Přednáška 7: Bilineární model a bayesovský přístup k maticové dekompozici (odhad počtu komponent?)

Ondřej Tichý

Bayesovské metody v machine learningu (BML)

April 8, 2018

Obsah přednášky

- bilineární model jako maticová dekompozice
- pár motivačních příkladů využití
- klasické možnosti řešení dekompozice
- bayesovská formulace maticové dekompozice
  - odhad počtu komponent
  - meze neurčitosti odhadů
- další předpoklady

Bilineární model

Co to je?

- linearita v 1D:

$$\begin{array}{cc} \text{ano} & f(x) = 4x + 7 \\ \text{ne} & f(x) = 4x^2 + 7 \\ \text{ne} & f(x) = \sin x + 7 \end{array}$$

- bilinearita ve 2D (lineární v každé proměnné):

$$\begin{array}{cc} \text{ano} & f(x,y) = 4xy + 7y + 3 \\ \text{ano} & f(x,y) = 3x + 5y + 9 \\ \text{ne} & f(x,y) = xy^2 + x + 7 \end{array}$$

Bilineární model

Maticová formulace

- v maticové formulaci nás bude zajímat bilineární model

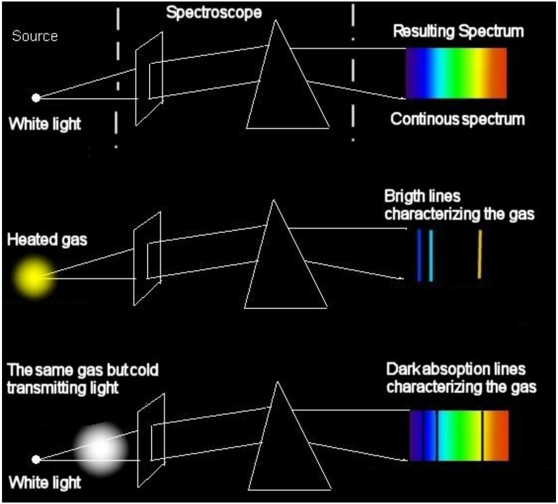
$$D = AX^T, \tag{1}$$

  kde $D \in \mathbf{R}^{p \times n}$, $A \in \mathbf{R}^{p \times r}$ a $X \in \mathbf{R}^{n \times r}$.
- r je vnitřní rozměr maticového součinu

Motivace

K čemu je to dobré?

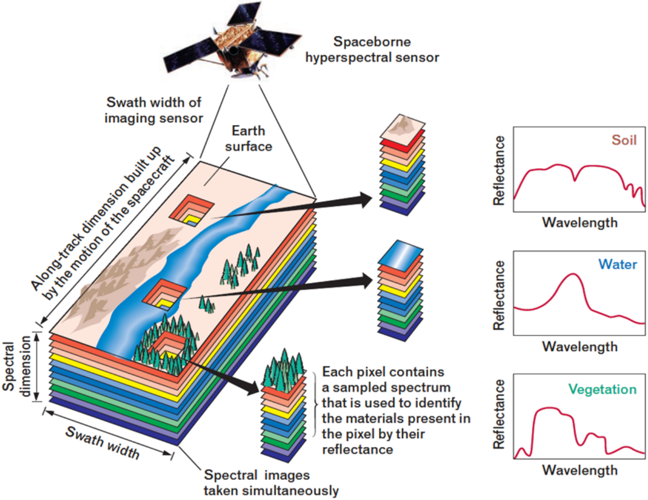
- astronomie - spektroskopie



Motivace

K čemu je to dobré?

