Metoda hlavních komponent

Václav Adamec
vadamec@mendelu.cz

Vícerozměrná data

- Extenze univarietních dat na více proměnných (p)
- Datová matice: n x p
- Hodnoty proměnných získány z jednoho subjektu (i)
  Předpoklad závislostí mezi proměnnými
- Rozsah MV souboru: n
- Studium MV souborů: umělé proměnné vzniklé lineární funkcí původních proměnných

$$x = w_1 y_1 + w_2 y_2 + \ldots + w_p y_p$$

- Váhy w_i zvoleny podle různých kritérií

Multivarietní rozdělení

- Y ~ $MVN_p(\mu,\Sigma)$; μ vektor populačních průměrů; Σ populační matice kovariancí
- Funkce MVN pravděp. hustoty:

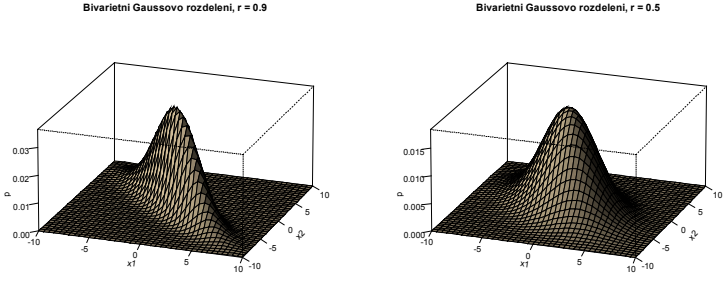
$$f(y) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{0.5}} \cdot e^{-(y-\mu)^T \Sigma^{-1}(y-\mu)/2}$$

- Mahalanobisova vzdálenost:

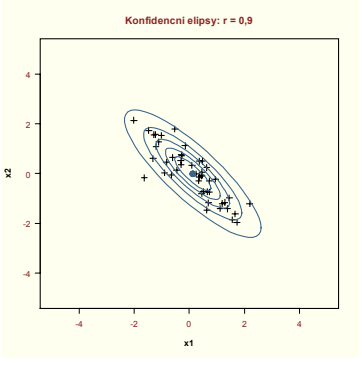
$$\Delta^2 = z^T z = (y-\mu)^T \Sigma^{-1} (y-\mu) \sim \chi^2_p$$

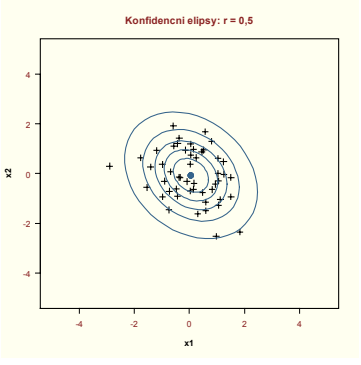
- Determinanty: Generalizovaná variance |S|, |Σ|
- Malý |Σ| výskyt kolinearity (lineárních závislostí)
- Velký |Σ| absence kolinearity (lineárních závislostí)

Bivarietní rozdělení



Konfidenční elipsy





Multivarietní rozdělení

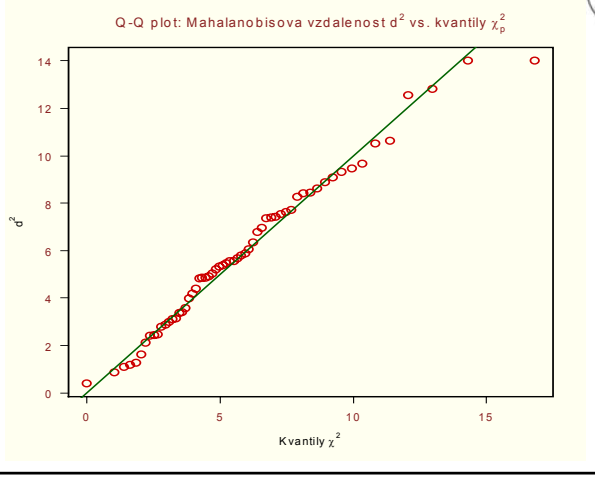
- Tvar elipsy MV rozdělení je dán Δ^2
- Hlavní podélná osa funkcí největšího charakteristického čísla λ_{max}
- Vedlejší příčná osa funkcí nejmenšího charakteristického čísla λ_{min}
- Univarietní normalita neznamená multivarietní normalitu
- Testy MVN problematičtější:
  - Testy elipsoidního tvaru bivarietních rozdělení
  - Multivarietní Q-Q plot
  - Omezeně testy (multivarietní SW test, atd.)

