - EXEY}{EX^2 - E^2X} = \frac{Cov(X,Y)}{D^2X}.$$

A lineáris függvények között az $\hat{a}X+\hat{b}$ egyenes az, amelyik a legkisebb négyzetes becslést adja a lineáris függvények közül. Ezt *regressziós egyenesnek* nevezzük.

3. fejezet

Faktoranalízis

3.1. Bevezetés

A faktoranalízis a változók számának csökkentésének legelterjedtebb módszere. Azzal foglalkozik, hogy több megfigyelhető változó által közvetített információt próbál magyarázni, kevesebb a háttérben meghúzódó, nem megfigyelhető változó, ún. faktorok segítségével. A módszer több egymással korreláló változó közötti összefüggéseket vizsgálja, az erősen összefüggő változócsoportokat szeretnénk egy közös faktorral magyarázni. Az így kapott faktorokat ezután megpróbálhatjuk értelmezni, hogy mit is jelenthetnek a valóságban. A faktoranalízis abban az esetben alkalmazható, amikor az eredetileg megfigyelt változók között magas korrelációt, ellenben alacsony parciális korrelációt tapasztalunk. Ennek megvizsgálásához alkalmazhatjuk például a Kaiser-Meyer-Olkin-tesztet (KMO-teszt).

A módszert alkalmazzák a pszichológiában, viselkedés- és társadalomtudományban, szociológiában, marketingben, és még sok további területen.

A faktoranalízis abban különbözik a főkomponensanalízistől, hogy a főkomponensek az eredeti változók lineáris kombinációi, ezek magukat az egyes változókat modellezik, míg itt az eredeti változókat fejezzük ki a faktorok lineáris kombinációjaként, ezekkel a változók közötti korrelációkat próbáljuk magyarázni. Ebben a részben a [3]-mas forrást használtam fel.

3.2. Kaiser-Meyer-Olkin-teszt

A faktorelemzés előtt meg kell vizsgálnunk, hogy a változók között milyen erősségű kapcsolat van, hogy megtudjuk, hogy egyáltalán alkalmasak-e a változók a faktoranalízisre. A módszert olyan esetekben lehet alkalmazni, amikor a megfigyelt változók között magas korrelációt tapasztalunk, hiszen éppen az az elemzés célja, hogy ezeket egymással korrelálatlan faktorokkal írjuk le. A változók közötti összefüggések mérésére több módszer is létezik, az alábbiakban ezek közül mutatnék be egy példát.

A leggyakrabban használt módszer a Kaiser-Meyer-Olkin-teszt, amely a parciális korrelációk segítségével kiszámol egy ún. KMO-indexet, amely megmutatja, hogy milyen típusú összefüggést tapasztalunk a változóink között, vagyis mekkora korrelációt, illetve parciális korrelációt mutatnak az adatok. Ennek segítségével el tudjuk dönteni, hogy érdemes-e a faktoranalízist végrehajtani.

A KMO-index képlete a következő:

$$KMO = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} R_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n} R_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij}}, i \neq j \tag{3.1}$$

ahol R_{ij} az i-edik és j-edik változó közötti korreláció, r_{ij} pedig az i-edik és j-edik változó közötti parciális korreláció. Ennek az indexnek az érték 0 és 1 között mozog, és a következőket szokás mondani a faktoranalízis alkalmazhatóságáról:

- $0.9 \leq KMO$ – csodálatos (marvelous)
- $0.8 \leq KMO < 0.9$ – dicséretes (meritorious)
- $0.7 \leq KMO < 0.8$ – közepes (middling)
- $0.6 \leq KMO < 0.7$ – mérsékelt (mediocre)
- $0.5 \leq KMO < 0.6$ – szánalmas (miserable)
- $KMO < 0.5$ – elfogadhatatlan (unacceptable)

