Optimalizace

9. Singulární rozklad (SVD)

Tomáš Kroupa Tomáš Werner

2022 LS

Fakulta elektrotechnická
ČVUT v Praze

K čemu slouží SVD

Singulární rozklad libovolné reálné matice $\mathbf{A}$ je tvaru

$$\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathsf{T}},$$

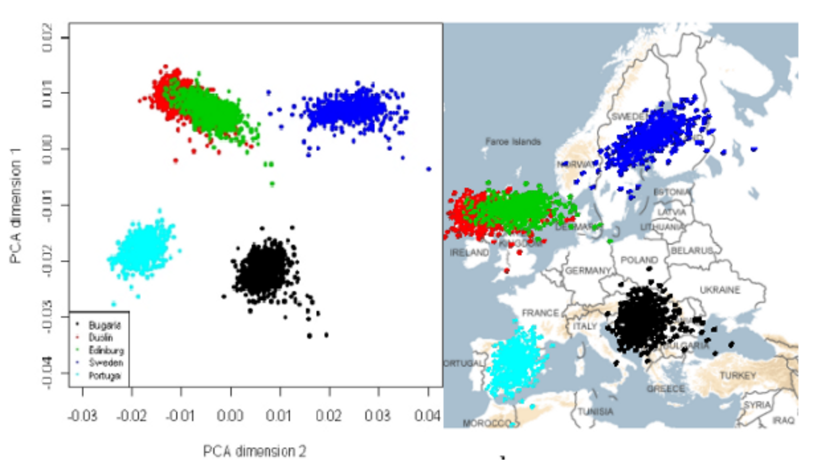
kde $\mathbf{S}$ je diagonální a $\mathbf{U}, \mathbf{V}$ jsou ortogonální matice.

Aplikace

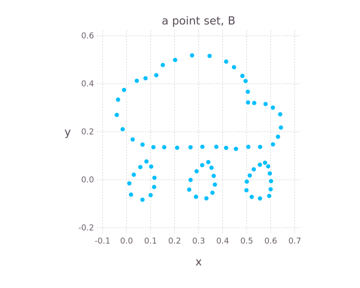
- PCA, nejbližší matice nižší hodnosti
- Rozpoznávání obličejů (eigenfaces)
- Latentní sémantická analýza
- Collaborative filtering
- Ortogonální Prokrustův problém

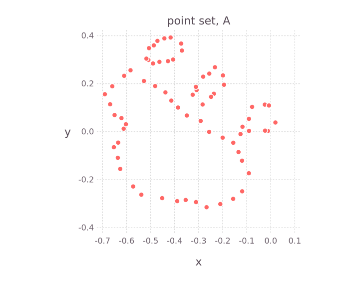
PCA

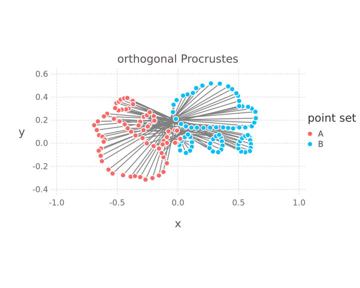
Databáze genomu ($m \approx 200\,000$ proměnných) pro $n = 1\,400$ Evropanů byla promítnuta na $k = 2$ hlavní komponenty:

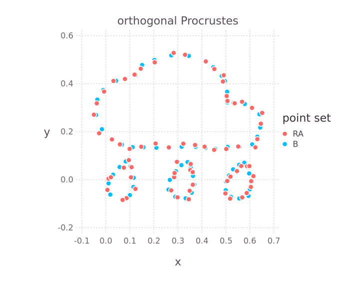


Hledáme optimální rotaci/zrcadlení









SVD teoreticky

Singulární rozklad matice

Věta (Existence SVD)

Nechť $\mathbf{A} \in \mathbb{R}^{m \times n}$ a $p := \min\{m, n\}$. Platí

$$\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathsf{T}},$$

kde diagonální matice $\mathbf{S} \in \mathbb{R}^{m \times n}$ má na diagonále singulární čísla $s_1, \ldots, s_p \geq 0$ a ortogonální matice

$$\mathbf{U} = \begin{bmatrix} \mathbf{u}_1 \cdots \mathbf{u}_m \end{bmatrix} \in \mathbb{R}^{m \times m}, \qquad \mathbf{V} = \begin{bmatrix} \mathbf{v}_1 \cdots \mathbf{v}_n \end{bmatrix} \in \mathbb{R}^{n \times n}$$

mají ve sloupcích levé/pravé singulární vektory.

Singulární čísla řadíme *sestupně*: $s_1 \geq \cdots \geq s_p$

SVD symetrické čtvercové matice A

Spektrální rozklad

$$\mathbf{A} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathsf{T}},$$

kde $\boldsymbol{\Lambda}$ má na diagonále vlastní čísla seřazena sestupně v absolutní hodnotě, $|\lambda_1| \geq \cdots \geq |\lambda_n|$.

