ADATREDUKCIÓ

Olyan statisztikai módszerek tartoznak ide, melyek lehetové teszik, hogy az adatmátrix méretét csökkentve kisebb költséggel értékelhessük ki a statisztikai sokaságot.
A redukált adatmennyiségbol levont statisztikai következtetések érvényesek maradnak az eredeti statisztikai sokaságra is.
A csökkentés vonatkozhat az esetszám csökkentésére és a változók számának a csökkentésére egyaránt.

- **Klaszteranalízis**
- **Ritkítás véletlenszám generálással**
- **Faktoranalízis**
- **Fokomponens-analízis**
- **Többdimenziós skálázás (MDS)**

FAKTORANALÍZIS

Nagyszámú, sztochasztikusan erosen összefüggo változónk van. A változók redundáns információt hordoznak. Ismeretlen, kisszámú faktorváltozót keresünk.

- **Hogyan lehet a változók által közösen magyarázott információt korrelálatlan faktorokkal kifejezni?**
- **A faktorok milyen mértékben magyarázzák az eredeti változókat?**
- **Mely változók vannak ugyanazokkal a faktorokkal kifejezve?**
- **Hogyan lehet ezek alapján a változóinkat csoportosítani?**
- **Mi lehet az egyes faktorok jelentése?**

FAKTORANALÍZIS

A VÁLTOZÓK KÖZÖTTI ÖSSZEFÜGGÉS EREJÉNEK MÉRÉSE

$$KMO = \frac{\sum_{i=1}^{p}\sum_{\substack{j=1 \\ i\neq j}}^{p} r_{ij}^2}{\sum_{i=1}^{p}\sum_{\substack{j=1 \\ i\neq j}}^{p} \rho_{ij}^2 + \sum_{i=1}^{p}\sum_{\substack{j=1 \\ i\neq j}}^{p} r_{ij}^2},$$

***Kaiser-Meyer-Olkin* mérték**

$$\rho_{ij} = \left| \frac{R_{ij}}{\sqrt{R_{ii} \cdot R_{jj}}} \right|$$

parciális korrelációs együttható

$$r_{ij} = R(X_i, X_j)$$

korrelációs együttható

$0.9 \leq KMO$	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)

FAKTORANALÍZIS

A VÁLTOZÓK KÖZÖTTI ÖSSZEFÜGGÉS EREJÉNEK MÉRÉSE

$$MSA_i = \frac{\sum_{i\neq j} r_{ij}^2}{\sum_{i\neq j} \rho_{ij}^2 + \sum_{i\neq j} r_{ij}^2} \quad (i = 1,2,\ldots,p)$$

measure of sampling adequacy

Az indulási *p* db változóból azokat érdemes elhagyni, amelyeknél az MSA_i érték a legkisebb.

Elvégezheto még a Bartlett-féle gömb próba.
Itt az a nullhipotézis, hogy a vizsgált változók függetlenek egymástól. Akkor érdemes továbbmenni, ha ez a próba nem szignifikáns!

FAKTORANALÍZIS

A *k*-FAKTOROS MODELL

Adottak az $X_1, X_2, \ldots, X_p$ **változók, a belolük alkotott *p*-dimenziós vektor** $\underline{X}$ **.**

$$\underline{X} = \underline{\underline{A}} \cdot \underline{F} + \underline{U} + \underline{m}$$

$\underline{\underline{A}}$ ***pxk*-as átviteli mátrix**

$\underline{F}$ ***k*-dimenziós közös faktor-vektor**

$\underline{U}$ ***p*-dimenziós egyedi faktor-vektor**

$E\underline{X} = \underline{m}$ **várható érték vektor**

FAKTORANALÍZIS

A *k*-FAKTOROS MODELL FELTÉTELEI

$F_1, F_2, \ldots, F_k$ **páronként korrelálatlanok** $R(F_i, F_j) = 0$

$E(F_i) = 0, \quad D^2(F_i) = 1$

$U_1, U_2, \ldots, U_p$ **páronként korrelálatlanok** $R(U_i, U_j) = 0$

$E(U_i) = 0, \quad D^2(U_i) = \Psi_{ii}$

$F_1, F_2, \ldots, F_k$ **és** $U_1, U_2, \ldots, U_p$ **páronként korrelálatlanok:**

$$R(U_i, U_j) = 0$$

FAKTORANALÍZIS

A *k*-FAKTOROS MODELL FELTÉTELEI

Egy k-faktoros modell pontosan akkor oldható meg, ha

$$\underline{\underline{\Sigma}} = \underline{\underline{A}} \cdot \underline{\underline{A}}^T + \underline{\underline{\Psi}}$$

