RDD - 02

22 February 2021 09:26

Minule: $E(X) = \mu$ ($p \times 1$)

PCA – teoreticky: $Cov(X) = \Sigma = U \Lambda U^T$: spekt. rozk.

$Y = U^T(X - \mu)$ – vektor hl. komponentov : súradnice X vzhľadom na $u_1, \dots, u_p$

($p \times 1$) $\rightarrow Y_j$ – j-ty hl. komponent

Veta: $D(Y_1) \geq D(h^T X)$ $\forall h \in \mathbb{R}^p$, $\|h\| = 1$. ($Y_1 = u_1^T X$)

Pre $j > 1$: $D(Y_2) \geq D(h^T X)$ $\forall h \in \mathbb{R}^p$: $\|h\| = 1$, $h \perp u_1$ ($Y_2 = u_2^T X$)

$D(Y_3) \geq \dots$, $h \perp u_1, u_2$, $h \perp u_3$ ($Y_3 = u_3^T X$)

Predpokladáme: $\mu = 0$ (pre jednoduchosť)

prakticky: $X_1, \dots, X_n$: náh. výber z rozdelenia s $E = \mu$ (neznáme) a $Var = \Sigma$ (neznáme)

realizácie: $x_1, \dots, x_n \in \mathbb{R}^p$

$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$, $\hat{\Sigma} = S = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T$: výberová kovariančná matica

q_j : vl. vektory S $\rightarrow$ $Q = (q_1, \dots, q_p)$: $p \times p$; ortog.

v_j : vl. hodnoty S

- ak $X_i \sim N(\mu, \Sigma)$, Σ-reg., potom $q_j \xrightarrow{n \to \infty} u_j$ $\forall j$

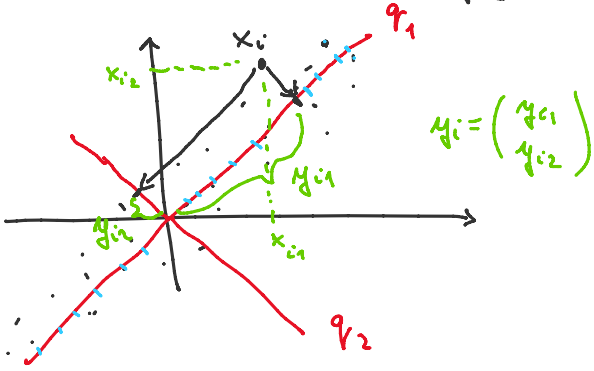
$v_j \rightarrow \lambda_j$

teoret.: $Y = U^T(X - \mu)$

výberovo: $x_i - \bar{x} = Q \cdot y_i$ $/ Q^T$ ($p \times 1$)

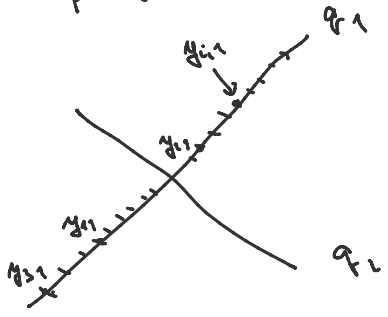
$y_i = Q^T(x_i - \bar{x})$ $i = 1, \dots, n$ ($p \times 1$, $p \times p$, $p \times 1$)

- y_i – súradnice centrovaného x_i vzhľadom na $q_1, \dots, q_p$



$y_1, \dots, y_n$ – **skóre**

$\begin{pmatrix} y_{1,j} \\ y_{2,j} \\ \vdots \\ y_{n,j} \end{pmatrix}$ – vektor skóre pre j-ty hl. komponent



$y_i = Q^T(x_i - \bar{x})$

Predp.: $x_1, \dots, x_n$ sú centrované: $\bar{x} = 0_n$

↳ častý prvý krok analýz

↳ $\tilde{x}_i = x_i - \bar{x} \quad \forall i$

↳ očistenie o priemer

$\Rightarrow \quad y_i = Q^T x_i \quad \forall i$

$$(y_1, \dots, y_n) = (Q^T x_1, \dots, Q^T x_n) \qquad \underset{n \times k}{X} = \begin{pmatrix} - x_1^T - \\ \vdots \\ - x_n^T - \end{pmatrix}$$

$$\underbrace{(y_1, \dots, y_n)}_{\text{ozn. } Y^T} = Q^T \underbrace{(x_1, \dots, x_n)}_{X^T}$$