- Když je $\mathbf{A}$ pozitivně semidefinitní, pak $\mathbf{A} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathsf{T}}$ je SVD a singulární čísla jsou vlastní čísla
- Pokud není, pak dostaneme SVD $\mathbf{A} = \mathbf{V}'\mathbf{S}\mathbf{V}^{\mathsf{T}}$ tak, že položíme $\mathbf{S} = |\boldsymbol{\Lambda}|$ a $\mathbf{V}'$ vznikne z $\mathbf{V}$ násobením -1 sloupců odpovídajích záporným vlastním číslům

Jak souvisí singulární čísla s vlastními čísly?

Spektrální rozklad matic $\mathbf{A}\mathbf{A}^\top$ a $\mathbf{A}^\top\mathbf{A}$ pomocí SVD

$$\mathbf{A}\mathbf{A}^\top = \mathbf{U}\mathbf{S}\mathbf{V}^\top\mathbf{V}\mathbf{S}^\top\mathbf{U}^\top = \mathbf{U}\mathbf{S}\mathbf{S}^\top\mathbf{U}^\top$$
$$\mathbf{A}^\top\mathbf{A} = \mathbf{V}\mathbf{S}^\top\mathbf{U}^\top\mathbf{U}\mathbf{S}\mathbf{V}^\top = \mathbf{V}\mathbf{S}^\top\mathbf{S}\mathbf{V}^\top$$

- Matice $\mathbf{A}\mathbf{A}^\top$ a $\mathbf{A}^\top\mathbf{A}$ mají stejná nenulová vlastní čísla
- Nenulová singulární čísla s_i matice $\mathbf{A}$ jsou odmocniny nenulových vlastních čísel λ_i matice $\mathbf{A}\mathbf{A}^\top$, neboť $\lambda_i = s_i^2$
- Levé singulární vektory $\mathbf{u}_i$ jsou vlastní vektory matice $\mathbf{A}\mathbf{A}^\top$
- Pravé singulární vektory $\mathbf{v}_i$ jsou vlastní vektory matice $\mathbf{A}^\top\mathbf{A}$

Jak počítat SVD?

Případ $m \gg n$

V této situaci je matice $\mathbf{A}^\top\mathbf{A} \in \mathbb{R}^{n\times n}$ malá a singulární vektory lze spočítat z jejího spektrálního rozkladu:

1. Platí $s_i = \sqrt{\lambda_i}$, kde $\lambda_i > 0$ jsou vlastní čísla matice $\mathbf{A}^\top\mathbf{A}$
2. Pravé singulární vektory $\mathbf{v}_i$ jsou vlastní vektory matice $\mathbf{A}^\top\mathbf{A}$
3. Levé singulární vektory dopočteme jako $\mathbf{u}_i = \frac{\mathbf{A}\mathbf{v}_i}{s_i}$

Existují speciální *SVD algoritmy*, které se vyhýbají explicitnímu výpočtu matic $\mathbf{A}\mathbf{A}^\top$ a $\mathbf{A}^\top\mathbf{A}$.

Různé verze SVD

$$\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^\mathsf{T} = s_1\mathbf{u}_1\mathbf{v}_1^\mathsf{T} + \cdots + s_p\mathbf{u}_p\mathbf{v}_p^\mathsf{T}$$

Plné SVD

$\mathbf{U} \in \mathbb{R}^{m\times m}$, $\mathbf{S} \in \mathbb{R}^{m\times n}$, $\mathbf{V} \in \mathbb{R}^{n\times n}$

Redukované SVD

$\mathbf{U} \in \mathbb{R}^{m\times p}$, $\mathbf{S} \in \mathbb{R}^{p\times p}$, $\mathbf{V} \in \mathbb{R}^{n\times p}$

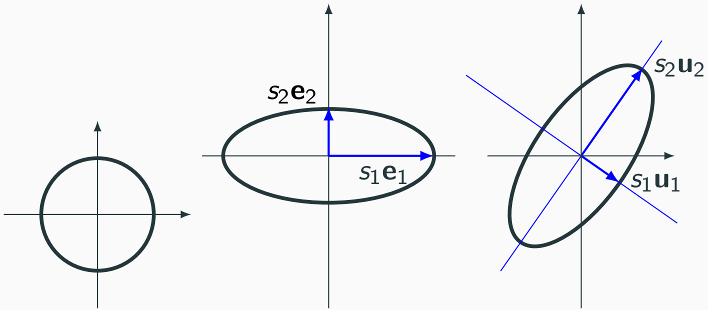
Rank-minimální SVD

$\mathbf{U} \in \mathbb{R}^{m\times r}$, $\mathbf{S} \in \mathbb{R}^{r\times r}$, $\mathbf{V} \in \mathbb{R}^{n\times r}$, kde $r := \operatorname{rank}\mathbf{S} \leq p$

Tedy r je počet nenulových singulárních čísel a $r = \operatorname{rank}\mathbf{A}$.

