RDD - 06

22 March 2021 14:42

$$\underset{p\times 1}{X} = \underset{\text{fact. loadings}}{A}\ \underset{q\times 1}{\mathbb{F}}^{q\times q \leftarrow \text{hidden factors}} + \underset{\text{uniqueness}}{\mathbb{U}}$$

Minule: Fakt. analýza

1.) Odhad A
   a) hl. faktory
   b) MLE

2.) Rotácie < varimax / kvartimax

3.) Odhad faktorových skóre: $\hat{F}^{(i)} = \hat{A}^T(\hat{A}\hat{A}^T + \hat{D})^{-1} x_i \quad , i = 1, \dots, n$

Kritika:

- nejednoznačnosť ⇒ príliš si môžeme vymýšľať interpretáciu
- normálnosť v MLE – nemusí byť v dátach
- numerické problémy – lok. maximá
- Heywood cases – degenerované prípady: niektorý kor. koef. > 1

Pr.: psych tests.csv

- výsledky študentov na 24 testoch (Chicago, 30-te roky)
  ↳ 2 školy

F1 : ~ 5–9 – slovný, jazykový faktor

F2: ~ 1–4, (13), 23 – vizuálna schopnosť, rozoznávanie vzťahov

F3: ~ 10–13 – rýchlosť

F4: ~ 14–18 – pamäť

↳ namiesto 24 čísel môžeme reprezentovať výsledky pomocou 4 faktorov

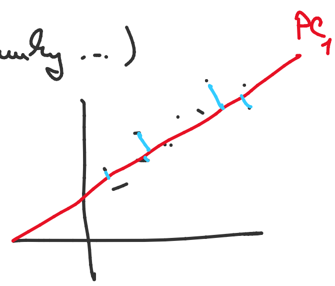
---

2. Nelineárne metódy extrakcie premenných

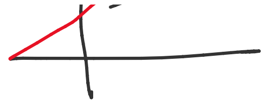
- nelineárne v dátach

Nelineárna PCA (nPCA)

- rozšíriť myšlienku PCA na nelineárne štruktúry (nie vždy iba priamky ...)
- PCA – max. rozptyl
  - najmenšia vzdialenosť dát od projekcií:
    nadroviny prechádzajúce "stredom" dát



nadroviny prechádzajúce "stredom" dát

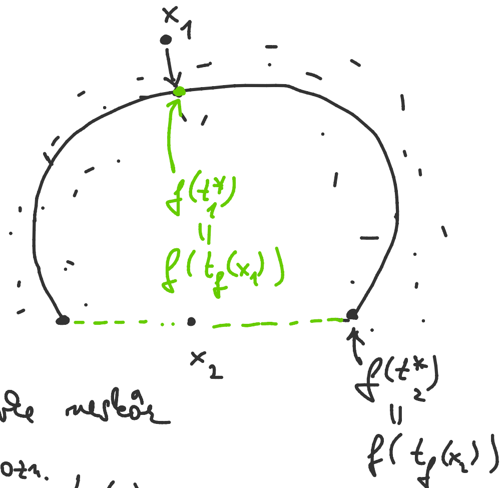


1. Hlavné krivky a plochy (principal curves and p. surfaces)

- hľadáme krivku, ktorá prechádza "stredom" dát
- historicky významné: Hastie (1984) ... → jedna z prvých nPCA
- hlavná krivka: hladká krivka, ktorá prechádza stredom dát

  hl. povrch: hladký povrch, — || —

krivka: $f: (a,b) \to \mathbb{R}^k$, $f(t) = \begin{pmatrix} f_1(t) \\ \vdots \\ f_k(t) \end{pmatrix}$



$x \in \mathbb{R}^k \Rightarrow$ najbližší bod na krivke f: $\|x - f(t)\| \to \min_t$

↳ $t^*: \|x - f(t^*)\| = \min_t \|x - f(t)\|$

- môže byť nejednoznačné: zvolíme ten, ktorý je na krivke neskôr

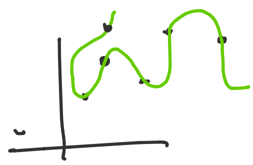
$$t^* = \max\{\tilde{t} \mid \|x - f(\tilde{t})\| = \min_t \|x - f(t)\|\} \overset{\text{ozn.}}{=} t_f(x)$$

výberovo/dátovo: $\hat{f}$ – hľadáme $\underset{\text{reconstruction error}}{RE}(X, f) = \sum_{i=1}^{n} \underbrace{\|x_i - f(t_f(x_i))\|^2}_{\text{vzdialenosť od projekcie na krivku}} \to \min_f$

hl. krivka

- minimalizujem vzdial. krivky od dát
- optimum: krivka prechádzajúca dátami

  ⇒ vyžadujeme isté vyhladenia



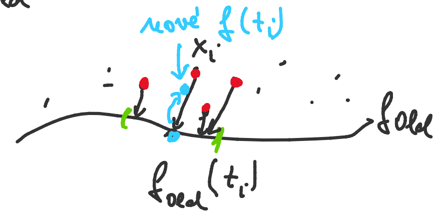
→ v praxi: krivka prechádzajúca stredom dát:

Algoritmus (Projection–Expectation)

- začneme s $f^{(0)} = PC_1$: priamka daná prvým hl. komp

(i) projekčný krok: pre danú f_{old} dopočítame t_f: $t_i = t_{f_{old}}(x_i)$ $\quad i = 1, \dots, n$

$\Rightarrow f(t_f)$: proj. x na f



(ii) priemerovací krok: pre dané $t_1, \dots, t_n$

– nová hodnota krivky v t_i: spriemerujeme x-ká, kt. sa premietajú do okolia t_i

Presne: • usporiadame t_i: $t_{(1)} < t_{(2)} < \dots < t_{(n)}$

a k nim: $x_{(1)}, x_{(2)}, \dots, x_{(n)}$

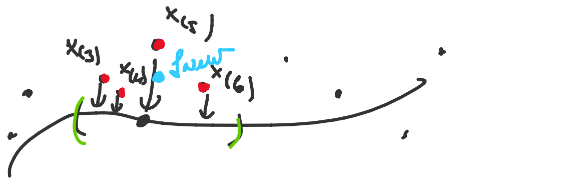
$N(t)$ = okolie f okolo t $\Rightarrow$ $N_{(i)} = \{ x_{(j)} \mid t_{(j)} \in N(t_i) \}$

tie $x_{(j)}$, kt. sa projektujú do okolia t_i

$$\Rightarrow f_{new}(t_i) = \frac{1}{|N_{(i)}|} \cdot \sum_{x_{(j)} \in N_{(i)}} x_{(j)}$$: priemer x-ov z $N_{(i)}$

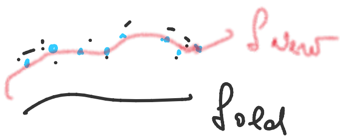
→ získame f_{new} v bodoch $t_{(1)}, \dots, t_{(n)}$

– medzi nimi: vyhladenie, napr. kubický splajn



$\Rightarrow N_{(5)} = \{ x_{(3)}, x_{(4)}, x_{(5)}, x_{(6)} \}$

• opakujeme (i), (ii)



• vylepšenia: zníženie vychýlenosti (biasu)

• hl. povrchy: 2-rozmerná analógia – projektujeme do 2-rozmeru

Úloha: food.txt – zloženie jedál