Ilustrační data

- Kraniální míry fotbalistů (Rencher, 1995):

V2	šířka hlavy
V3	obvod hlavy
V4	předo - zadní míra v úrovni očí
V5	výška oči - temeno
V6	výška uši - temeno
V7	šířka čelisti

Multivarietní Q-Q plot

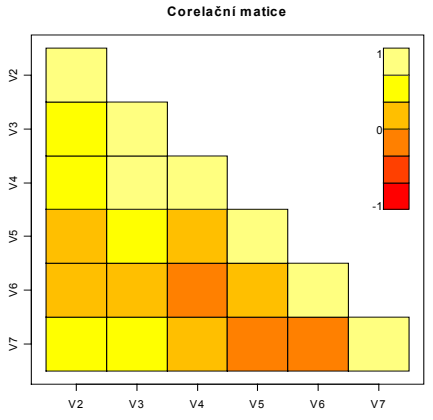


Kovarianční a korelační matice

- Za podmínky normality užitečné
- Symetrické matice p x p
- Výběrové (S, R) vs. Populační (Σ)
- Hlavní diagonála: variance (S, Σ) , jedničky (R)
- Mimo diagonální prvky: kovariance (S, Σ) , korelační koef. (R)
- $\sigma_{ij} = \sigma_{ji}$, $r_{ij} = r_{ji}$
- R má redukovanou škálu

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} & \sigma_{16} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \sigma_{24} & \sigma_{25} & \sigma_{26} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{34} & \sigma_{35} & \sigma_{36} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 & \sigma_{45} & \sigma_{46} \\ \sigma_{51} & \sigma_{52} & \sigma_{53} & \sigma_{54} & \sigma_5^2 & \sigma_{56} \\ \sigma_{61} & \sigma_{62} & \sigma_{63} & \sigma_{64} & \sigma_{65} & \sigma_6^2 \end{bmatrix} \qquad R = \begin{bmatrix} 1 & r_{12} & r_{13} & r_{14} & r_{15} & r_{16} \\ r_{21} & 1 & r_{23} & r_{24} & r_{25} & r_{26} \\ r_{31} & r_{32} & 1 & r_{34} & r_{35} & r_{36} \\ r_{41} & r_{42} & r_{43} & 1 & r_{45} & r_{46} \\ r_{51} & r_{52} & r_{53} & r_{54} & 1 & r_{56} \\ r_{61} & r_{62} & r_{63} & r_{64} & r_{65} & 1 \end{bmatrix}$$

Korelační matice



Metoda hlavních komponent

- Účel:
- Hledání lineární f-ce proměnných maximalizující celkovou varianci
- Zjednodušení struktury dat, redukce dimenze souboru (počtu proměnných)
- Výběr žádaných (podobných, nepodobných) proměnných
- Studium struktury disperse MV souboru nebo lineárních závislostí
- Regrese hlavních komponent (řešení kolinearity v matici regresorů)

Rozklad na vlastní čísla

- Většinou rozkládáme S, R nebo distanční matici (D)
- Vždy symetrická čtvercová matice A = A^T
- Definujeme diagonální matici $\Lambda_{(p \times p)}$ a matici korespondujících vlastních vektorů $E_{(p \times p)}$.

$$A = E \Lambda E^T$$

- Platí: $$\left| Ay - \Lambda y \right| = 0$$

- Matice Λ obsahuje p vlastních čísel λ_i uspořádaných sestupně
- Matice E obsahuje p sloupců vlastních vektorů e_i, kde každý sloupec přináleží jednomu vlastnímu číslu

Vlastnosti vlastních čísel

- Součin vlastních čísel: $\prod_{i=1}^{p} \lambda_i = |A|$
- Součet vlastních čísel: $\sum_{i=1}^{p} \lambda_{iR} = tr(R) = p$

  $\sum_{i=1}^{p} \lambda_{iS} = tr(S) = \sum_{i=1}^{p} s_i^2$
- Vlastní čísla λ_i ve vztahu $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$
- Počet nulových λ_i udává počet lineárních závislostí v y, singularitu E
- Podíl λ_i k součtu všech vlastních čísel udává procento celkové variance vysvětlené λ_i
- Podíly variance lze kumulovat

Vlastnosti vlastních vektorů

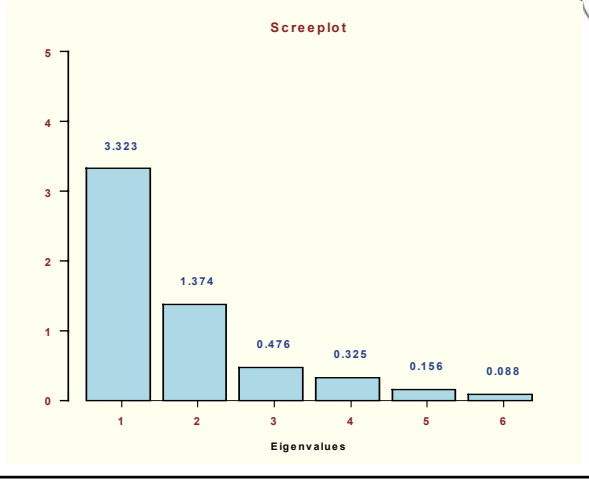
- Mají jednotkovou délku $e_i^T e_i = \sum e_i^2 = 1, \forall_i$
- Jsou vzájemně ortogonální $e_i^T e_j = \sum e_i e_j = 0, \forall i \neq j$
- Matice E je pak ortonormální $E^T = E^{-1} \quad EE^T = I$
- Hodnoty vlastního vektoru e_i vyjadřují míru participace korespondující proměnné na varianci (závislosti)