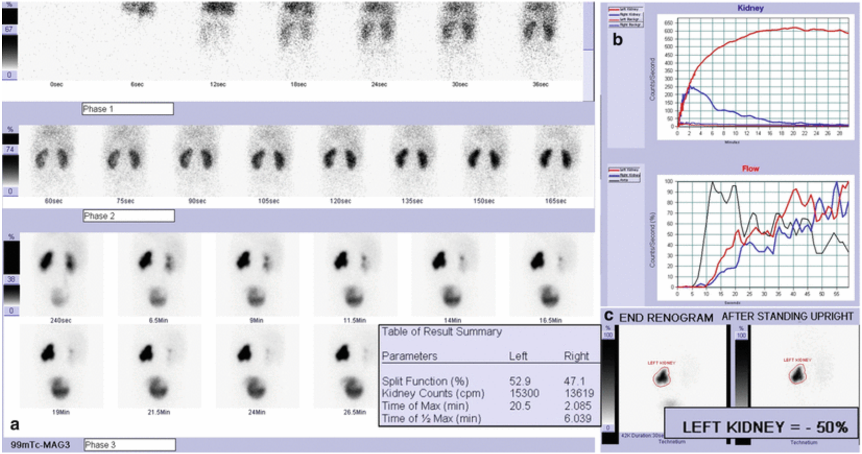
- hyperspektrální snímkování



Motivace

K čemu je to dobré?

- nukleární medicína



- zpracování textu, zpracování řeči, komprese, a další

Formulace modelu

Reprezentace “obrázků”

- sloupce matice X představují váhové vektory (např. aktivitu zdroje v daném čase)
- sloupce matice A představují (např.) zdrojové obrázky uložené ve sloupcích:

$$\begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} \tag{2}$$

Formulace modelu

Maticová formulace

- každý obrázek (sloupec matic D) $\mathbf{d}_t$ je lineární kombinace jednotlivých zdrojových obrázků:

$$\mathbf{d}_t = \mathbf{a}_1 x_{1,t} + \mathbf{a}_2 x_{2,t} + \cdots + \mathbf{a}_r x_{r,t} \tag{3}$$

  - měřená data: $D = [\mathbf{d}_1, \ldots, \mathbf{d}_t, \ldots, \mathbf{d}_n]$
  - zdrojové obrázky: $A = [\mathbf{a}_1, \ldots, \mathbf{a}_r]$
  - váhy zdrojových obrázků: $X = \left[\underline{\mathbf{x}}_1^T, \ldots, \underline{\mathbf{x}}_t^T, \ldots, \underline{\mathbf{x}}_n^T\right]^T$
- maticový zápis:

$$D = AX^T. \tag{4}$$

pozn.: $D_{i,j} = \sum_{k=1}^{r} A_{i,k} X_{k,j}$

Formulace modelu

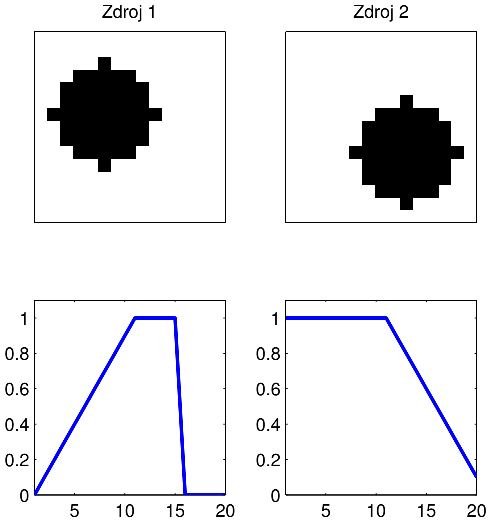
Předpoklady a neznalosti

- neznámý počet zdrojů r
- pozitivita (často)
- šum
- rotační neurčitost

$$\widehat{A}\widehat{X}^T = \underbrace{\widehat{A}R}_{\tilde{A}}\underbrace{R^{-1}\widehat{X}^T}_{\tilde{X}} \tag{5}$$