Ezen értékeknek az eredetére nem találtam semmilyen hivatkozást, ezért szimulációval megvizsgáltam R-ben az indexet. Alá is tudjuk támasztani ezeket, ha megvizsgáljuk R-ben a KMO-teszt által kapott értékeket, illetve a faktorelemzéssel kapott közelítés minőségét. Ehhez a [4]-es forrást használtam. A vizsgálat előtt töltsük be R-ben a *clusterGeneration* csomagot, amelyben a *genPositiveDefMat* függvény segítségével tudunk véletlenszerűen generálni kovarianciamátrixokat. Ez a függvény úgy működik, hogy először véletlen sajátvektorokat generál 1 és 10 között az egyenletes eloszlás szerint, majd ezeket egy szintén véletlenül generált ortogonális mátrixszal egy véletlen pozitív definitté transzformálja. Végül visszatérési értékként megadja a véletlen kovarianciamátrixot és a sajátértékeit. A 3.1-es algoritmus segítségével generálunk egy p-dimenziós korrelációs mátrixot, amelyet korrelációs-, majd parciális korrelációs mátrixszá transzformálunk.

```
gen<-function(p)
{
R <-genPositiveDefMat(p)
R <- R$Sigma
r<-sqrt(diag(R));R <- R/(r %o% r); rm(r); diag(R)<-1
P <- solve(R)
P <- -P ; p<-sqrt(-diag(P))
P <- P/(p %o% p); rm(p); diag(P)<-1
return(list(R=R,P=P))}
```

3.1. ábra. Korrelációs és parciális korrelációs mátrixok generálása

Ezzel a függvénnyel a 3.2-es algoritmusban generálunk 5000 db 5-dimenziós mátrixot, kiszámoljuk ezekre a Kaiser-Meyer-Olkin-index értékét, majd elvégezzük a faktoranalízist 2 faktorral, és kiszámolunk egy távolságot, ami mutatja, hogy a faktoranalízissel mennyire sikerült jól modellezni a változók közötti kapcsolatot.

Ezt megcsináljuk 5000 alkalommal, majd ábrázoljuk, hogy mit tapaszta-

```
for(k in 1:N)
{ w<-gen(5)
P<-w$P
R<-w$R
M<-factanal(cov=R,fac=2,rot="none")
kiszamoljuk a KMO-t ----------
p<-dim(R)[1]
KMO<-(sum(R^2)-p)/(sum(R^2)+sum(P^2)-2*p)
kiszamoljuk a korrelaciotol valo tavolsagot --
L <- M$loadings
class(L) <- NULL
L2<-sum((R-(L%*%t(L)+diag(M$uni)))^2)/(sum(R^2)-p)
tav<-c(KMO,L2)
d[k,]<-tav
}
```

3.2. ábra. A KMO-index és a faktoranalízis kapcsolata

lunk, ezt látjuk a 4.1-es ábrán. Az ábráról leolvasható, hogy mikor a KMO-index értéke 0.5 alatt van, akkor igen nagy a mátrixok távolsága, a modellünk tehát rossz. Minél nagyobb azonban ez a szám, annál jobb a közelítésünk, tehát mondhatjuk, hogy a KMO-index valóban jól használható a faktoranalízis alkalmazhatóságának kiderítéséhez, azaz arra, hogy megvizsgáljuk a kiinduló változók közötti összefüggéseket.

5. Megjegyzés. *A 4.1-es ábra a KMO függvényében mutatja azt, hogy egy véletlen korrelációmátrix faktoranalízise mennyire jól közelíti a korrelációmátrixot. A regressziós vonal lokális regresszióval készült. Ez azt mutatja, hogy kicsi KMO-érték esetén a KMO-érték és az L2-érték közötti kapcsolat igen laza, míg nagyobb KMO-érték esetén szorosabb.*

6. Megjegyzés. *Az R a véletlen ortogonális transzformációkat véletlen kétdimenziós alterekbeli Givens-rotációkkal állítja elő.*

3.3. A faktoranalízis modellje

Ebben a részben a [6]-os forrás alapján a faktoranalízis modelljét vizsgáljuk.