Tabulkové vyjádření rozkladu

č	Lambda	% Variance	Kumul %
1	3.323	57.871	57.871
2	1.374	23.931	81.802
3	0.476	8.290	90.091
4	0.325	5.654	95.745
5	0.156	2.725	98.470
6	0.088	1.530	100.000

Suma λ_i = 5.997
Součin λ_i = 0.001994

Grafické vyjádření rozkladu



První dva vlastní vektory

- Variance v matici S:

V2	V3	V4	V5	V6	V7
0.0426	0.8088	0.1002	0.3459	0.0380	0.0324

- Vlastní vektory (1. a 2.):

	E1	E2
V2	-0.2074	0.142
V3	-0.8728	0.219
V4	-0.2613	0.231
V5	-0.3259	-0.891
V6	-0.0656	-0.222
V7	-0.1279	0.187

Hlavní komponenty

- Princip: Výpočet nových proměnných (hlavních komponent), které zachovávají varianci, ale eliminují kovariance.
- Výpočet:

$$Z = YE$$

$$pc_1 = z_1 = e_1^T y = e_{11} y_1 + e_{12} y_2 + \ldots + e_{1p} y_p$$

$$pc_2 = z_2 = e_2^T y = e_{21} y_1 + e_{22} y_2 + \ldots + e_{2p} y_p$$

$$\ldots$$

$$pc_p = z_p = e_p^T y = e_{p1} y_1 + e_{p2} y_2 + \ldots + e_{pp} y_p$$

- Variance PC:

$$Var(pc) = \begin{bmatrix} \lambda_1 & 0 & . & 0 \\ 0 & \lambda_2 & . & 0 \\ . & . & . & 0 \\ 0 & 0 & 0 & \lambda_p \end{bmatrix}$$

Hlavní komponenty

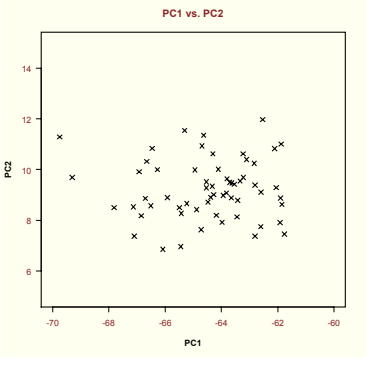
- Hlavní komponenty pc_i jsou vzájemně ortogonální
- Variance pc_i are jsou maximální pro i=1, ale postupně klesají
- Hlavní komponenty nulových λ_i jsou téměř konstantní
- Nulové λ_i důležité pro detekci lineárních závislostí

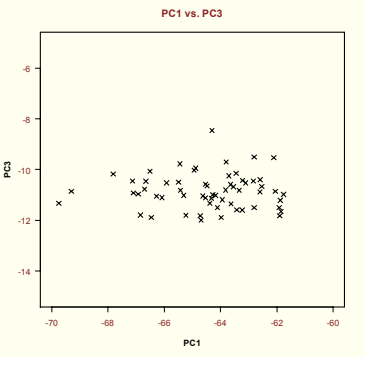
- Poslední vlastní vektor:

V2	V3	V4	V5	V6	V7
0.731	-0.238	0.358	0.113	-0.235	-0.460

- Na „téměř“ lineární závislosti se podílí především V2 a V7

Grafy hlavních komponent





Počet vybraných vlastních čísel

- Kritéria:
  - Vlastní čísla vysvětlující nejméně 80–90 % variance
  - Visuální posouzení grafu úpatí Λ
  - Nadprůměrné λ_i , $\lambda_i > 1.0$
  - Asymptotický věrohodnostní test
  - Metoda “broken stick“ (Jackson, 1993)
  - Počet zvolený podle nejvyššího počtu metod

Poznámky

- PCA je vztažena ke škále proměnných (rozdílná pro S a R)
- Multivarietní normalita výhodou
- Výstupy PCA ovlivněny extrémy v datech
- Koeficienty PC regrese jsou vychýlené, obtížně interpretovatelné

As far as the laws of mathematics refer to reality, they are not certain; as far as they are certain, they do not refer to reality.

Albert Einstein