4. előadás - Többváltozós statisztikai alapok

Sztochasztikus rendszerek matematikája

2016. szeptember 26.

Alapfogalmak: minta és adatmátrix

- Legyen $\underline{x} = (x_1, \dots, x_p)^\intercal$ p-dimenziós valószínűségi vektorváltozó, azaz olyan véletlen vektor, melynek minden koordináta-változója egy-egy 1-dimenziós valószínűségi változó. Tegyük fel, hogy adott n darab $\underline{x}^1, \dots, \underline{x}^n \in \mathbb{R}^p$ független realizáció. Ez a p-**dimenziós** n **elemű i.i.d. minta**.

- Azt a $p \times n$ dimenziós mátrixot, melynek oszlopai a mintaelemek, **adatmátrix**nak nevezzük. A pontos jelölés:

$$X = \begin{pmatrix} x_1^1 & x_1^2 & \dots & x_1^n \\ x_2^1 & x_2^2 & \dots & x_2^n \\ \vdots & \vdots & & \vdots \\ x_p^1 & x_p^2 & \dots & x_p^n \end{pmatrix}$$

Alapfogalmak: várható érték vektor

- Tegyük fel, hogy az $\underline{X}$ p-dimenziós valószínűségi vektorváltozó minden komponensének létezik a várható értéke. Ekkor

$$\underline{\mu} = (\mu_1, \dots, \mu_p)^\intercal = \mathrm{E}\underline{X} = (\mathrm{E}X_1, \dots, \mathrm{E}X_p)^\intercal$$

  a v.v.v. **várható érték vektora**.

- n-elemű statisztikai minta esetén $\underline{\mu}$ mintából való becslése az **átlagvektor**, mely a komponensek átlagainak vektora. Azaz

$$\overline{\underline{X}} = \frac{1}{n} \sum_{i=1}^{n} \underline{X}_i$$

  A becslés torzítatlan és erősen konzisztens.

Alapfogalmak: kovariancia mátrix

- Legyen az $\underline{X}$ p-dimenziós valószínűségi vektorváltozó olyan, mely minden komponensének létezik a második momentuma. A Cauchy-Schwarz egyenlőtlenség alapján ekkor bármely két komponense szorzatának is létezik a várható értéke, így a kovarianciájuk is. Az ezen értékekből felépített mátrix lesz $\underline{X}$ **kovariancia-mátrixa**, azaz

$$\Sigma = (\sigma_{ij})_{i,j=1}^{p} = \mathrm{E}[(\underline{X} - \mathrm{E}\underline{X})(\underline{X} - \mathrm{E}\underline{X})^\intercal] = (\mathrm{cov}(X_i, X_j))_{i,j=1}^{p}$$

- Σ n-elemű mintából való becslése a **tapasztalati kovariancia-mátrix**:

$$S_n = \frac{1}{n}\sum_{i=1}^{n}(\underline{X}_i - \overline{\underline{X}})(\underline{X}_i - \overline{\underline{X}})^\intercal$$

A korrigált változat:

$$S_n^* = \frac{n}{n-1} S_n$$

Alapfogalmak: szimmetria és definitség

Mind S_n, mind S_n^* szimmetrikus és pozitív szemidefinit, azaz

- $S_n^\top = S_n$ és $S_n^{*\top} = S_n^*$
- bármely $\underline{a} \in \mathbb{R}^p$ esetén $\underline{a}^\top S_n \underline{a} = \sum_{i=1}^{p} \sum_{j=1}^{p} a_i a_j (S_n)_{i,j} \geq 0$ (S_n^*-ra hasonlóan)

Ebből következik, hogy létezik ún. **spektrálfelbontás**uk, azaz

$$S_n = V \Lambda V^\top = \sum_{i=1}^{p} \lambda_i v_i v_i^\top$$

ahol $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ a sajátértékek diagonális mátrixa, $V = (\underline{v}_1, \ldots, \underline{v}_p)$ a normált sajátvektorok alkotta ortogonális mátrix (azaz $VV^\top = V^\top V = I$). A sajátértékek a szimmetria miatt valósak, a pozitív szemidefinitség miatt pedig nemnegatívak. Ugyanez igaz S_n^* esetén is.