7. Definíció. Legyen x p-dimenziós valószínűségi változó m várható értékkel, és Σ kovarianciamátrixszal. Azt mondjuk, hogy x leírható a *k-faktormodellel*, ha előállítható

$$x = Af + u + m \tag{3.2}$$

alakban, ahol A $(p \times k)$-s ún. *átviteli (loading) mátrix*, f k-dimenziós közös faktor-vektor (faktor score), u p-dimenziós egyedi faktor-vektor.
$f_1, f_2, ..., f_k$ páronként korrelálatlanok, azaz $R(f_i, f_j) = 0$, $Ef_i = 0$, $D^2 f_i = 1$.
$u_1, u_2, ..., u_p$ páronként korrelálatlanok, azaz $R(u_i, u_j) = 0$, $Eu_i = 0$, $D^2 u_i = \psi_{ii}$.
$f_1, f_2, ..., f_k$ és $u_1, u_2, ..., u_p$ páronként korrelálatlanok, azaz $R(f_i, u_j) = 0$.

Ennek alapján a koordinátánkénti alakra azt kapjuk, hogy

$$x_i = \sum_{j=1}^{k} a_{ij} f_j + u_i + m_i \qquad\qquad i = 1, ..., p \tag{3.3}$$

Mivel u_i és f_j korrelálatlanok és m_j konstans, ezért x_i szórásnégyzetére azt kapjuk, hogy

$$\sigma_{ii} = \sum_{j=1}^{k} a_{ij}^2 + \psi_{ii} \tag{3.4}$$

8. Definíció. A 3.4-es képletben az egyes szórások két részre bonthatók: az egyik rész a közös faktoroktól függ ezt nevezzük *kommunalitásnak*:

$$h_i^2 = \sum_{j=1}^{k} a_{ij}^2 \tag{3.5}$$

A másik rész az egyedi faktoroktól függ, ez a ψ_{ii} tag, amit *egyedi varianciának* nevezünk.

A faktormodellben célunk az ismeretlen A és Ψ mátrixok meghatározása. Ehhez a 3.2-es egyenlet segítségével felírhatjuk a következő egyenlőséget:

$$\Sigma = AA^T + \Psi \tag{3.6}$$

Tehát ha x leírható a 3.2-es faktormodellel, akkor a kovarianciamátrix 3.6-os alakú.

A továbbiakban az egyszerűség kedvéért tegyük fel, hogy $m = 0$.

1. Lemma. *Ha egy x valószínűségi változó leírható a k-faktormodellel, akkor tetszőleges átskálázás után is leírható marad a k-faktormedellel, valamint az A átviteli mátrix sorainak tetszőleges elforgatása után is átviteli mátrix marad.*

Bizonyítás. Az átskálázás ekvivalens egy D diagonális mátrixszal való szorzással. Legyen $y = Dx$. Ha $x = Af + u$, akkor $y = DAf + Du$, és így $Eyy^T = D\Sigma D = DAA^TD + D\Psi D$. Tehát az y valószínűségi változó átviteli mátrixa DA lesz, f és Du továbbra is korrelálatlanok, $D\Psi D$ továbbra is diagonális marad. Ezáltal a változó továbbra is leírható a k-faktormodellel, tehát az állítás igaz.

Az x változóról tudjuk, hogy 3.2-es alakú, így tetszőleges V ortonormált mátrixra

$$x = (AV)(V^T f) + u.$$

Mivel f helyett Vf is eleget tesz a 7-es definíció feltételeinek, ezért x felírható egy olyan k-faktormodellel is, ahol az átviteli mátrix AV. □

Tehát a lemma alapján a k-faktormodell csak forgatástól eltekintve lehet egyértelmű.

1. Lemma. *Egy 0 várható értékű, Σ kovarianciamátrixú x valószínűségi változó akkor és csak akkor írható le a k-faktormodellel, ha létezik olyan Ψ nemnegatív elemű diagonális mátrix, amelyre $\Sigma - \Psi$ k-adrangú pozitív szemidefinit mátrix.*

Bizonyítás. Ha x leírható a k-faktormodellel, akkor a 7-es definícióból, és a 3.6-os egyenlőségből az állítás nyilvánvaló.

