Optimalizace

PCA jako úloha na největší stopu

Tomáš Kroupa Tomáš Werner

2024

Fakulta elektrotechnická
ČVUT v Praze

Motivace

Databáze `iris` obsahuje $n = 150$ kosatců. U každého je uveden jeho druh a $m = 4$ charakteristiky jeho kališního/okvětního lístku.

Redukce dimenze a vizualizace

Můžeme rovnou natrénovat klasifikátor kosatců, ale před tím je vhodné získat představu o rozložení dat v prostoru dimenze 4.

- Datové vektory $\mathbf{a}_1, \ldots, \mathbf{a}_{150} \in \mathbb{R}^4$ popisující 4 charakteristiky kosatců promítneme na afinní podprostory dimenzí $k = 1, 2, 3$
- Chyba proložení udává vhodnost takové redukce dimenze
- Souřadnice promítnutých bodů lze zobrazit

Stopa matice

Stopa matice

Stopa čtvercové matice $\mathbf{A} \in \mathbb{R}^{n \times n}$ je číslo

$$\operatorname{tr} \mathbf{A} = a_{11} + \cdots + a_{nn}.$$

Vlastnosti

1. $\operatorname{tr}(\mathbf{A} + \mathbf{B}) = \operatorname{tr} \mathbf{A} + \operatorname{tr} \mathbf{B}, \quad \operatorname{tr}(\alpha \mathbf{A}) = \alpha \operatorname{tr} \mathbf{A}$
2. $\operatorname{tr}(\mathbf{A}^{\mathsf{T}}) = \operatorname{tr} \mathbf{A}$
3. $\operatorname{tr}(\mathbf{AB}) = \operatorname{tr}(\mathbf{BA})$, kde $\mathbf{A} \in \mathbb{R}^{m \times n}$ a $\mathbf{B} \in \mathbb{R}^{n \times m}$
4. $\operatorname{tr} \mathbf{A} = \lambda_1 + \cdots + \lambda_n$, kde $\mathbf{A}$ je symetrická

Skalární součin matic

Pro matice $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$ definujeme skalární součin

$$\langle \mathbf{A}, \mathbf{B} \rangle = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} b_{ij}.$$

Vlastnosti

- Pro $m = 1$ nebo $n = 1$ je to standardní skalární součin vektorů
- Platí $\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{tr}(\mathbf{A}^\mathsf{T} \mathbf{B}) = \mathrm{tr}(\mathbf{B}^\mathsf{T} \mathbf{A})$

Norma matice

Frobeniova norma matice $\mathbf{A} \in \mathbb{R}^{m \times n}$ je

$$\|\mathbf{A}\| = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2}.$$

Vlastnosti

$$\|\mathbf{A}\| = \sqrt{\langle \mathbf{A}, \mathbf{A} \rangle} = \sqrt{\mathrm{tr}(\mathbf{A}\mathbf{A}^\mathsf{T})} = \sqrt{\lambda_1 + \cdots + \lambda_m},$$

kde $\lambda_1 \geq \cdots \geq \lambda_m \geq 0$ jsou vlastní čísla PSD matice $\mathbf{A}\mathbf{A}^\mathsf{T}$.

Příklad

Ortogonální projektory ze spektrálního rozkladu

Ve spektrálním rozkladu symetrické matice řádu n

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathsf{T}} = \lambda_1 \mathbf{v}_1 \mathbf{v}_1^{\mathsf{T}} + \cdots + \lambda_n \mathbf{v}_n \mathbf{v}_n^{\mathsf{T}}$$

jsou dva různé ortogonální projektory $\mathbf{v}_i \mathbf{v}_i^{\mathsf{T}}$ a $\mathbf{v}_j \mathbf{v}_j^{\mathsf{T}}$ kolmé,

$$\langle \mathbf{v}_i \mathbf{v}_i^{\mathsf{T}}, \mathbf{v}_j \mathbf{v}_j^{\mathsf{T}} \rangle = \operatorname{tr}(\mathbf{v}_i \underbrace{\mathbf{v}_i^{\mathsf{T}} \mathbf{v}_j}_{0} \mathbf{v}_j^{\mathsf{T}}) = 0.$$