Többdimenziós standard normális eloszlás

Definíció

Legyenek $\xi_1, \ldots, \xi_p$ független $N(0,1)$ eloszlású valószínűségi változók. Ekkor a $\xi = (\xi_1, \ldots, \xi_p)^\intercal$ valószínűségi vektorváltozót p-**dimenziós standard normális eloszlású** v.v.v-nak nevezzük. Jelölés: $\xi \sim \mathcal{N}_p(\underline{0}, I_{p\times p})$.

- $\mathrm{E}\xi = \underline{0}$, hiszen $\mathrm{E}\xi_i = 0$ minden $i = 1, \ldots, p$ esetén.
- $\Sigma_\xi = I_{p\times p}$, hiszen $D^2(\xi_i) = 1$ minden $i = 1, \ldots, p$ esetén és $\mathrm{cov}(\xi_i, \xi_j) = 0$ minden $i \neq j$ esetén.
- Sűrűségfüggvény:

$$f_\xi(\underline{x}) = \frac{1}{(\sqrt{2\pi})^p} \exp\left\{-\frac{1}{2}\underline{x}^\intercal \underline{x}\right\}$$

ahol $\underline{x} = (x_1, \ldots, x_p)^\intercal \in \mathbb{R}^p$.

Többdimenziós normális eloszlás

Definíció

Legyenek $\underline{m} \in \mathbb{R}^p$, $A \in \mathbb{R}^{p\times p}$ és $\xi \sim \mathcal{N}_p(\underline{0}, I_{p\times p})$. Ekkor az $\eta = A\xi + \underline{m}$ valószínűségi vektorváltozót p-**dimenziós normális eloszlású** v.v.v-nak nevezzük. Jelölés: $\eta \sim \mathcal{N}_p(\underline{m}, \Sigma_\eta)$.

- $\mathrm{E}\eta = \underline{m}$.
- $\Sigma_\eta = AA^\intercal$ szimmetrikus pozitív szemidefinit mátrix.
- Ha Σ_η invertálható, akkor létezik sűrűségfüggvény:

$$f_\eta(\underline{y}) = \frac{1}{(\sqrt{2\pi})^p \cdot (\det\Sigma_\eta)^{1/2}} \exp\left\{-\frac{1}{2}(\underline{y}-\underline{m})^\intercal \Sigma_\eta^{-1}(\underline{y}-\underline{m})\right\}$$

ahol $\underline{y} = (y_1, \ldots, y_p)^\intercal \in \mathbb{R}^p$.

Fordított irány

A dolog visszafelé is működik, azaz ha $\eta \sim \mathcal{N}_p(\underline{m}, \Sigma_\eta)$ adott és Σ_η invertálható, akkor vegyük a

$$\Sigma_\eta = AA^\intercal$$

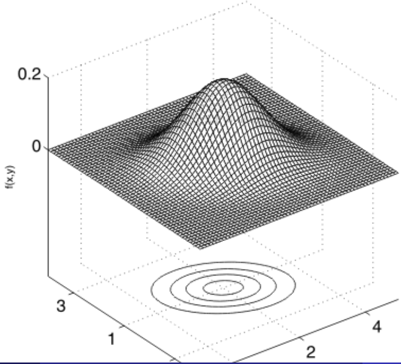
felbontást. Ez létezik, hiszen a spektrálfelbontásból $A = V\Lambda^{1/2}$ például jó választás lenne. A baj az, hogy ez még nem lesz egyértelmű felírás. Viszont, mivel Σ_η szimmetrikus és pozitív definit (hiszen feltettük, hogy invertálható), így az $A = V\Lambda^{1/2}V^\intercal$ felírás már egyértelmű. Ezt szokás $\Sigma_\eta^{1/2}$ jelöléssel is használni. Ekkor könnyen látszik, hogy a

$$\xi = A^{-1}(\eta - \underline{m})$$

lineáris transzformáció **standard normális eloszlású** valószínűségi vektorváltozót eredményez.

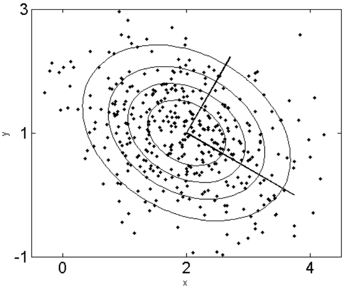
Szemléltetés: 2-dimenzió

A kétdimenziós standard normális eloszlás sűrűségfüggvényének képe egy harang alakú felület, mely tulajdonképpen a haranggörbe saját tengelye körüli megforgatottja. A szintvonalak körök lesznek. (Azaz függetlenek a komponensek és azonos szórásúak.)