Ha létezik olyan Ψ mátrix, ami eleget tesz a feltételeknek, akkor tudjuk, hogy a $\Sigma - \Psi$ k-adrangú pozitív szemidefinit mátrix felírható AA^T alakban, ahol A $(p \times k)$-s mátrix. Ezzel a lemmát bizonyítottuk. □

1. Tétel. *Legyen az x 0 várható értékű, Σ kovarianciamátrixú p-dimenziós valószínűségi változó leírható a k-faktormodellel. Tegyük fel, hogy a 3.6-os egyenlőségben szereplő Ψ ismert. Ekkor az átviteli mátrix - forgatástól eltekintve - egyértelműen meghatározott.*

Bizonyítás. Azt kell bizonyítanunk, hogy ha léteznek A és B mátrixok, melyekre

$$\Sigma - \Psi = AA^T = BB^T,$$

akkor $B = AZ$, ahol Z egy ortomormált mátrix. Írjuk fel az A mátrix szinguláris felbontását:

$$A = VSU^T.$$

A V ortonormált mátrix oszlopai legyenek $v_1, v_2, ..., v_p$. Ekkor tudjuk, hogy U oszlopait A ortogonális vektorokba viszi, ezért V oszlopait A^T viszi ortogonális vektorokba, tehát:

$$(A^T v_i, A^T v_j) = 0,$$

és ebből

$$(v_i, AA^T v_j) = 0.$$

Tehát V oszlopvektorai az $AA^T = \Sigma - \Psi$ mátrix sajátvektorai. Mivel $AA^T = BB^T$, ezért V oszlopait B^T is ortogonális vektorokba viszi, így B szinguláris felbontása

$$B = VTW^T$$

alakú. $AA^T = BB^T$ miatt

$$VS^2W^T = VT^2V^T,$$

ebből $S^2 = T^2$. Mivel a szinguláris értékek nemnegatívak és nagyság szerint rendezettek, ezért $S = T$ adódik. Így

$$B = VSW^T = VSU^TUW^T = AUW^T.$$

Mivel U és W is $(k \times k)$ortogonális mátrixok, ezért $Z = UW^T$ is ortonormált mátrix, azaz forgatás, ezzel a tételt beláttuk. □

7. Megjegyzés. *Az eddigiek alapján világos, hogy A még ismert Ψ esetén sem egyértelmű. Lawley [2]-es cikkében javasolta először azt az ötletet, hogy keressünk olyan A mátrixot, amelyre*

$$A^T C A$$

diagonális valamely C $(p \times p)$-s diagonális mátrixra, amelynek főátlójában nemnegatív elemek állnak, monoton növő sorrendben. Ha itt $C = I$ egységmátrix, akkor főfaktorizációról beszélünk. Ortogonális átviteli mátrix keresése ekvivalens a főfaktorizációval.

3.4. Alkalmazás

A faktoranalízis a változók számának csökkentésének módszere. Ahhoz, hogy alkalmazni tudjuk ezt, keresnünk kell egy olyan adathalmazt, amelyben elemek tulajdonságait többféle változó szerint mérik. Ilyen adathalmaz lehet például egy pszichológiai felmérés, amely egyes személyiségi jellemzőket vizsgál, vagy sportban csapatok statisztikáiból lehet következtetéseket levonni: melyik jellemzi leginkább a csapatok erősségét, melyik elhanyagolható, vagy melyek állnak szorosabb kapcsolatban egymással? Én most az utóbbira szeretnék egy példát mutatni.

Az amerikai futballban a csapatokat rengeteg különböző statisztika alapján állítják sorrendbe. Ilyen például a meccsenként szerzett pontok átlaga, a meccsenként megtett yardok átlaga, a sikeres harmadik próbálkozások száma, az eladott labdák száma, vagy a meccsenként passzolt yardok átlaga. A szakértők szerint ezek a statisztikák mutatják legjobban az egyes csapatok támadósorainak erejét: nyilván minél nagyobb számot látunk egy oszlopban, annál jobb a csapatnak azon mutatója.

