

# Analýza hlavních komponent a faktorová analýza

Mirko Navara  
Centrum strojového vnímání  
katedra kybernetiky FEL ČVUT  
Karlovo náměstí, budova G, místnost 104a  
<http://cmp.felk.cvut.cz/~navara>

5. listopadu 2012

**Předpoklad:** Náhodný vektor  $(X_1, \dots, X_k)$  má  $k$ -rozměrné normální rozdělení s vektorem středních hodnot  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_k)$  a kovarianční maticí  $\boldsymbol{\Sigma}_X$ .

**Vstup:** Náhodný výběr

$$((x_{11}, \dots, x_{1k}), \dots, (x_{n1}, \dots, x_{nk}))$$

rozsahu  $n \geq k$ , který vyjádříme maticí

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}.$$

hodnosti  $k$ .

Max. věrohodný odhad (za předpokladu normálního rozdělení chyb) = odhad metodou nejmenších čtverců = parametry empirického rozdělení  $\text{Emp}(X_1, \dots, X_k)$ :

$$\begin{aligned} \hat{\boldsymbol{\mu}} &= \bar{\mathbf{x}} = (\bar{x}_1, \dots, \bar{x}_k), \\ \mathbf{S}_X &:= \hat{\boldsymbol{\Sigma}}_X = \text{cov}(\text{Emp}(X_1, \dots, X_k)) = \frac{1}{n} \mathbf{X}^T \mathbf{X}, \\ s_{im} &:= (\mathbf{S}_X)_{im} = \frac{1}{n} \sum_{j=1}^n (x_{ji} - \bar{x}_i)(x_{jm} - \bar{x}_m). \end{aligned}$$

**Předpoklad:**  $X_1, \dots, X_k$  jsou normované.

**Pozn.:** Někdy se tento předpoklad vynechává, ale musí se jednat o souměřitelné veličiny (nikoli měřené v různých jednotkách). Lineární transformace měřítka ovlivňuje výsledek. Předpokládáme alespoň, že jsou centrované, tj. s nulovými středními hodnotami, jinak by se zkomplikovaly vzorce.

**Důsledek:**  $\mathbf{S}_X$  = kovarianční matice = korelační matice,  $\forall i: s_{ii} = 1$ .

Sloupce matice  $\mathbf{X}$  generují lineární podprostor  $\mathcal{L} \subseteq \mathbb{R}^n$ .

**Předpoklad:**  $\dim \mathcal{L} = k$ . (Pokud ne, můžeme nepotřebné sloupce vypustit.)

V  $\mathcal{L}$  lze zvolit jinou bázi, k níž přejdeme lineární transformací  $\mathbf{T} \in \mathbb{R}^{k \times k}$ ,  $\mathbf{T}$  regulární.

Vektory nové báze = sloupce matice  $\mathbf{Z} := \mathbf{X} \mathbf{T}$ .

**Předpoklad (BÚNO):**  $\mathbf{T}$  je ortonormální.

**Důsledek:**  $\mathbf{T}^{-1} = \mathbf{T}^T$ .

Korelační matice nových bázových vektorů je

$$\frac{1}{n} \mathbf{Z}^T \mathbf{Z} = \frac{1}{n} (\mathbf{X} \mathbf{T})^T \mathbf{X} \mathbf{T} = \frac{1}{n} \mathbf{T}^T \underbrace{\mathbf{X}^T \mathbf{X}}_{n \mathbf{S}_X} \mathbf{T} = \mathbf{T}^T \mathbf{S}_X \mathbf{T}.$$

$\mathbf{S}_X$  je symetrická pozitivně definitní  $\Rightarrow$  její vlastní čísla jsou reálná kladná.

**Předpoklad:**  $\mathbf{S}_X$  má navzájem různá vlastní čísla.

**Důsledek:**  $T$  lze zvolit tak, že

$$D := T^T S_X T = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 & 0 \\ 0 & \lambda_2 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \lambda_{k-1} & 0 \\ 0 & 0 & \dots & 0 & \lambda_k \end{pmatrix}$$

je diagonální, její diagonální prvky jsou vlastní čísla matice  $S_X$  řazená sestupně,  $\lambda_1 > \lambda_2 > \dots > \lambda_k$ .