Cvičná data

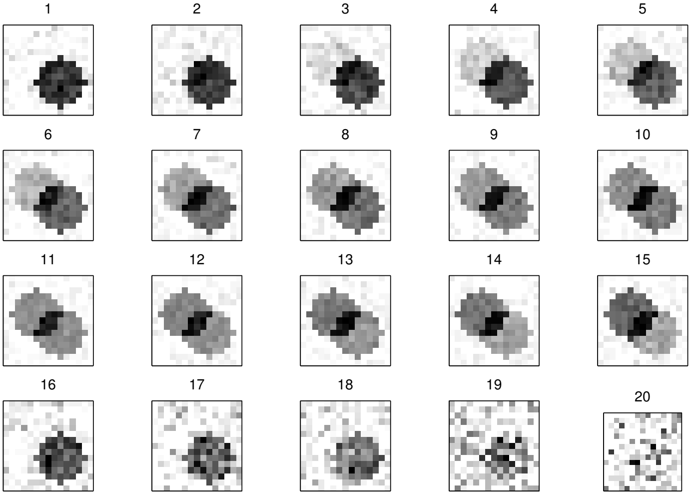
Simulované hodnoty



Cvičná data

Data

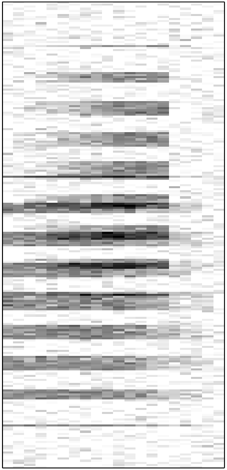
- vynásobíme AX^T a přidáme šum:



Cvičná data

Data

- matice D

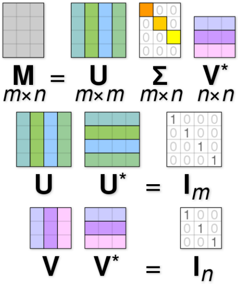


Souvislost s PCA

PCA a SVD

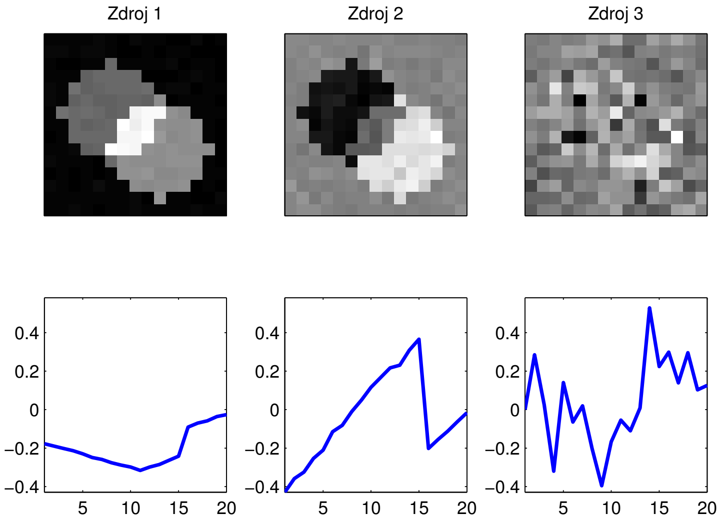
- hledáme projekci $A = DU$, kde U je matice vlastních vektorů kovarianční matice $C_D = \frac{1}{2}\sum(d_i - \overline{d})(d_i - \overline{d})^T$, tzn. $C_D = U \Lambda U^T$.
- defacto hledáme souřadnou soustavu, ve které je variance jednotlivých komponent maximalizována
- SVD (singular value decomposition) přístup:

$$D = USV^T = \underbrace{US}_{A}\underbrace{V^T}_{X^T}, \tag{6}$$



Souvislost s PCA

Cvičný příklad



Souvislost s PCA

Poznámky

- velmi rychlé
- neřekne nám “správný” počet komponent
- šum a pozitivita
- rotační neurčitost

Non-negative matrix factorization (NMF)

Ještě bez obtěžování pana Bayese...