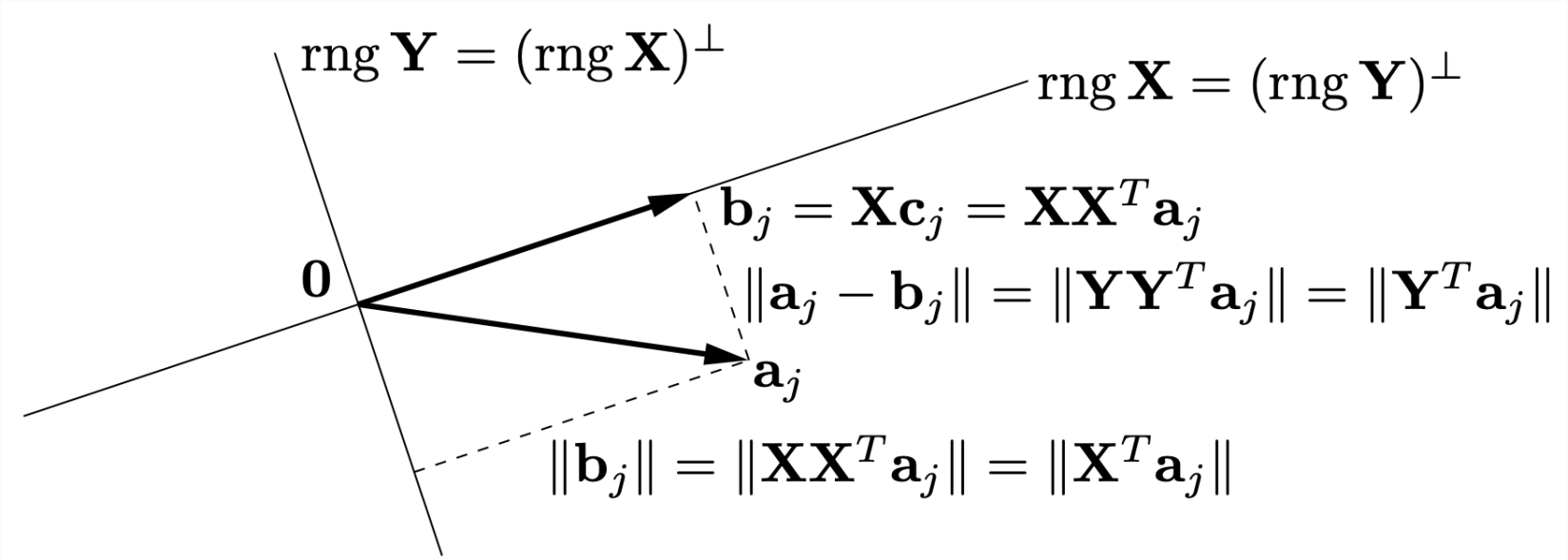
Ortogonální projektor $\mathbf{v}_i \mathbf{v}_i^{\mathsf{T}}$ má vlastní čísla $\lambda_1 = 1$ a $\lambda_i = 0$ pro $i = 2, \ldots, n$, tedy jeho Frobeniova norma je

$$\|\mathbf{v}_i \mathbf{v}_i^{\mathsf{T}}\|^2 = \operatorname{tr}(\mathbf{v}_i \underbrace{\mathbf{v}_i^{\mathsf{T}} \mathbf{v}_i}_{1} \mathbf{v}_i^{\mathsf{T}}) = \operatorname{tr}(\mathbf{v}_i \mathbf{v}_i^{\mathsf{T}}) = 1.$$

Úloha na největší stopu

Proložení bodů lineárním podprostorem

Pro vektory $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^m$ hledáme lineární podprostor dimenze $k \leq m$ vyjádřený ortonormální bází $\mathbf{X} \in \mathbb{R}^{m \times k}$ *minimalizující součet čtverců kolmých vzdáleností* $\mathbf{a}_1, \ldots, \mathbf{a}_n$ od $\operatorname{rng} \mathbf{X}$.