**Trik:**  $\forall i$ : za  $i$ -tý sloupec matice  $T$  volíme jednotkový vlastní vektor odpovídající vlastnímu číslu  $\lambda_i$ .

Řešení je jediné, až na znaménka (orientaci) sloupců matice  $T$ .

**Pozn.:** Nejsou-li vlastní čísla matice  $S_X$  navzájem různá, řešení není jednoznačné.

**Důsledek:**

$$S_X = T D T^T = T D^{1/2} D^{1/2} T^T = \underbrace{(T D^{1/2})}_{V} \underbrace{(T D^{1/2})^T}_{V^T},$$

kde

$$D^{1/2} = \begin{pmatrix} \sqrt{\lambda_1} & 0 & \dots & 0 & 0 \\ 0 & \sqrt{\lambda_2} & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \sqrt{\lambda_{k-1}} & 0 \\ 0 & 0 & \dots & 0 & \sqrt{\lambda_k} \end{pmatrix},$$

$$D^{1/2} D^{1/2} = D.$$

**Pozn.:** Dostali jsme vyjádření

$$S_X = V V^T,$$

což je možné právě pro symetrické pozitivně semidefinitní matice. Pokrok oproti

$$S_X = \frac{1}{n} X^T X = \left( \frac{1}{\sqrt{n}} X \right)^T \left( \frac{1}{\sqrt{n}} X \right)$$

je v tom, že  $V$  je čtvercová.

## 1 Analýza hlavních komponent

**(Principal Component Analysis, PCA)**

Náhodný vektor  $(X_1, \dots, X_k)$  transformujeme lineární transformací na náhodný vektor  $(Z_1, \dots, Z_k) = (X_1, \dots, X_k) T$ , který popisuje stejné náhodné proměnné v jiné bázi.

Jeho výběrová korelační matice  $D$  je diagonální.

Z výběrového souboru (=trénovací množiny) se zdá, že

1.  $Z_1, \dots, Z_k$  jsou nekorelované,
2.  $Z_1$  popisuje největší část rozptylu, kterou lze popsat jednou proměnnou,  
 $Z_2$  popisuje největší možnou část zbývajících rozptylu  
 atd.

Pokud vezmeme jen  $m < k$  prvních proměnných  $Z_1, \dots, Z_m$ , odpovídajících největším vlastním číslům korelační matice  $S_X$ , dostaneme popis, který v jistém smyslu optimálně aproximuje původní data, ale má menší dimenzi  $m$ , takže se s ním lépe pracuje.

K tomu nám stačí, aby prvních  $m$  vlastních čísel korelační matice  $S_X$  bylo navzájem různých.

$Z_i$  je  **$i$ -tá hlavní komponenta**, popisuje část rozptylu úměrnou

$$\frac{\lambda_i}{\sum_{p=1}^k \lambda_p} = \frac{\lambda_i}{k},$$

kde jmenovatel

$$\sum_{p=1}^k \lambda_k = \text{tr } \mathbf{D} = \text{tr } \mathbf{S}_{\mathbf{X}} = \sum_{p=1}^k 1 = k.$$

Jedná se o stopu matice  $\mathbf{D}$  (=součet diagonálních prvků), která se lineární transformací souřadnic nemění, a stopa původní matice  $\mathbf{S}_{\mathbf{X}}$  je  $k$ , neboť korelační matice typu  $k \times k$  má na diagonále jedničky.

## 1.1 Příklad aplikace

Fotografie obličejů lze uspokojivě popsat několika desítkami souřadnic (hlavních komponent), zatímco původní snímky mají dimenzi rovnou počtu pixelů, ale jsou velmi korelované.