- hledáme rozklad datové matice D jako

$$D \approx AX^T \tag{7}$$

- minimalizujeme vzdálenost levé a pravé strany rovnice

$$\min ||D - AX^T||_2^2, \tag{8}$$

  kde $||B - C||_2^2 = \sum_{i,j} (B_{i,j} - C_{i,j})^2$.
- problém je konvexní buď v A nebo X, nikoliv v obou najednou $\Rightarrow$ nelze očekávat nalezení globálního minima.
- široká škála numerických metod pro řešení (gradient descent, konjugované gradienty, atd.)
- kompromis: “multiplicative update rules”, Lee and Seung 2001 (verze, kterou si ukážeme)

Non-negative matrix factorization (NMF)

Algoritmus

- inicializace A a X
- iterujeme

$$X^T \leftarrow X^T \circ \left(\left(A^T D \right) \oslash \left(A^T A X^T \right) \right) \tag{9}$$

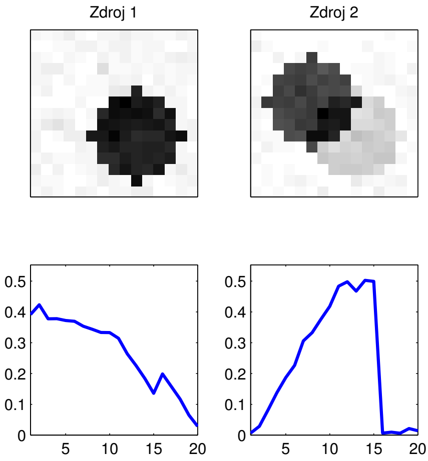
$$A \leftarrow A \circ \left((DX) \oslash \left(A X^T X \right) \right) \tag{10}$$

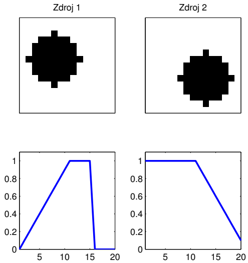
  kde $\circ$ je násobení “po prvcích” a $\oslash$ je dělení “po prvcích”.
- máme výsledné odhady $\widehat{A}$ a $\widehat{X}$

Non-negative matrix factorization (NMF)

Cvičný příklad

- start z náhodně vygenerovaných matic
- $r = 2$

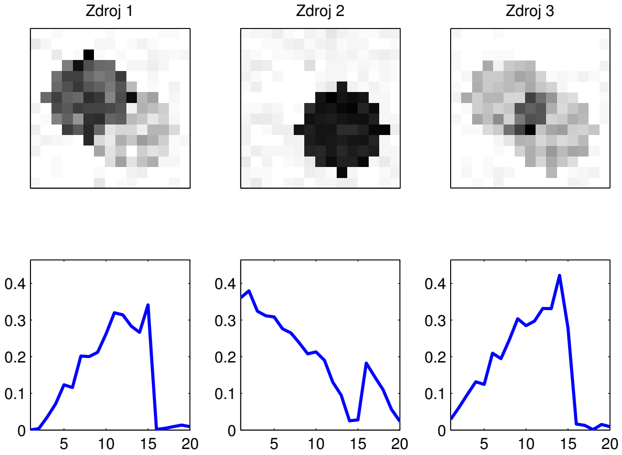




Non-negative matrix factorization (NMF)

Cvičný příklad

- start z náhodně vygenerovaných matic
- $r = 3$



Non-negative matrix factorization (NMF)

Hlavní (ne)výhody

- velice jednoduché na implementaci
- velice rychlé na výpočet
- máme pouze bodový odhad A a X
- výsledek silně závisí na počtu komponent
- stále rotační neurčitost, výsledek je nejednoznačný

Matematická vsuvka

Diagonála, stopa, Kronekerův součin

- operátor diagonály, $\mathrm{diag}()$:

$$\mathrm{diag}\left(\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}\right) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix} \tag{11}$$

$$\mathrm{diag}\left(\begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}\right) = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \tag{12}$$

- operátor stopa, součet diagonálních prvků čtvercové matice, $\mathrm{tr}()$:

$$\mathrm{tr}\left(A\right) = \sum_i A_{i,i} \tag{13}$$

- pro rozměrově kompatibilní matice platí:

$$\mathrm{tr}\left(ABC\right) = \mathrm{tr}\left(CAB\right) \tag{14}$$

- Kronekerův součin dvou matic:

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1m}B \\ a_{21}B & a_{22}B & & \vdots \\ \vdots & & \ddots & \vdots \\ a_{n1}B & \cdots & \cdots & a_{nm}B \end{pmatrix} \tag{15}$$

- maticové normální rozdělení matice $A \in \mathbf{R}^{p \times r}$:

$$\mathcal{N}_A(\mu_A, \Sigma_p \otimes \Phi_r) = \frac{1}{(2\pi)^{\frac{pr}{2}} |\Sigma_p|^{\frac{r}{2}} |\Phi_r|^{\frac{p}{2}}} \times \\ \times \exp\left(-\frac{1}{2}\mathrm{tr}\left[\Sigma_p^{-1}(A - \mu_A)(\Phi_r^{-1})^T(A - \mu_A)^T\right]\right), \quad (16)$$

- jeho momenty:

$$\widehat{A} = \mu_A, \quad (17)$$

$$\widehat{AA^T} = \mathrm{tr}(\Phi_r)\Sigma_p + \mu_A\mu_A^T, \quad (18)$$

$$\widehat{A^TA} = \mathrm{tr}(\Sigma_p)\Phi_r + \mu_A^T\mu_A. \quad (19)$$

Bayesovská formulace

Variační PCA [Bishop, 1999]

- model pozorování

$$D = AX^T + E, \tag{20}$$

kde $E_{i,j} = \mathcal{N}(0, \omega^{-1}), \ \forall i, \forall j.$

- model dat

$$f\left(D|A, X, \omega\right) = \mathcal{N}\left(AX^T, \omega^{-1} I_p \otimes I_n\right), \tag{21}$$

kde ω je (neznámá) variance šumu s apriorním modelem

$$f(\omega) = \mathcal{G}\left(\vartheta_0, \rho_0\right). \tag{22}$$

Bayesovská formulace

Variační PCA [Bishop, 1999]

- předpokládejme známé (zvolené) r
- apriorní modely pro A a X jsou voleny jako

$$f(A) = \mathcal{N}(\mathbf{0}, I_p \otimes I_r) \tag{23}$$

a

$$f(X) = \mathcal{N}(\mathbf{0}, I_n \otimes I_r)\,. \tag{24}$$

- máme model dat $f(\mathbf{y}|\theta)$,
- máme q podmíněně nezávislých proměnných, tzn. pro jejich arpiorna platí

$$f\left(\theta_1, \theta_2, \ldots, \theta_q\right) = \prod_{i=1}^{q} f(\theta_i), \tag{25}$$

- pak aposteriorní rozdělení $\tilde{f}(\theta_i|\mathbf{y})$ pro $i$tý parametr získáme VB aproximací jako

$$\tilde{f}\left(\theta_i|\mathbf{y}\right) \propto \exp\left[\mathrm{E}_{\tilde{f}(\theta_{\setminus i}|\mathbf{y})}\left(\ln f(\theta_1, \theta_2, \ldots, \theta_q, \mathbf{y})\right)\right], \ \forall i. \tag{26}$$

Variační PCA

Řešení

- odhad variance šumu: $\tilde{f}(\omega|D,r) = \mathcal{G}_\omega(\vartheta, \rho)$, kde

$$\vartheta = \vartheta_0 + \frac{np}{2}, \tag{27}$$

$$\rho = \rho_0 + \frac{1}{2}\operatorname{tr}\left(DD^T - \widehat{A}\widehat{X}^T D^T - D\widehat{X}\widehat{A}^T\right) + \frac{1}{2}\operatorname{tr}\left(\widehat{A^T A X^T X}\right) \tag{28}$$

s momentem $\widehat{\omega} = \frac{\vartheta}{\rho}$.