$\underline{\underline{\Sigma}} = \mathrm{cov}(X_i, X_j)$ $\underline{X}$ **kovarianciamátrixa**

$\underline{\underline{\Psi}}$ $\underline{U}$ **kovarianciamátrixa**

Van *p(p+1)/2* egyenlet, és *p(k+1)* ismeretlen

***(p+1)/2 > k+1* esetben az egyenletrendszer túldefiniált**

***(p+1)/2 < k+1* esetben az egyenletrendszer aluldefiniált**

- Maximum likelyhood módszer
- Fokomponens-analízis
- A legkisebb négyzetek módszere
- **:**

FAKTORANALÍZIS

A *k*-FAKTOROS MODELL KOORDINÁNTÁNKÉNT

$$X_i = \sum_{j=1}^{k} a_{ij} F_j + U_i + m_i$$

F **koordinátái mindegyik X_i eloállításában szerepelnek**

U **koordinátái közül csak U_i szerepel X_i eloállításában**

$$D^2 X_i = \sum_{j=1}^{k} a^2_{ij} + \Psi_{ii}$$

Az X_i varianciája

A kummuláns

Az egyedi variancia

$\frac{\sum_{j=1}^{k} a^2_{ij}}{D^2 X_i}$ **Ez az arány azt fejezi ki, hány %-ot magyaráznak a közös faktorok.**

FAKTORANALÍZIS

A FAKTOROK FORGATÁSA (ROTÁCIÓ)

$$\underline{\underline{\Sigma}} = \underline{\underline{A}} \cdot \underline{\underline{A}}^T + \underline{\underline{\Psi}} = \underline{\underline{A}}\underline{\underline{E}}\underline{\underline{A}}^T + \underline{\underline{\Psi}} = \underline{\underline{A}}\underline{\underline{G}}\underline{\underline{G}}^T\underline{\underline{A}}^T + \underline{\underline{\Psi}} = \underline{\underline{A}}^* \cdot \underline{\underline{A}}^{*T} + \underline{\underline{\Psi}}$$

$\underline{\underline{A}}^* = \underline{\underline{A}}\underline{\underline{G}}$ **Az új átviteli mátrix**

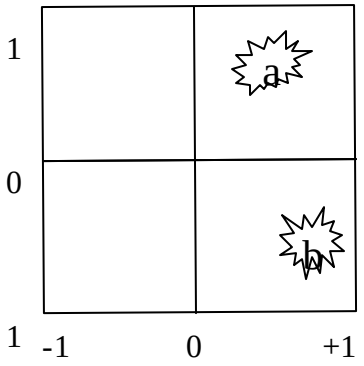
$\underline{F}^* = \underline{G}\underline{F}$ **Az új faktorvektor**

***G* ügyes megválasztásával a modell jobban magyarázható lesz!**

- Varimax — **azon változók száma kevés lesz, melyekhez sok faktor szerepel nagy súllyal**
- Quartimax — **a magyarázó faktorok számát minimalizálja**
- Equamax — **a két eljárás keverékét végzi**

A FAKTOROK FORGATÁSA (ROTÁCIÓ)

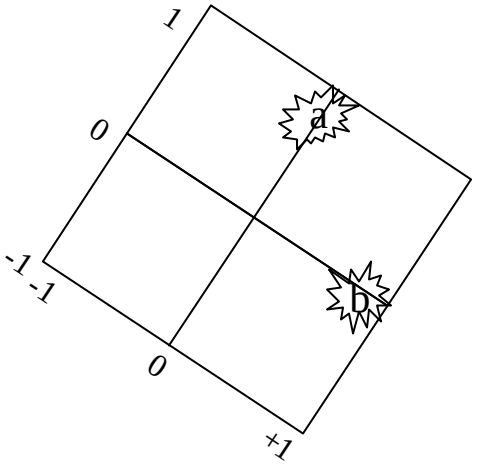
A rotáció szemléltetése egy egyszeru kétdimenziós példán:



Az eredeti változók **a** és **b** csoportja a rotáció nélkül kapott mindkét faktoron jelentos faktorsúllyal rendelkezik.