Úloha PCA

Minimalizuj $\sum_{i=1}^{n} \|\mathbf{a}_i - \mathbf{X}\mathbf{X}^\mathsf{T}\mathbf{a}_i\|^2$ za podmínky $\mathbf{X} \in \mathbb{R}^{m \times k}$, $\mathbf{X}^\mathsf{T}\mathbf{X} = \mathbf{I}$

Největší stopa matice na podprostoru

Úloha PCA ekvivalentně

Maximalizuj $\sum_{i=1}^{n} \|\mathbf{X}^\mathsf{T}\mathbf{a}_i\|^2$ za podmínky $\mathbf{X} \in \mathbb{R}^{m\times k}$, $\mathbf{X}^\mathsf{T}\mathbf{X} = \mathbf{I}$

- Uvažujeme matici $\mathbf{A} = [\mathbf{a}_1 \dots \mathbf{a}_n] \in \mathbb{R}^{m\times n}$
- Z definice Frobeniovy normy a stopy dostaneme

$$\sum_{i=1}^{n} \|\mathbf{X}^\mathsf{T}\mathbf{a}_i\|^2 = \|\mathbf{X}^\mathsf{T}\mathbf{A}\|^2 = \operatorname{tr}(\mathbf{X}^\mathsf{T}\mathbf{A}\mathbf{A}^\mathsf{T}\mathbf{X})$$

PCA jako úloha na největší stopu

Maximalizuj $\operatorname{tr}(\mathbf{X}^\mathsf{T}\mathbf{A}\mathbf{A}^\mathsf{T}\mathbf{X})$ za podmínky $\mathbf{X} \in \mathbb{R}^{m\times k}$, $\mathbf{X}^\mathsf{T}\mathbf{X} = \mathbf{I}$

Řešení úlohy na největší stopu

Věta

Nechť $\mathbf{B} \in \mathbb{R}^{m \times m}$ je symetrická matice se spektrálním rozkladem $\mathbf{B} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\mathsf{T}$, vlastními čísly $\lambda_1 \geq \cdots \geq \lambda_m$ a $k \leq m$. Platí

$$\max \left\{ \operatorname{tr}(\mathbf{X}^\mathsf{T}\mathbf{B}\mathbf{X}) \mid \mathbf{X} \in \mathbb{R}^{m \times k},\ \mathbf{X}^\mathsf{T}\mathbf{X} = \mathbf{I} \right\} = \lambda_1 + \cdots + \lambda_k$$

a maxima se nabývá pro $\mathbf{X} = [\mathbf{v}_1 \ldots \mathbf{v}_k]$.

V úloze PCA je $\mathbf{B} = \mathbf{A}\mathbf{A}^\mathsf{T}$, přičemž

- $\mathbf{B}$ je pozitivně semidefinitní a
- její vlastní čísla splňují $\lambda_1 \geq \cdots \geq \lambda_m \geq 0$.

PCA kuchařka

PCA jako instance úlohy na nejmenší stopu

$$\max\left\{\operatorname{tr}(\mathbf{X}^\mathsf{T}\mathbf{A}\mathbf{A}^\mathsf{T}\mathbf{X}) \mid \mathbf{X} \in \mathbb{R}^{m\times k},\, \mathbf{X}^\mathsf{T}\mathbf{X} = \mathbf{I}\right\}$$

1. Najdeme spektrální rozklad $\mathbf{A}\mathbf{A}^\mathsf{T} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\mathsf{T} \in \mathbb{R}^{m\times m}$ při řazení $\lambda_1 \geq \cdots \geq \lambda_m$ a $\mathbf{V} = [\underbrace{\mathbf{v}_1 \cdots \mathbf{v}_k}_{\mathbf{X}}\ \underbrace{\mathbf{v}_{k+1} \cdots \mathbf{v}_m}_{\mathbf{Y}}]$
2. Sloupce matice $\mathbf{X}$ tvoří ortonormální bázi hledaného lineárního podprostoru dimenze k
3. Sloupce matice $\mathbf{Y}$ jsou ortonormální bází jeho ortogonálního doplňku dimenze $m - k$
4. Chyba proložení je $\lambda_{k+1} + \cdots + \lambda_m$

Co když prokládáme afinním podprostorem?