Z nich lze i rekonstruovat původní fotografii (známe-li hlavní komponenty) zpětnou transformací

$$(X_1, \dots, X_k) = (Z_1, \dots, Z_k) \mathbf{T}^T,$$

resp.

$$(X_1, \dots, X_k) \approx (Z_1, \dots, Z_m, 0, \dots, 0) \mathbf{T}^T,$$

pokud jsme provedli projekci na  $m$  hlavních komponent.

## 2 Faktorová analýza

Snaží se vysvětlit závislost náhodných veličin  $X_1, \dots, X_k$  pomocí lineární závislosti na jiných náhodných veličinách  $F_1, \dots, F_m$ ,  $m \leq k$ , zvaných **faktory**.

**Model:**

$$\begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix} = \mathbf{V} \begin{pmatrix} F_1 \\ \vdots \\ F_m \end{pmatrix} + \begin{pmatrix} \mathcal{E}_1 \\ \vdots \\ \mathcal{E}_k \end{pmatrix},$$

kde  $\mathbf{V} \in \mathbb{R}^{k \times m}$  je neznámá matice,

$\mathcal{E}_1, \dots, \mathcal{E}_k$  jsou náhodné veličiny (chyby, šum).

**Předpoklady:**  $X_1, \dots, X_k$  jsou normované,

$F_1, \dots, F_m$  jsou ortogonální a **normované**, tedy ortonormální,

$\mathcal{E}_1, \dots, \mathcal{E}_k$  jsou nezávislé navzájem i na faktorech a centrované (=s nulovou stř. hodnotou).

Značení:

$$\mathbf{X} := \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix}, \quad \mathbf{F} := \begin{pmatrix} F_1 \\ \vdots \\ F_m \end{pmatrix}, \quad \mathcal{E} := \begin{pmatrix} \mathcal{E}_1 \\ \vdots \\ \mathcal{E}_k \end{pmatrix},$$

$$\mathbf{X} = \mathbf{V} \mathbf{F} + \mathcal{E},$$

kovarianční (=korelační) matice

$$\begin{aligned} \Sigma_{\mathbf{X}} &= \text{E}(\mathbf{X} \mathbf{X}^T) = \text{E}((\mathbf{V} \mathbf{F} + \mathcal{E})(\mathbf{V} \mathbf{F} + \mathcal{E})^T) = \text{E}(\mathbf{V} \mathbf{F} \mathbf{F}^T \mathbf{V}^T) + \text{E}(\mathcal{E} \mathcal{E}^T) = \\ &= \mathbf{V} \underbrace{\text{E}(\mathbf{F} \mathbf{F}^T)}_{\mathbf{I}_k} \mathbf{V}^T + \Sigma_{\mathcal{E}} = \mathbf{V} \mathbf{V}^T + \Sigma_{\mathcal{E}}, \end{aligned}$$

kde  $\Sigma_{\mathcal{E}}$  je (diagonální) kovarianční matice vektoru chyb. Tu pro výpočet nahradíme výběrovou korelační maticí  $\mathbf{S}_{\mathbf{X}}$  vypočtenou z náhodného výběru.

Pokud  $m = k$ , žádné chyby připouštět nemusíme, protože existuje rozklad  $\mathbf{S}_{\mathbf{X}} = \mathbf{V} \mathbf{V}^T$ . (Našli jsme ho při analýze hlavních komponent, jediný rozdíl je, že zde je měřítko upraveno tak, že normovaná není transformace, ale náhodné veličiny  $F_1, \dots, F_k$ .)

Pokud  $m < k$ , hledáme rozklad

$$\mathbf{S}_{\mathbf{X}} = \mathbf{V} \mathbf{V}^T + \Sigma_{\mathcal{E}},$$

kde  $\mathbf{V} \in \mathbb{R}^{k \times m}$  je hodnosti  $m < k$  a  $\Sigma_{\mathcal{E}}$  je diagonální.

**Problémy:**

1. Obtížná úloha vyžadující iterační postupy.
2. Nejednoznačnost faktorů (stejný podprostor má různé ortonormální báze – **rotace faktorů**).
3. Volba vhodné dimenze  $m$ .