A FAKTOROK FORGATÁSA (ROTÁCIÓ)

A rotáció szemléltetése egy egyszeru kétdimenziós példán:



Az eredeti változók **a** csoportja csak a rotációval kapott egyik, a **b** csoport pedig csak a másik faktoron rendelkezik jelentos faktorsúllyal.

FOKOMPONENSANALÍZIS

A faktoranalízis speciális esete. Dimenziószám csökkentésre használható. Az eredetileg p változóval jellemzett statisztikai sokaságot k<<p változóval (fokomponensekkel) jellemezzük. A k-dimenziós statisztikai elemzések következtetései a p-dimenziós sokaságra is érvényesek lesznek. Ezzel jelentos költséget lehet megtakarítani.
Lehetoség van a p>3 dimenziós sokaságot (ha k<4) pontfelho grafikonon szemléltetni.
A fokomponensek terében a változók korrelálatlanok lesznek.

A fokomponens-transzformáció:

$$\underline{F} = \underline{\underline{G}}^T (\underline{X} - \underline{m}) \quad \text{Þ} \quad \underline{X} = \underline{\underline{G}} \cdot \underline{F} + \underline{m}$$

$\underline{F} = (F_1, F_2, \ldots, F_p)$ **a fokomponens-vektor**

$\underline{\underline{G}} = (\underline{g}_1, \underline{g}_2, \ldots, \underline{g}_p)$ **a foirányok mátrixa**

FOKOMPONENSANALÍZIS

A FOKOMPONENS-MODELL TULAJDONSÁGAI

• **A fokomponensek korrelálatlanok:** $R(F_i, F_j) = 0$

• A

$E(F$ 0

• A

$T = \sum_{i=1} D^2 x_i = \sum_{i=1} \lambda_i$

$\frac{\lambda_i}{T}$ **megmutatja, hány %-ot magyaráz** F_i

A fokomponensek csökkeno jelentoséguek:

F $_1$ **magyaráz a legtöbbet,** F_2 **a második legtöbbet,...,** F_p **magyaráz a legkevesebbet** ***T*****-bol.**

FOKOMPONENSANALÍZIS

A FOKOMPONENS-MODELL TULAJDONSÁGAI

• **A foírányok jelentése:** $\underline{\underline{\Sigma}} \underline{g}_i = \lambda_i \underline{g}_i$

$\underline{g}_1$ **ebben az irányban a legnagyobb a variancia**

$\underline{g}_2$ **ebben az irányban a legnagyobb a variancia a** $\underline{g}_1$ **irányra meroleges irányok között**

$\vdots$

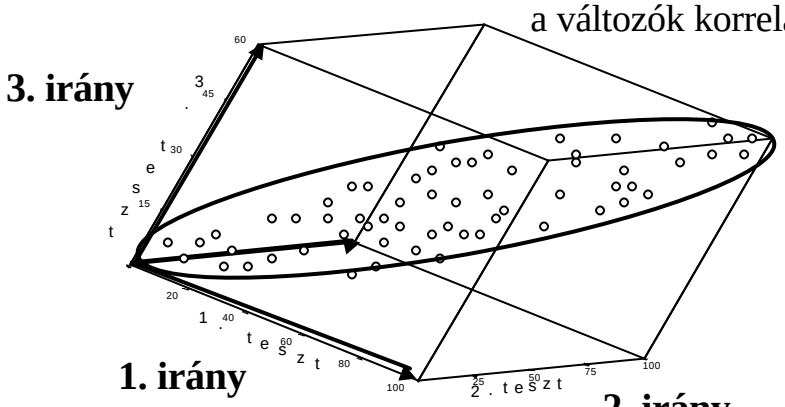
• **Dimenziócsökkentés:**

Ha $\underline{X}$ **helyett az elso k fofaktor-alkotta vektorral számolunk,**

az elvesztett információ csupán $1 - \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{T}$