Tvrzení

Afinní podprostor dimenze k, který minimalizuje součet čtverců vzdáleností k vektorům $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^m$, obsahuje jejich těžiště

$$\overline{\mathbf{a}} = \frac{1}{n}(\mathbf{a}_1 + \cdots + \mathbf{a}_n).$$

1. Vektory $\mathbf{a}_i$ posuneme tak, aby měly těžiště v $\mathbf{0}$:

$$\mathbf{a}_1 - \overline{\mathbf{a}}, \ldots, \mathbf{a}_n - \overline{\mathbf{a}}$$

2. Posunuté vektory proložíme lineárním prostorem X dimenze k
3. Hledaný afinní podprostor je $X + \overline{\mathbf{a}}$

Příklad

PCA pro $m = 3$ a $n = 4$

Vektory $\mathbf{a}_1 = (1, 3, 0)$, $\mathbf{a}_2 = (2, 1, 1)$, $\mathbf{a}_3 = (-1, 3, 0)$
a $\mathbf{a}_4 = (2, -3, 0)$ se zřejmě příliš neliší v poslední souřadnici.
Přesvědčí nás o tom PCA pro dimenzi $k = 2$ (rovina).

- Vezmeme matici vystředěných vektorů $\mathbf{A}$
- $\mathbf{A}\mathbf{A}^{\mathsf{T}} = \begin{bmatrix} 6 & -8 & 1 \\ -8 & 24 & 0 \\ 1 & 0 & 0.75 \end{bmatrix} = \mathbf{V}\,\mathrm{diag}(27.0, 3.3, 0.4)\,\mathbf{V}^{\mathsf{T}}$
- Chyba proložení podprostorem s bází $\mathbf{v}_1, \mathbf{v}_2$ je $\lambda_3 = 0.4$
- Relativní chyba proložení $\frac{\lambda_3}{\lambda_1+\lambda_2+\lambda_3} \approx 0.01$

Příklad – `iris` (1)

Matice $\mathbf{A} \in \mathbb{R}^{4\times 150}$ má v každém z $n = 150$ sloupců měření $m = 4$ proměnných, od nichž jsme odečetli $\bar{\mathbf{a}}$. Volíme dimenzi $k = 2$.