8. Megjegyzés. *Természetesen ezek a statisztikák egyáltalán nem mutatják pontosan a csapatok sorrendjét, de elég jó képet adnak arról, hogy melyik*

csapatoknak erős a passzjátékuk, ami ma talán az egyik legfontosabb eleme a játéknak. Ahhoz, hogy teljes képet kapjunk, további statisztikákat kellene vizsgálnunk. Ilyen lehetne például a meccsenként futott yardok átlaga, a védelmek által meccsenként engedett passzolt, illetve futott yardok átlaga, vagy a szerzett touchdown-ok száma. A teljes képhez jóval több változót kellene alkalmazni, ez további munkával lehetséges, most csak a csapatok passzjátékát vizsgálom.

Ezeket szeretném jobban megvizsgálni a faktoranalízis segítségével: szeretném két faktorral magyarázni az ezen statisztikák közti összefüggéseket. Ehhez ismét az R statisztikai progamcsomagot kell segítségül hívni.

Töltsük be a 4-es ábrán látható táblázatot R-be egy *nfl* nevű változóba. A táblázat az NFL (National Football League) 32 csapatának 2014. évi statisztikáit tartalmazza. Az első oszlopban a meccsenként szerzett pontok átlaga szerepel (PPG), a másodikban a meccsenként összesen elért yardok átlaga (YPG), a harmadikban a sikeres harmadik kísérletek száma (TRD), a negyedikben az eladott és szerzett labdák számának különbsége (TO), az ötödikben pedig a meccsenként elért passzolt yardok átlaga (PYG). Ezek jellemzik leginkább egy csapat passzjátékát, ezért ezen statisztikák közötti összefüggéseket vizsgálom meg részletesebben. A táblázat sorai a YPG szerint vannak rendezve.

Miután az adatokat bevittük, meg tudjuk vizsgálni az egyes változók közötti korrelációt is. Az alábbi korrelációs mátrixot kapjuk 4 tizedesjegyre kerekítve:

$$\mathbf{R} = \begin{pmatrix} 1.0000 & 0.8767 & 0.7941 & 0.4244 & 0.6911 \\ 0.8767 & 1.0000 & 0.7493 & 0.2216 & 0.8214 \\ 0.7941 & 0.7493 & 1.0000 & 0.3179 & 0.5827 \\ 0.4244 & 0.2216 & 0.3179 & 1.0000 & 0.0362 \\ 0.6911 & 0.8214 & 0.5827 & 0.0362 & 1.0000 \end{pmatrix}$$

A mátrixból leolvasható, hogy az első három változó között nagyon magas korrelációt tapasztalhatunk, ami nem is meglepő, hiszen ahhoz, hogy egy

csapat pontokat tudjon elérni, sok yardot kell megtenniük, és sokszor kerülhetnek olyan helyzetbe, hogy harmadik kísérletre kell elérniük a továbbhaladás lehetőségét. Ami kicsit talán meglepőbb lehet az az, hogy a negyedik oszlop milyen alacsony korrelációt mutat a többivel, sőt az ötödik változóval majdhogynem korrelálatlanok. Ez azt jelenti, hogy a labdaszerzések-labdavesztések nem állnak szoros kapcsolatban a passzolt yardok mennyiségével, de még a szerzett pontok mennyiségével sem különösebben magas a kapcsolat. Ez talán azzal magyarázható, hogy itt bejön a védelmek szerepe is, valamint a labdát nemcsak passz esetén, hanem futás esetén is el lehet veszíteni. Ez is indokolja a kétfaktoros modellezés jogosságát.

