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 $\boldsymbol{\Sigma}_{\boldsymbol{X}}$.
Vstup: Náhodný výběr

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

rozsahu $n \geq k$, který vyjádříme maticí

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

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

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}} = (\bar{x}_1, \dots, \bar{x}_k) ,$$

$$\boldsymbol{S}_{\boldsymbol{X}} := \hat{\boldsymbol{\Sigma}}_{\boldsymbol{X}} = \mathrm{cov}(\mathrm{Emp}(X_1, \dots, X_k)) = \frac{1}{n} \boldsymbol{X}^T \boldsymbol{X} ,$$

$$s_{im} := (\boldsymbol{S}_{\boldsymbol{X}})_{im} = \frac{1}{n} \sum_{j=1}^{n} (x_{ji} - \bar{x}_i)\,(x_{jm} - \bar{x}_m) .$$

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: $\boldsymbol{S}_{\boldsymbol{X}}$ = kovarianční matice = korelační matice, $\forall i : s_{ii} = 1$.
Sloupce matice $\boldsymbol{X}$ 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 $\boldsymbol{T} \in \mathbb{R}^{k\times k}$, $\boldsymbol{T}$ regulární.
Vektory nové báze = sloupce matice $\boldsymbol{Z} := \boldsymbol{X}\,\boldsymbol{T}$.
Předpoklad (BÚNO)**:** $\boldsymbol{T}$ je ortonormální.
Důsledek: $\boldsymbol{T}^{-1} = \boldsymbol{T}^T$.
Korelační matice nových bázových vektorů je

$$\frac{1}{n}\,\boldsymbol{Z}^T\,\boldsymbol{Z} = \frac{1}{n}\,(\boldsymbol{X}\,\boldsymbol{T})^T\,\boldsymbol{X}\,\boldsymbol{T} = \frac{1}{n}\,\boldsymbol{T}^T\,\underbrace{\boldsymbol{X}^T\,\boldsymbol{X}}_{n\,\boldsymbol{S_X}}\,\boldsymbol{T} = \boldsymbol{T}^T\,\boldsymbol{S_X}\,\boldsymbol{T}\,.$$

$\boldsymbol{S_X}$ je symetrická pozitivně definitní $\Rightarrow$ její vlastní čísla jsou reálná kladná.
Předpoklad: $\boldsymbol{S_X}$ má navzájem různá vlastní čísla.
Důsledek: $\boldsymbol{T}$ lze zvolit tak, že

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

je diagonální, její diagonální prvky jsou vlastní čísla matice $\boldsymbol{S_X}$ řazená sestupně, $\lambda_1 > \lambda_2 > \dots > \lambda_k$.
Trik: $\forall i$: za i-tý sloupec matice $\boldsymbol{T}$ volíme jednotkový vlastní vektor odpovídající vlastnímu číslu λ_i.
Řešení je jediné, až na znaménka (orientaci) sloupců matice $\boldsymbol{T}$.
Pozn.: Nejsou-li vlastní čísla matice $\boldsymbol{S_X}$ navzájem různá, řešení není jednoznačné.
Důsledek:

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

kde

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

$$\boldsymbol{D}^{1/2}\,\boldsymbol{D}^{1/2} = \boldsymbol{D}\,.$$

Pozn.: Dostali jsme vyjádření

$$\boldsymbol{S_X} = \boldsymbol{V}\,\boldsymbol{V}^T\,,$$

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

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

je v tom, že $\boldsymbol{V}$ je čtvercová.

1 Analýza hlavních komponent

(**Principal Component Analysis**, **PCA**)
Náhodný vektor $(X_1, \dots, X_k)$ transformujeme lineární transformací na náhodný vektor $(Z_1, \dots, Z_k) = (X_1, \dots, X_k)\,\boldsymbol{T}$, který popisuje stejné náhodné proměnné v jiné bázi.
Jeho výběrová korelační matice $\boldsymbol{D}$ 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 zbývajícího rozptylu
   atd.

Pokud vezmeme jen $m < k$ prvních proměnných $Z_1, \dots, Z_m$, odpovídajících největším vlastním číslům korelační matice $\boldsymbol{S_X}$, 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 $\boldsymbol{S_X}$ bylo navzájem různých.
Z_i je ***i*-tá hlavní komponenta**, popisuje část rozptylu úměrnou

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

kde jmenovatel

$$\sum_{p=1}^{k} \lambda_k = \operatorname{tr} \boldsymbol{D} = \operatorname{tr} \boldsymbol{S_X} = \sum_{p=1}^{k} 1 = k\,.$$

Jedná se o stopu matice $\boldsymbol{D}$ (=součet diagonálních prvků), která se lineární transformací souřadnic nemění, a stopa původní matice $\boldsymbol{S_X}$ 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

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

resp.

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

pokud jsme provedli projekci na m hlavních komponent.

2 Faktorová analýza

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

Model:

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

kde $\boldsymbol{V} \in \mathbb{R}^{k \times m}$ je neznámá matice,
$\mathcal{E}_1, \ldots, \mathcal{E}_k$ jsou náhodné veličiny (chyby, šum).
Předpoklady: $X_1, \ldots, X_k$ jsou normované,
$F_1, \ldots, F_m$ jsou ortogonální a **normované**, tedy ortonormální,
$\mathcal{E}_1, \ldots, \mathcal{E}_k$ jsou nezávislé navzájem i na faktorech a centrované (=s nulovou stř. hodnotou).
Značení:

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

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

kovarianční (=korelační) matice

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

kde $\boldsymbol{\Sigma}_{\mathcal{E}}$ je (diagonální) kovarianční matice vektoru chyb. Tu pro výpočet nahradíme výběrovou korelační maticí $\boldsymbol{S}_{\boldsymbol{X}}$ vypočtenou z náhodného výběru.
Pokud $m = k$, žádné chyby připouštět nemusíme, protože existuje rozklad $\boldsymbol{S}_{\boldsymbol{X}} = \boldsymbol{V}\,\boldsymbol{V}^T$. (Našli jsme ho při analýze hlavních komponent, jediný rozdíl je, že zde je měřítko upraveno tak, že normovaná není transformace, ale náhodné veličiny $F_1, \ldots, F_k$.
Pokud $m < k$, hledáme rozklad

$$\boldsymbol{S}_{\boldsymbol{X}} = \boldsymbol{V}\,\boldsymbol{V}^T + \boldsymbol{\Sigma}_{\mathcal{E}} ,$$

kde $\boldsymbol{V} \in \mathbb{R}^{k \times m}$ je hodnosti $m < k$ a $\boldsymbol{\Sigma}_{\mathcal{E}}$ 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.