- ohad matice A: $\tilde{f}(A|D,r) = \mathcal{N}_A(\mu_A, I_p \otimes \Sigma_A)$, kde

$$\Sigma_A = \left(\widehat{\omega}\widehat{X^T X} + I_r\right)^{-1}, \; \mu_A = \widehat{\omega} D\widehat{X}\Sigma_A. \tag{29}$$

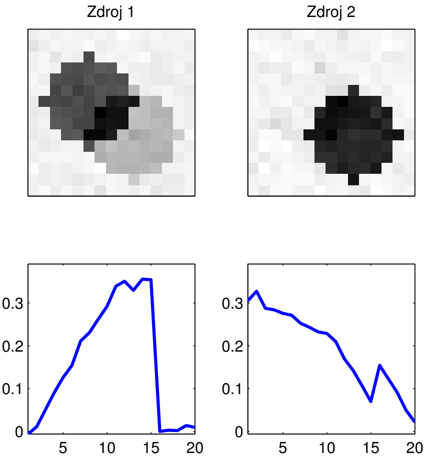
- ohad matice X: $\tilde{f}(X|D,r) = \mathcal{N}_X(\mu_X, I_n \otimes \Sigma_X)$, kde

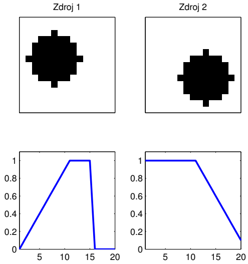
$$\Sigma_X = \left(\widehat{\omega}\widehat{A^T A} + I_r\right)^{-1}, \; \mu_X = \widehat{\omega} D^T\widehat{A}\Sigma_X. \tag{30}$$

Variační PCA

Cvičný příklad

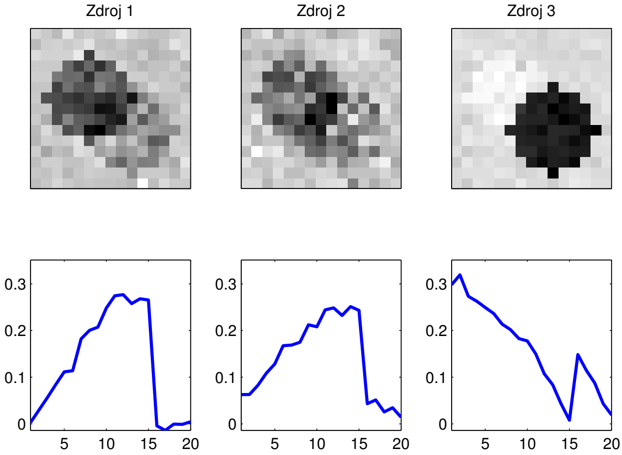
- ukážeme si případ pro $r = 2$ a $r = 3$:





Variační PCA

Cvičný příklad

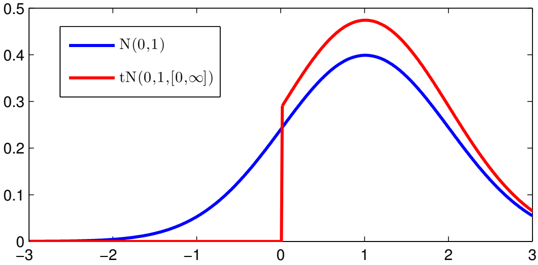


Variační PCA s pozitivitou

Pozitivita

- ořezané normální rozdělení:
  - místo $\mathcal{N}(\mu, \sigma)$ můžeme použít $t\mathcal{N}(\mu, \sigma, [0, \infty])$

$$t\mathcal{N}_x(\mu, \sigma, [0, \infty]) \propto \mathcal{N}_x(\mu, \sigma)\chi(x > 0) \tag{31}$$



  - v našem případě:

$$f(A) = t\mathcal{N}_A\left(\mathbf{0}, I_p \otimes I_r, [0, \infty]\right) \tag{32}$$

$$f(X) = t\mathcal{N}_X\left(\mathbf{0}, I_n \otimes I_r, [0, \infty]\right) \tag{33}$$

Variační PCA s pozitivitou

Cvičný příklad