Először azonban vizsgáljuk meg, hogy mit mond a 3.2-es részben leírt Kayser-Meyer-Olkin-teszt a faktoranalízis alkalmazhatóságáról. A korrelációmátrixot már ismerjük, ebből kiszámíthatóak a parciális korrelációk, ezáltal pedig a következő érték adódik:

$$KMO = 0.7735,$$

ami viszonylag jónak mondható, tehát a faktoranalízis elvégezhető.

```
M3<-factanal(nfl,fac=2, score="reg",rot="none")
plot(M3$sc,t="n")
text(M3$sc[,1],M3$sc[,2],1:32)
```

3.3. ábra. Faktoranalízis a 4-es táblázatban szereplő adatokra

A 3.3-mas algoritmussal faktoranalízist végezhetünk a 4-es táblázatban szereplő adatokra. Ennek eredményét a 4.2-es ábrán láthatjuk. Leolvasható, hogy a táblázat elején szereplő csapatok inkább az ábra jobb oldalán, míg a táblázat alján szereplő csapatok inkább az ábra bal oldalán foglalnak helyet. Ha jobban megvizsgáljuk, akkor látható, hogy a csapatok sorrendje az első faktor alapján nagyjából megegyezik a PPG alapján számított sorrenddel. A másik faktort már jóval nehezebb értelmezni, ugyanis az eszerint kiugró csapatok statisztikái között nem látni különösebb kapcsolatot. A 13-mas és

1-es csap TO mutatója nagyon alacsony, azonban például a szintén alacsony értékkel rendelkező 32-es csapat nem lóg ki a többi közül a második faktor alapján, tehát nem mondhatjuk, hogy ez főként a TO-t modellezi. A második faktor értelmezéséhez más módszert kell segítségül hívnunk.

Ehhez használjuk a főkomponensanalízist, amelynél a faktoranalízishez hasonlóan a változók számának csökkentése a cél.

```
M2<-princomp(nfl,cor=TRUE)
```

3.4. ábra. Főkomponensanalízis a 4-es táblázatban szereplő adatokra

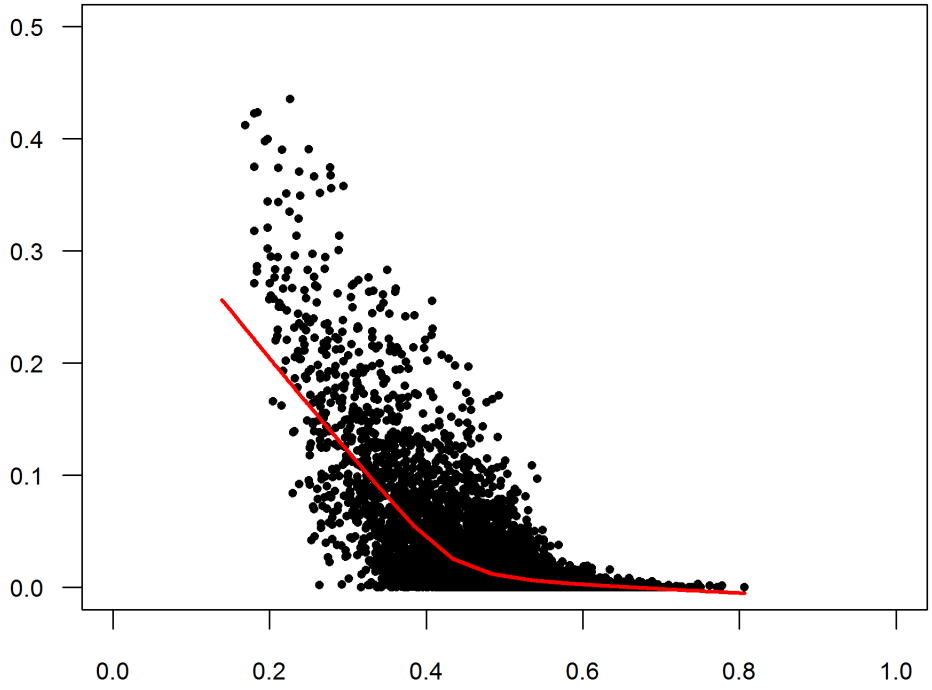
A 3.4-es algoritmussal végezzük el a főkomponensanalízist az adatainkra, majd hasonlítsuk össze a kapott eredményt a faktoranalízis eredményével. Az eredmény a 4.3-mas ábrán látható. Ha összehasonlítjuk a két ábrát, akkor nem látunk jelentős különbségeket: az egyes csapatoknak megfelelő számok nagyjából ugyanott helyezkednek el mindkét esetben. Az R-eredmény alapján a főkomponensanalízis esetében az első főkomponens a szórás 67 százalékát, az első két főkomponens pedig a szórás 87.6 százalékát magyarázza. Ezen az ábrán is jól látszik, hogy a TO-változó jól elkülönül a többitől, amelyeket az első főkomponens vagy első faktor jól modellez.