↳ ozn. $\underset{n \times k}{Y} = \begin{pmatrix} - y_1^T - \\ \vdots \\ - y_n^T - \end{pmatrix} \qquad \Rightarrow \quad Y^T = Q^T X^T$

$$\boxed{Y = XQ}$$

$x_1, \dots, x_n \in \mathbb{R}^k \longrightarrow y_1, \dots, y_n \in \mathbb{R}^k$

Cieľ: redukovať dimenziu – vezmeme iba niektoré hl. komp.: tie s najvyššou varianciou

$$\underset{\lambda_1}{D(Y_1)} \geq \underset{\lambda_2}{D(Y_2)} \geq \dots \geq \underset{\lambda_k}{D(Y_k)} \qquad \mathbb{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_k \end{pmatrix}$$

→ vezmeme prvých q súradnice $Y_1, \dots, Y_q$

$$\alpha_j = \frac{\lambda_1 + \dots + \lambda_j}{\lambda_1 + \dots + \lambda_k} \in [0,1], \quad j = 1, \dots, k$$

↑ časť celkovej variancie, ktorú vysvetľuje prvých j hlavných komponentov

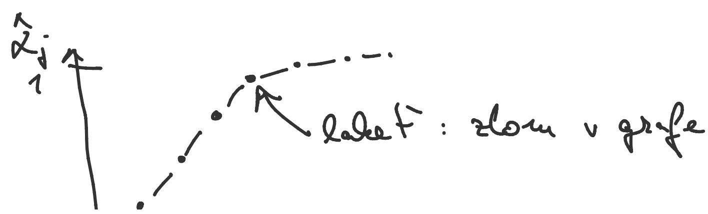
– vysvetlená časť variancie

výberovo $\hat{\alpha}_j = \frac{\nu_1 + \dots + \nu_j}{\nu_1 + \dots + \nu_k}$ ($\hat{\alpha}_j$ konverguje k α_j pre $N(\,)$ dáta)

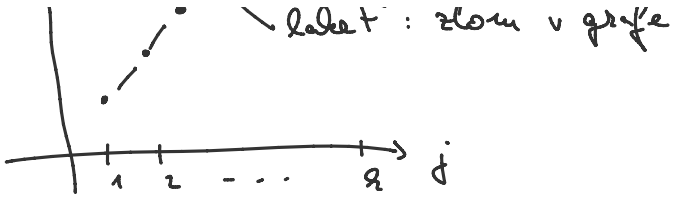
voľba # hl. komponentov:

1) elbow (lakťový) diagram

– vezmem tie hl. komponenty, ktoré sú do bodu zlomu.



- vezmem tie hl. komponenty, ket. sú do bodu zlomu



2.) najmenšie j také, že $\hat{\lambda}_j \geq c$
(pre zvolené c, napr. 0,9)

3.) Kaiserovo pravidlo: najmenšie j, že $\hat{\lambda}_j \geq$ priemer $(\hat{\lambda}_1, \dots, \hat{\lambda}_q)$

Rko:

food.txt - zloženie jedál
- pod. elipce: koľko tukov (celkov, ...) obsahuje na 1g jedla
↳ tieto uvažujeme

- PCA nie je invariantná na zmenu mierky
(ak zmeníme ju na cm, zmení sa výsledok)
→ pokus o riešenie: dáta normalizujeme / štandardizujeme
- predelíť každú premennú odmocninou z jej disperzie ($X_i \to X_i / \sqrt{D(X_i)}$)
- potom majú všetky disperzie 1 ($D(X_i) = 1 \ \forall X_i$)
- vlastne pracujeme s $R = \widehat{\text{cor}}(X)$ namiesto S
→ ak sú všetky premenné rovnakého typu a v rovnakej mierke, tak to problém nie je

- reali sme: $u_j = \begin{pmatrix} u_{j1} \\ \vdots \\ u_{jk} \end{pmatrix} \to \sim S(X_1, Y_j)$ $\to \sim S(X_k, Y_j)$

→ PC1: "tiež" táp. korelácia s Energiou, Tukmi a Nasýt. tukmi
⇒ jedlo s vysokou hodnotou PC_1 (s vysokým 1. zložkou)
bude menej energetické a menej mastné

→ PC2: viac uhľovodíkov, menej cholest. } jedlo s vysokou PC2
viac energie, menej bielkovín } sú skôr uhľovodíkové

Aplikácie PCA:
- vykreslenie, menejrozmerná reprezentácia
- vstup do iných metód: napr. zníženie # vysvetľujúcich prem. v regresii
- Principal component regression