Lineární transformace s regulární maticí A jednotkové kružnice

$$\mathbf{A} = \mathbf{USV}^\mathsf{T} = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 \end{bmatrix} \begin{bmatrix} s_1 & 0 \\ 0 & s_2 \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^\mathsf{T} \\ \mathbf{v}_2^\mathsf{T} \end{bmatrix} = s_1 \mathbf{u}_1 \mathbf{v}_1^\mathsf{T} + s_2 \mathbf{u}_2 \mathbf{v}_2^\mathsf{T}$$



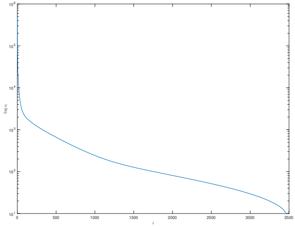
Obrazem jednotkové kružnice je natočená elipsa s délkami os s_1,s_2.

Singulární čísla některých matic

- Ortogonální matice má všechna singulární čísla rovna 1
- Hilbertova matice $\mathbf{A} = \left[\frac{1}{i+j-1}\right]_{ij}$ řádu n je regulární a např. pro $n = 100$ platí

$$s_1 = 2.1827, \ldots, s_{100} \approx 10^{-17}$$

- Černobílý obrázek psa je matice 3456×4608 s plnou řádkovou hodností, její singulární čísla jsou v grafu s $\log_{10}$ stupnicí



SVD v optimalizaci

Nejbližší matice nižší hodnosti z redukovaného SVD

$$\mathbf{A} \in \mathbb{R}^{m\times n}, \;\; \mathbf{A} = \mathbf{USV}^\mathsf{T}, \;\; s_1 \geq \cdots \geq s_p, \;\; p = \min\{m, n\}$$

Věta (Eckart-Young)

Nechť $k \leq p$. Řešením úlohy

$$\min\left\{\|\mathbf{A} - \mathbf{B}\|^2 \mid \mathbf{B} \in \mathbb{R}^{m\times n}, \operatorname{rank} \mathbf{B} \leq k\right\}$$

je matice

$$\mathbf{B}^* = \mathbf{US}_k\mathbf{V}^\mathsf{T} = s_1\mathbf{u}_1\mathbf{v}_1^\mathsf{T} + \cdots + s_k\mathbf{u}_k\mathbf{v}_k^\mathsf{T},$$

kde $\mathbf{S}_k$ vznikne z $\mathbf{S}$ vynulováním singulárních čísel $s_{k+1}, \ldots, s_p$.

Měříme kvalitu aproximace

Tvrzení

Pro každou matici $\mathbf{A} \in \mathbb{R}^{m \times n}$ hodnosti r platí

$$\|\mathbf{A}\| = \sqrt{s_1^2 + \cdots + s_r^2}.$$

Relativní chybu aproximace maticí $\mathbf{B}^*$ hodnosti nejvýše k lze zjistit ze singulárních čísel matice $\mathbf{A}$:

- Pro $k = 1, \ldots, r-1$ dostaneme

$$\frac{\|\mathbf{A} - \mathbf{B}^*\|}{\|\mathbf{A}\|} = \sqrt{\frac{s_{k+1}^2 + \cdots + s_r^2}{s_1^2 + \cdots + s_r^2}}$$

- Pro $r \leq k \leq p$ triviálně platí $\mathbf{B}^* = \mathbf{A}$ a chyba je 0

Komprese obrázku psa pomocí SVD

$k = 5$, chyba 17%, $5 \times (3456 + 4608)$

$k = 20$, chyba 9%, $20 \times (3456 + 4608)$

$k = 200$, chyba 4%, $200 \times (3456 + 4608)$

Originál 3456×4608

PCA pomocí Eckart-Youngovy věty

Matice $\mathbf{A} = \begin{bmatrix}\mathbf{a}_1 \cdots \mathbf{a}_n\end{bmatrix} \in \mathbb{R}^{m \times n}$ má ve sloupcích datové vektory $\mathbf{a}_i$ a jejich těžiště je $\bar{\mathbf{a}} = \mathbf{0}$. Promítáme je na podprostor dimenze k.