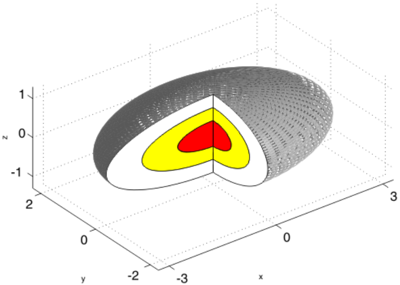
Szemléltetés: 2-dimenzió

Nem standard esetben ennek eltorzítottját kapjuk, melynek szintvonalai ekkor ellipszisek lesznek, $\underline{m}$ középponttal és $\sqrt{\lambda_1}, \sqrt{\lambda_2}$ hosszúsággal arányos, $\underline{v}_1, \underline{v}_2$ irányú tengelyekkel. Ezek az ún. koncentráció ellipszisek, melyek azt mutatják meg, hogy belülről kifelé haladva egyre kisebb valószínűséggel esnek a pontok az ellipszisekbe.



Szemléltetés: 3-dimenzió

Ebben az esetben a felületet már nem tudjuk felrajzolni, csak a szintvonalakat: azon pontok, melyeken a sűrűségfüggvény azonos értéket vesz fel, egy-egy ellipszoidon helyezkednek el. Ezek a koncentráció ellipszoidok, $\underline{m}$ középponttal és $\sqrt{\lambda_1}, \sqrt{\lambda_2}, \sqrt{\lambda_3}$ hosszúsággal arányos, $\underline{v}_1, \underline{v}_2, \underline{v}_3$ irányú tengelyekkel. Olyan, mint egy héjszerkezet, középen a legsűrűbb az anyag, kifelé pedig folyamatosan ritkul.



Paraméterbecslés: ML-módszer

Legyen $\underline{X}_1, \dots, \underline{X}_n$ független minta az $\underline{X} \in \mathcal{N}_p(\underline{m}, C)$ vektorra, és tegyük fel, hogy $n \geq p$.

Feladat: a paraméterek, azaz $\underline{m}$ és C becslése, feltéve, hogy C pozitív definit.

Eszköz: maximum-likelihood (ML) becslés

Likelihood függvény:

$$L_{\underline{m},C}(\underline{X}_1, \dots, \underline{X}_n) =$$

$$= \prod_{k=1}^{n} \frac{1}{(\sqrt{2\pi})^p \cdot (\det C)^{1/2}} \exp\left\{-\frac{1}{2}(X_k - \underline{m})^\intercal C^{-1}(X_k - \underline{m})\right\} =$$

$$= \frac{1}{(2\pi)^{\frac{np}{2}} (\det C)^{\frac{n}{2}}} \cdot \exp\left\{-\frac{1}{2}\sum_{k=1}^{n}(\underline{X}_k - \underline{m})^\intercal C^{-1}(\underline{X}_k - \underline{m})\right\}$$

Az ML-becslés tulajdonságai

- A várható érték becslése torzítatlan, míg a kovariancia-mátrix becslése aszimptotikusan torzítatlan, a korrigált változat persze torzítatlan lesz.
- A becsléseink konzisztensek és hatásosak.
- Alsó korlát adható az

$$I_n(\theta_0) = \frac{1}{n}\mathrm{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\ln L_{\theta_0}(\underline{X})\right)\left(\frac{\partial}{\partial\theta}\ln L_{\theta_0}(\underline{X})\right)^\top\right]$$

Fisher-információs mátrix segítségével, a

$$D_\theta^2(T) \geq I_n(\theta_0)^{-1}, \quad \theta \in \Theta$$

Cramér-Rao egyenlőtlenség alapján, ahol θ_0 és θ az igazi és a futó paramétervektorokat jelölik.

Főkomponens-analízis

Azt vizsgáljuk, hogyan lehet nagy dimenziós megfigyelésekből kisebb dimenziósat készíteni úgy, hogy ne veszítsünk sok információt. Mivel Σ szimmetrikus és pozitív definit,

$$\Gamma^{\mathsf{T}}\Sigma\Gamma = \Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix},$$

ahol $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ Σ sajátértékei. Legyen $\underline{\gamma}_{(i)}$ a Γ i-edik oszlopa.