Összességében tehát azt mondhatjuk, hogy egy csapat passzjátékának erősségét a TO-változó nem jellemzi jól, azonban a másik négy változó igen.

9. Megjegyzés. *A faktoranalízis elég sok paraméterrel dolgozik. A loading mátrixon kívül az egyedi varianciák is lényegében szabad paraméterekként viselkednek. Azért dolgoztunk* 5 *változóval, mert ez a legkisebb modell, amelyben két faktor lehetséges,* 5 *változós modellben viszont* 2 *faktor a legtöbb, ami lehetséges.*

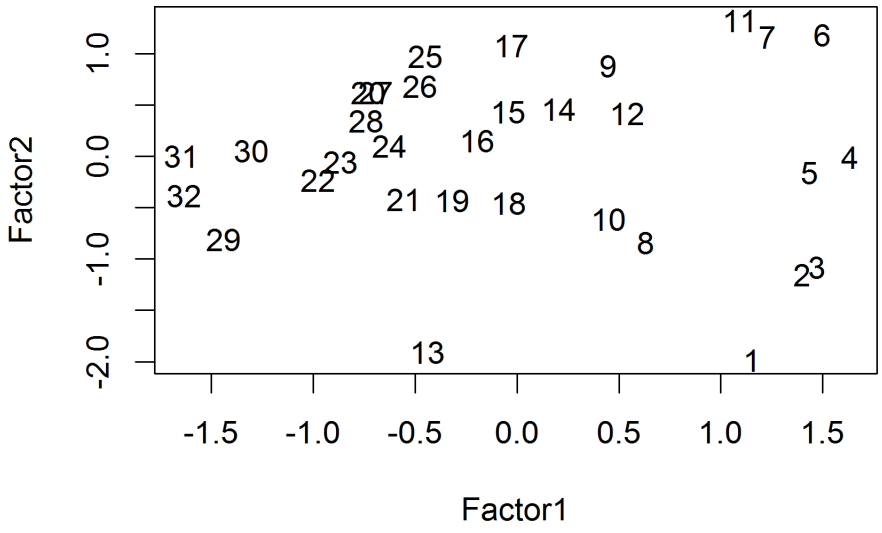
4. fejezet

Ábrák

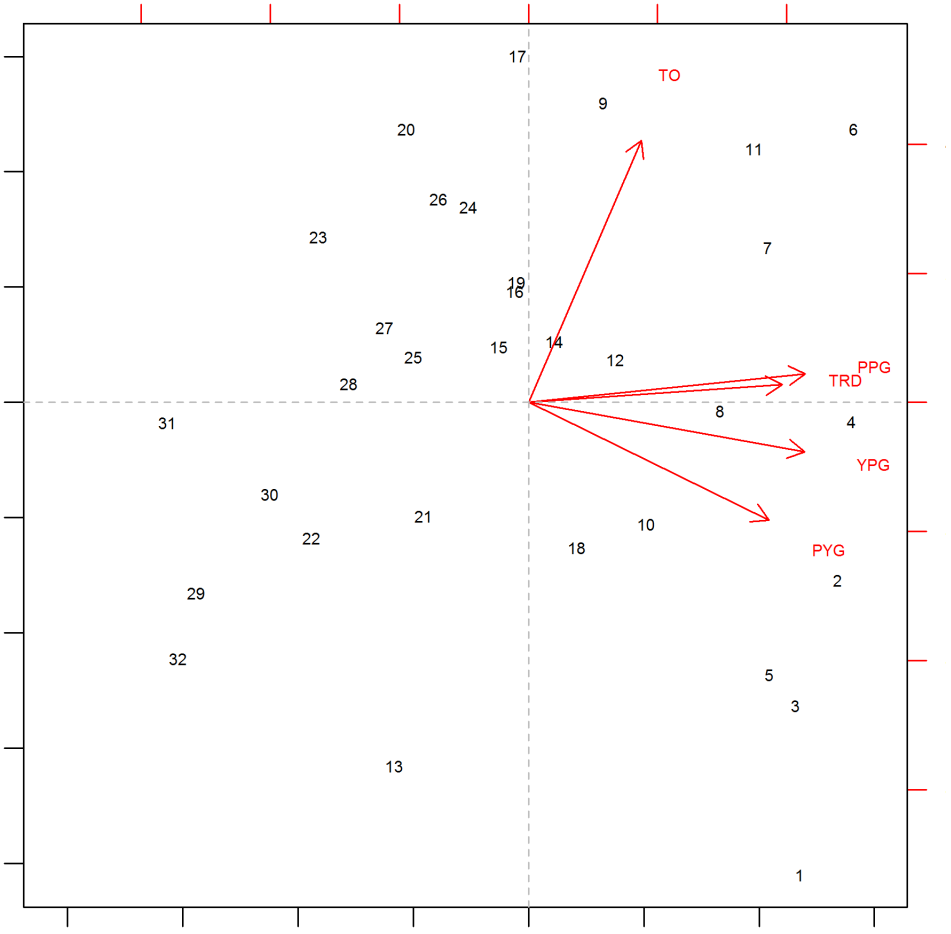


4.1. ábra. A KMO-teszt megbízhatósága

PPG	YPG	TRD	TO	PYG
25.1	411.4	48	−13	297.8
27.2	411.1	45	0	301.6
28.6	406.6	41	−5	305.9
30.1	402.9	44	5	291.3
29.6	396.8	44	−8	272.2
30.4	386.1	47	14	266.3
29.2	383.6	47	6	236.5
23.8	378.2	44	5	284.6
24.6	375.8	42	10	203.1
23.8	367.2	43	−2	267.0
29.2	365.5	44	12	257.6
25.6	364.9	41	2	238.7
18.8	358.6	32	−12	252.9
24.2	350.1	40	2	233.1
22.8	348.0	40	0	213.8
21.2	346.7	42	3	219.4
23.2	344.6	39	12	209.5
21.8	341.6	45	−5	256.1
20.1	340.8	39	7	251.9
19.1	327.4	40	7	191.4
19.9	327.1	38	−5	237.0
17.7	326.6	39	−11	184.1