Řešení

1. Spočítáme redukované SVD $\mathbf{A} = \mathbf{USV}^\mathsf{T}$ a seřadíme sestupně singulární čísla, $s_1 \geq \cdots \geq s_p$
2. Označíme $\mathbf{U} = [\underbrace{\mathbf{u}_1 \cdots \mathbf{u}_k}_{\mathbf{U}_k} \, \mathbf{u}_{k+1} \cdots \mathbf{u}_m]$
3. Sloupce matice $\mathbf{U}_k$ jsou ortonormální bází hledaného podprostoru dimenze k, neboť jsou to vl. vektory matice $\mathbf{AA}^\mathsf{T}$
4. Souřadnice promítnutých bodů jsou $\mathbf{U}_k\mathbf{A}$
5. Optimální hodnota úlohy (absolutní chyba) je $s_{k+1}^2 + \cdots + s_p^2$

Výběrová kovarianční matice

Matice $\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 \cdots \mathbf{a}_n \end{bmatrix} \in \mathbb{R}^{m \times n}$ má datové vektory $\mathbf{a}_i$ ve sloupcích a předpokládáme, že výběrové průměry veličin jsou nulové, $\overline{\mathbf{a}} = \mathbf{0}$.

Definice

Výběrová kovarianční matice je matice $\frac{1}{n}\mathbf{A}\mathbf{A}^\top$. Na diagonále má *výběrové rozptyly* a jinde *výběrové kovariance*.

Stejnou interpretaci má matice $\frac{1}{n}\mathbf{A}^\top\mathbf{A}$, pokud jsou normalizované datové vektory uspořádány do řádků matice $\mathbf{A} \in \mathbb{R}^{n \times m}$.

Příklad (1)

$n = 3$ recenzenti hodnotí $m = 4$ filmy body $0, \dots, 5$.
Po normalizaci průměry dostaneme $\mathbf{A}$ a kovarianční matici:

$$\begin{bmatrix} 5 & 4 & 1 \\ 5 & 5 & 0 \\ 0 & 0 & 5 \\ 1 & 0 & 4 \end{bmatrix} \qquad \mathbf{A} = \begin{bmatrix} 1.67 & 0.67 & -2.33 \\ 1.67 & 1.67 & -3.33 \\ -1.67 & -1.67 & 3.33 \\ -0.67 & -1.67 & 2.33 \end{bmatrix}$$

$$\tfrac{1}{3}\mathbf{A}\mathbf{A}^{\mathsf{T}} = \begin{bmatrix} 2.89 & 3.89 & -3.89 & -2.56 \\ 3.89 & 5.56 & -5.56 & -3.89 \\ -3.89 & -5.56 & 5.56 & 3.89 \\ -2.56 & -3.89 & 3.89 & 2.89 \end{bmatrix}$$

Hodnoty kovariancí naznačují, že existují $k = 2$ typy filmů.

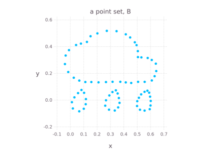
Příklad (2)

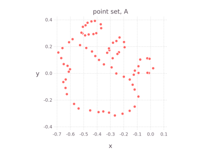
- Pomocí SVD matice $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^\mathsf{T}$ dostaneme levé singulární vektory $\mathbf{U} = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 & \mathbf{u}_4 \end{bmatrix}$, jsou seřazeny podle singulárních čísel $s_1 = 7.05$, $s_2 = 1$, $s_3 = s_4 = 0$
- PCA pro $k = 2$: klademe $\mathbf{U}_2 := \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 \end{bmatrix}$ a matice souřadnic datových vektorů promítnutých do $\mathbb{R}^2$ je

$$\mathbf{U}_2^\mathsf{T}\mathbf{A} = \begin{bmatrix} -2.88 & -2.88 & 5.74 \\ -0.71 & 0.71 & 0 \end{bmatrix}$$

- Chyba aproximace je $s_3 + s_4 = 0$ a proto jsou data pouze *dvourozměrná*

Ortogonální Prokrustův problém





- Matice $\mathbf{A} = \begin{bmatrix}\mathbf{a}_1 \cdots \mathbf{a}_n\end{bmatrix} \in \mathbb{R}^{m\times n}$ a $\mathbf{B} = \begin{bmatrix}\mathbf{b}_1 \cdots \mathbf{b}_n\end{bmatrix} \in \mathbb{R}^{m\times n}$
- Hledáme ortogonální matici $\mathbf{X} \in \mathbb{R}^{m\times m}$ minimalizující

$$\sum_{i=1}^{k} \|\mathbf{X}\mathbf{a}_i - \mathbf{b}_i\|^2 = \|\mathbf{X}\mathbf{A} - \mathbf{B}\|^2$$

Optimální řešení

$\mathbf{X}^* = \mathbf{U}\mathbf{V}^\mathsf{T}$, kde $\mathbf{B}\mathbf{A}^\mathsf{T} = \mathbf{U}\mathbf{S}\mathbf{V}^\mathsf{T}$ je redukované SVD