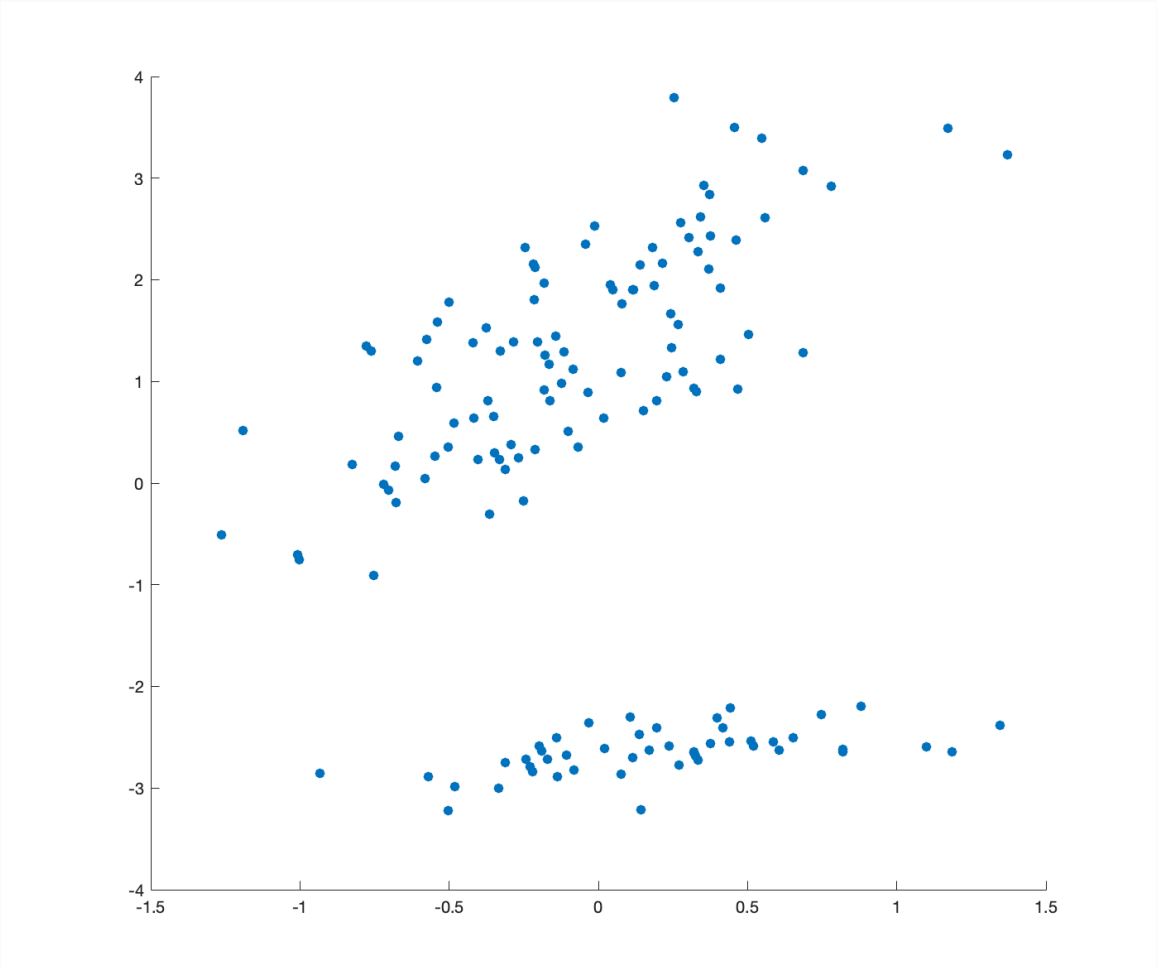
Řešení

- $\mathbf{A}\mathbf{A}^\mathsf{T} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\mathsf{T}$, kde $\mathbf{\Lambda} = \operatorname{diag}(629.50,\ 36.10,\ 11.70,\ 3.53)$
- Hledaný podprostor má bázi $\mathbf{X} = [\mathbf{v}_1\ \mathbf{v}_2] \in \mathbb{R}^{4\times 2}$
- Chyba je $\lambda_3 + \lambda_4$
- Relativní chyba proložení je

$$\frac{\lambda_3 + \lambda_4}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4} \approx 0.02$$

Příklad – `iris` (2)

- Chceme si zobrazit první dvě hlavní komponenty
- Zobrazíme si souřadnice promítnutých bodů v $\mathbb{R}^2$
- Nalezneme je ve sloupcích matice $\mathbf{X}^\mathsf{T}\mathbf{A} \in \mathbb{R}^{2\times 150}$



Shrnutí PCA

- Ortonormální báze nalezeného podprostoru je v $\mathbf{X} \in \mathbb{R}^{m \times k}$
- Ortogonální projekce $\mathbf{a}_i$ na ten podprostor je $\mathbf{b}_i = \mathbf{X}\mathbf{X}^\mathsf{T}\mathbf{a}_i$
- Vektor souřadnic bodu $\mathbf{b}_i$ v ortonormální bázi $\mathbf{X}$ je $\mathbf{X}^\mathsf{T}\mathbf{a}_i$
- Matice souřadnic těch bodů je $\mathbf{X}^\mathsf{T}\mathbf{A} \in \mathbb{R}^{k \times n}$

Využití

1. Komprese: $\mathbf{A}$ má mn prvků, $\mathbf{X}$ a $\mathbf{X}^\mathsf{T}\mathbf{A}$ mají $(m+n)k$ prvků
2. Redukce dimenze: Body $\mathbf{X}^\mathsf{T}\mathbf{A}$ jsou v menší dimenzi než $\mathbf{A}$
3. Vizualizace: Pro $k \leq 3$ lze souřadnice v $\mathbf{X}^\mathsf{T}\mathbf{A}$ zobrazit
4. Rozpoznávání: Body v $\mathbf{X}^\mathsf{T}\mathbf{A}$ jsou vhodnější pro klasifikaci