18.7	324.6	30	6	216.6
19.4	319.8	40	8	238.0
22.1	318.8	40	−3	198.9
21.4	318.5	37	7	225.9
20.3	315.5	39	−1	202.8
20.2	314.7	35	−2	212.5
15.9	303.7	30	−10	213.2
17.3	292.0	37	−8	206.1
15.6	289.6	32	−6	187.6
15.8	282.2	34	−15	204.7



4.2. ábra. A 4-es táblázatban szereplő adatokra vonatkozó faktoranalízis eredménye



4.3. ábra. A 4-es táblázatban szereplő adatokra vonatkozó főkomponensanalízis eredménye

Irodalomjegyzék

[1] Vargha András. A parciális korrelációs együttható értelmezési problémái a többdimenziós normalitás feltételének sérülése esetén. *Statisztikai Szemle*, 2011. 5

[2] D. Lawley. Estimation in factor analysis under various initial assumptions. *British journal of statistical Psychology*, 11(1):1–12, 1958. 17

[3] Márkus László. Főkomponens és faktor analízis előadás, 2014. 10

[4] W. Qiu and H. Joe. *clusterGeneration: Random Cluster Generation (with Specified Degree of Separation)*, 2015. R package version 1.3.4. 12

[5] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. 4

[6] Móri Tamás és Székely Gábor. *Többváltozós statisztikai analízis*, pages 95–116. Műszaki Könyvkiadó, 1986. 14