Definíció

Az $\underline{Y} = \Gamma^{\mathsf{T}}(\underline{X} - \underline{\mu})$ vektort főkomponensvektornak, $Y_{(i)} = \underline{\gamma}_{(i)}^{\mathsf{T}}(\underline{X} - \underline{\mu})$ i-edik főkomponensnek nevezzük.

Tulajdonságok

- $EY_{(i)} = 0$
- $D^2 Y_{(i)} = \lambda_i$
- $cov(Y_{(i)}, Y_{(j)}) = 0$, ha $i \neq j$
- $\sum_{i=1}^{p} D^2 Y_{(i)} = tr(\Sigma)$
- $\prod_{i=1}^{p} D^2 Y_{(i)} = |\Sigma|$

Állítás

$$D^2 Y_{(1)} = \max_{\underline{a}^\intercal \underline{a}=1} D^2(\underline{a}^\intercal \underline{X}) = \max_{\underline{a}^\intercal \underline{a}=1} \underline{a}^\intercal \Sigma \underline{a}$$

$$D^2 Y_{(k+1)} = \max_{\underline{a}^\intercal \underline{a}=1, \quad cov(\underline{a}^\intercal \underline{X}, Y_{(i)})=0 \quad (i=1,\ldots,k)} D^2(\underline{a}^\intercal \underline{X})$$

Következmény

Tehát ha adott egy 100 dimenziós vektorunk, és csak 5 dimenziósat szeretnénk használni, abból az eredeti a legjobban akkor állítható vissza, ha ez az 5-dim magyarázóvektor az első 5 főkomponensből áll :

Tétel

Legyen $\sigma_i^2 = \min_{\alpha,\underline{\beta}} E\left(X_i - \alpha - \sum_{j=1}^{k} \beta_i \underline{b}_i^\intercal \underline{X}\right)^2$. *Ekkor a* $\sum_{i=1}^{p} \sigma_i^2$ *hibát minimalizáló* k*-dimenziós* $(\underline{b}_1^\intercal \underline{X}, \ldots, \underline{b}_k^\intercal \underline{X})^\intercal$ *vektor az első* k *főkomponensből áll.*

Végrehajtás a gyakorlatban

- A $\frac{\lambda_1+\ldots+\lambda_k}{\lambda_1+\ldots+\lambda_p}$ hányadost variancia-hányadrésznek nevezzük, általában k-t úgy választjuk, hogy ez nagy legyen (szokásos értéke 0,9).
- A főkomponensanalízis nem skalárinvariáns, ezért szokás a kovarianciamátrix helyett a korrelációs mátrixszal elvégezni.
- Ha Σ sajátértékei különbözőek, akkor ML-becslésük S sajátértékeivel és sajátvektoraival (V oszlopaival ...)

Faktormodell

Definíció

$\underline{X}$ leírható a k-faktormodellel, ha

$$\underline{X} = A\underline{F} + \underline{W} + \underline{\mu},$$

ahol A egy $p \times k$-as rögzített mátrix, $\underline{F}$ k-dim. v.v.v. és $\underline{W}$ p-dim v.v.v. $\underline{0}$ várható érték vektorral, a vektorok komponensei korrelálatlanok, $\underline{F}$ komponenseinek szórása 1, formálisan

$$\begin{aligned} E\underline{F} &= \underline{0}, & E\underline{F}\,\underline{F}^\intercal &= I_k \\ E\underline{W} &= \underline{0}, & E\underline{W}\,\underline{W}^\intercal &= D \\ E\underline{F}\,\underline{W}^\intercal &= 0. \end{aligned}$$

$\underline{F}$ a közös faktorok, $\underline{W}$ az egyedi faktorok.

Koordinátánként

$$X_i = \sum_{i=1}^{k} A_{ij} F_j + W_i + \mu_i.$$

Ekkor X_i varianciájára

$$\Sigma_{ii} = \sigma_i^2 = \sum_{j=1}^{k} A_{ij}^2 + D_{ii},$$

$$\Sigma = AA^{\mathsf{T}} + D.$$

Tehát X_i varianciájából a $\sum_{j=1}^{k} A_{ij}^2$ részt magyarázzák a közös faktorok (kommunalitás), D_{ii} pedig az egyedi variancia.