Související úlohy

Speciální případ: úloha PCA pro $k = 1$

- V této situaci prokládáme data jen přímkou a hledáme tedy jednotkový vektor $\mathbf{x} \in \mathbb{R}^m$ maximalizující kritérium

$$\|\mathbf{x}^\mathsf{T}\mathbf{A}\|^2 = \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{A}^\mathsf{T}\mathbf{x}$$

- Tento problém je instancí obecnější úlohy na maximalizaci kvadratické formy na sféře:

Tvrzení

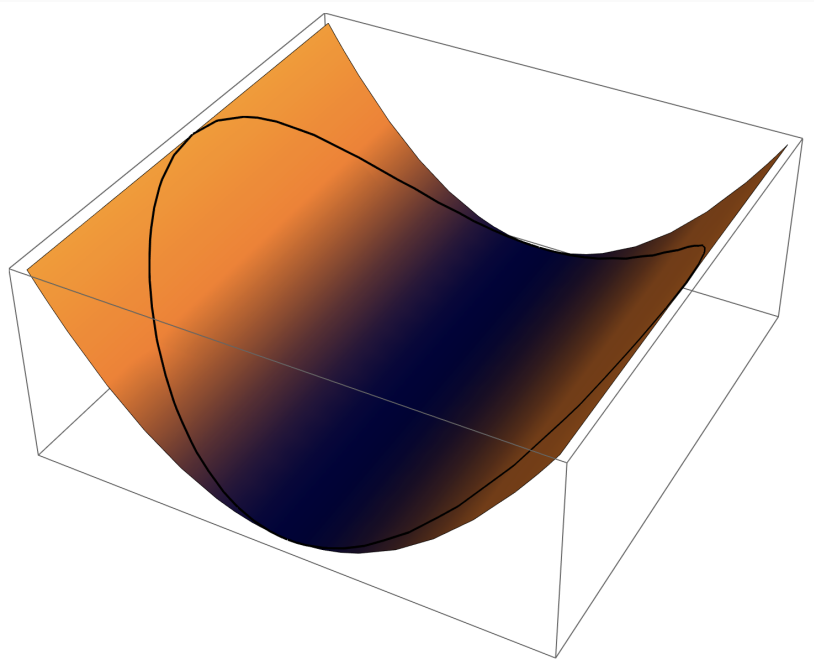
Nechť $\mathbf{B} \in \mathbb{R}^{m\times m}$ je symetrická s vlastními čísly $\lambda_1 \geq \cdots \geq \lambda_m$ a ortonormální bází vlastních vektorů $\mathbf{v}_1, \ldots, \mathbf{v}_m$. Potom platí

$$\max\{\mathbf{x}^\mathsf{T}\mathbf{B}\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^m,\ \|\mathbf{x}\| = 1\} = \lambda_1$$

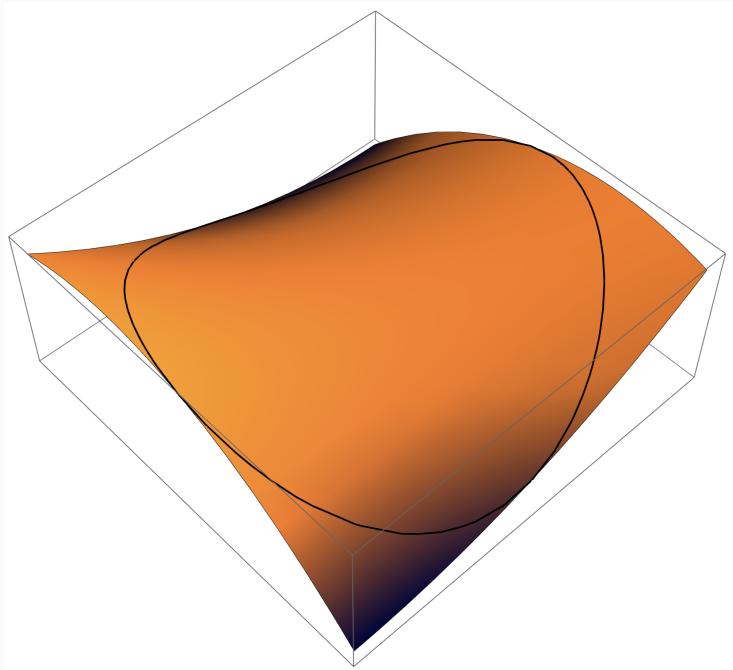
kde maxima se nabývá pro $\mathbf{v}_1$.