FOKOMPONENSANALÍZIS

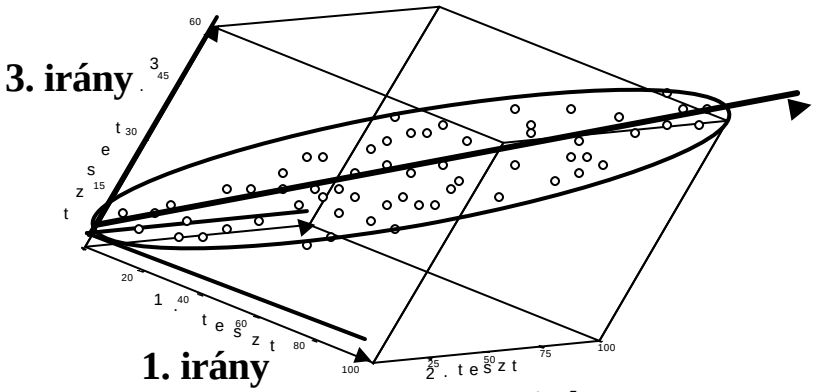
Tengelyek nem derékszögeket zárnak be: a változók korreláltak!



FOKOMPONENSANALÍZIS

Ebben az irányban tudunk legjobban differenciálni a pontok között. A fokomponensek hosszát (fontosságát) az ún. sajátértékkel (*eigenvalue*) jellemezzük, ami az értelmezett variancia.

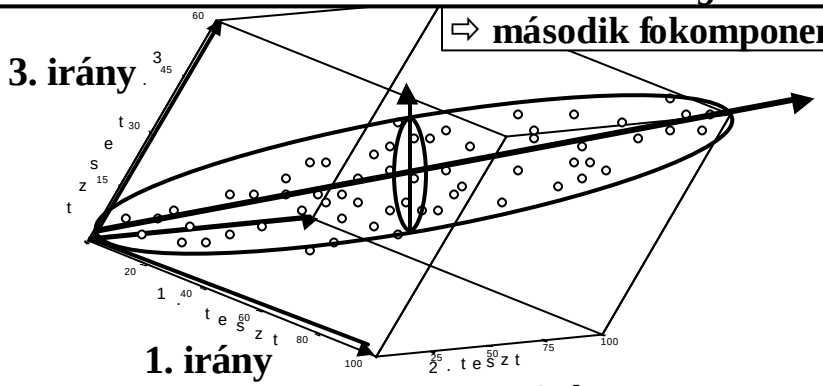
első fokomponens



FOKOMPONENSANALÍZIS

Az eljárást folytatni lehetne a harmadik fokomponens megkeresésével, de ennek a konkrét esetben már nincs értelme, mivel ebben az irányban már jelentéktelen a szóródás ⇨ ***az adatok leírására 2 dimenzió elegendő!***

⇨ **második fokomponens**



TÖBBDIMENZIÓS SKÁLÁZÁS

Egy p-dimenziós sokaságot lehet egy k=1,2 vagy 3 dimenziós Euklideszi ponthalmazzal vizualizálni.
A ponthalmaz távolságviszonyai az eredeti sokaság eseteinek távolságviszonyaival nagymértékben egyezik.

A vizualizálás révén tanulmányozható a statisztikai sokaság térbeli struktúrálódása. Jellegzetes tömörülések, irányok fedezhetők fel az elkészült scatter-grafikonon.

$$\underline{\underline{X}} = \begin{pmatrix} x_1^1 & x_1^2 & \cdots & x_1^p \\ x_2^1 & x_2^2 & \cdots & x_2^p \\ \vdots & \vdots & \ddots & \vdots \\ x_n^1 & x_n^2 & \cdots & x_n^p \end{pmatrix} \quad \begin{matrix} \underline{x}^{(1)} & \text{1. eset} \\ \underline{x}^{(2)} & \text{2. eset} \\ \cdot & \\ \underline{x}^{(n)} & \text{n. eset} \end{matrix}$$

TÖBBDIMENZIÓS SKÁLÁZÁS

$\underline{\underline{D}} = (d(\underline{x}^{(i)}, \underline{x}^{(j)}))$ **Az esetvektorok egymástól vett *nxn*-es távolságmátrixa**

Megkonstruálhatók olyan $\underline{z}_1, \underline{z}_2, \ldots, \underline{z}_n$ ***k*=1,2 vagy 3 dimenziós vektorok, melyek *nxn*-es Euklideszi távolság-mátrixa nagymértékben hasonló** $\underline{\underline{D}}$ **-hez.**

$\sum_{i=1}^{n}\sum_{j=1}^{n}\left(d(\underline{x}^{(i)}, \underline{x}^{(j)}) - \rho(\underline{z}_i, \underline{z}_j\right)^2$ **„kicsi”**