Příklady

- $\mathbf{B} = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}$
- Forma $2x_1^2$
- Vlastní čísla 2 a 0
- Vlastní vektory $(1, 0)$ a $(0, 1)$



- $\mathbf{B} = \begin{bmatrix} -2 & 2 \\ 2 & 1 \end{bmatrix}$
- Forma $-2x_1^2 + x_2^2 + 4x_1x_2$
- Vlastní čísla 2 a -3
- Vl.vektory $(1, 2)$ a $(-2, 1)$

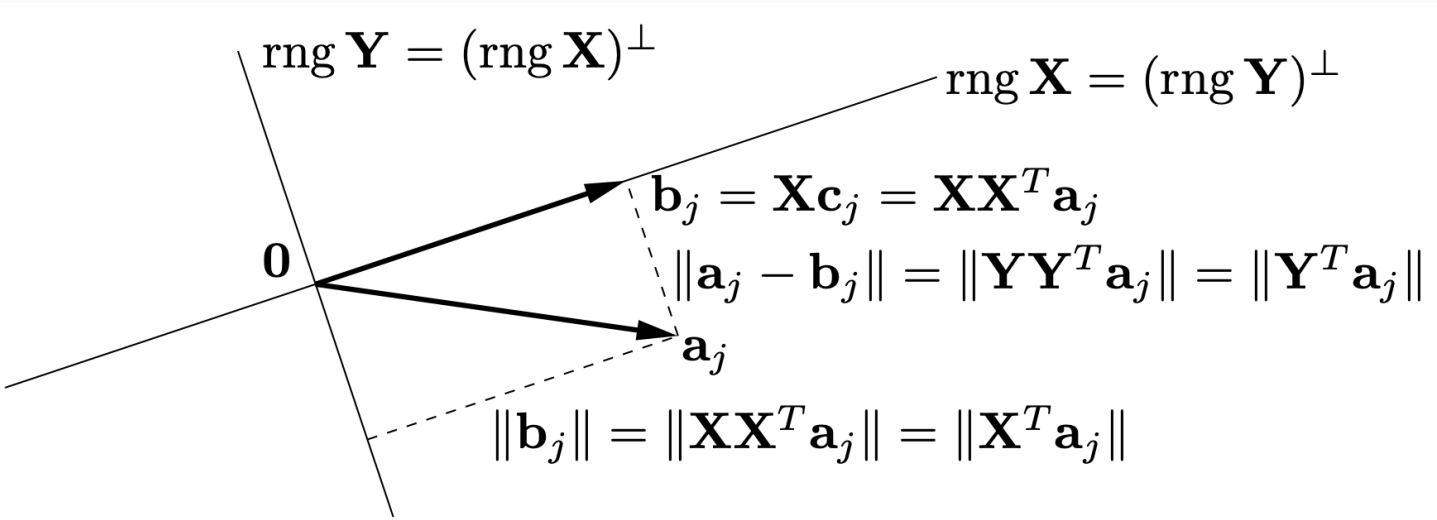


Nejbližší matice nižší hodnosti

Tato úloha je ekvivalentní úloze PCA:

Low rank approximation pro matici $\mathbf{A} = [\mathbf{a}_1 \dots \mathbf{a}_n]$

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



Optimální řešení je $\mathbf{B} = \mathbf{X}\mathbf{X}^\mathsf{T}\mathbf{A}$.

Poznámky

- Ve statistice se $\frac{1}{n}\mathbf{A}\mathbf{A}^\mathsf{T}$ nazývá empirická kovarianční matice pro data $\mathbf{a}_1, \ldots, \mathbf{a}_n$ splňující $\bar{\mathbf{a}} = \mathbf{0}$ a úloha PCA vede na maximalizaci variance promítnutých dat
- PCA lze aternativně řešit pomocí SVD (singulárního rozkladu)
- Semestrální projekt Jana Rutterleho
  https://ruttejan.github.io/PCA/
- *Why are Big Data Matrices Approximately Low Rank?*
  https://arxiv.org/abs/1705.07474