Lineárna Regresia 1

Prednáška č.10

Lukáš Lafférs

KM FPV UMB
www.lukaslaffers.com

Scvrkávacie metódy (shrinkage methods)

Máme príliš veľa regresorov?

Ako využiť túto informáciu?

PCA

100 pozorovaní

40 premenných

X je matica 100 krát 40

Začíname

Nastavíme u_1 (vektor rozmeru 40) tak, aby

$var(Xu_1) \rightarrow \max$

Okrem toho budeme požadovať, aby $u_1^T u_1 = 1$ (aby bol vektor u_1 jednoznačne definovaný).

Xu_1 je **prvý hlavný komponent**

Pokračujeme

Nastavíme u_2 tak, aby

$var(Xu_2) \to \max$

zatiaľčo $u_2^T u_1 = 0$ a zároveň $u_2^T u_2 = 1$.

Xu_2 je **druhý hlavný komponent**

Pokračujeme

Týmto spôsobom pokračujeme a vieme reprezentovať každý bol v oblaku bodov matice X ako jednoznačnú lineárnu kombináciu hlavných komponentov.

Pointou je, že napríklad 5 hlavných komponentov môže vysvetliť až 99% variácie v X. To znamená, že nepotrebujeme 40 čísel na reprezentovanie jedného porozorania ale stačí nám napr. len 5.

Hlavné komponenty

- sú ortogonálne čo je výhodné, ak ich používame ako prediktory v regresii, lebo pridaním ďalšieho hlavného komponentu sa nám nezmenia odhady parametrov. Okrem toho odhady sú numericky stabilnejšie.
- môžu šetriť pamäť alebo miesto na disku.
- môžu ale nemusia byť interpretovateľné. Niekedy tým, že vidíme akej lineárnej kombinácii u_1 zodpovedá prvý komponent, tak môžme pochopiť o čo ide, a to nám vie pridať vhľad do problematiky.
- nám môžu pomôcť odhaliť zhluky podobných bodov. V 40 rozmeroch nás častokrát naša intuícia opúšťa.

Príklad 1

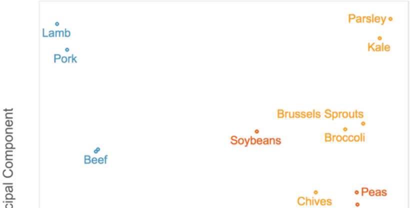
Príklad 1 - Factor Loadings

	PC1	PC2	PC3	PC4
Fat	-0.45	0.66	0.58	0.18
Protein	-0.55	0.21	-0.46	-0.67
Fiber	0.55	0.19	0.43	-0.69
Vitamin C	0.44	0.70	-0.52	0.22

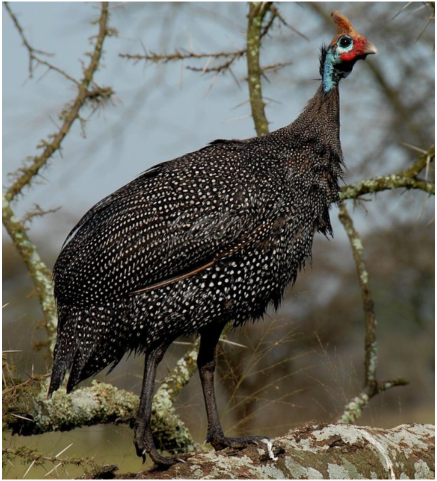
Obr. 2: Zdroj:
https://www.quora.com/What-is-an-intuitive-explanation-for-PCA

Príklad 1


Vitamin C
Parsley
Kale
Broccoli
Cauliflower
Soybeans
Yam
Guinea Hen
(Vitamin C) - (Fat)
Parsley
Kale
Broccoli
Cauliflower
Cabbage
Spinach
Yam
Sweet Corn
Guinea Hen
Bluefish
Mackerel
Chicken
Beef
Pork
Lamb
(Vitamin C + Fiber) - (Fat)
Parsley
Peas
Lotus Root
Chives
Cauliflower
Soybeans
Eggplant
Sweet Corn
Mushrooms
Haddock
Guinea Hen
Bluefish
Mackerel
Chicken
Beef
Pork
Lamb


cipal Component
Lamb
Pork
Beef
Parsley
Kale
Brussels Sprouts
Soybeans
Broccoli
Chives
Peas

Guinea Hen



Obr. 4: Zdroj: wiki

Príklad 2

Obr. 5: Zdroj: `https://github.com/gbuesing/pca/tree/master/examples`

Príklad 2 - rekonštrukcia pomocou 36 hlavných komponentov

Obr. 6: Zdroj: `https://github.com/gbuesing/pca/tree/master/examples`

Príklad 2 - 'Eigenfaces'

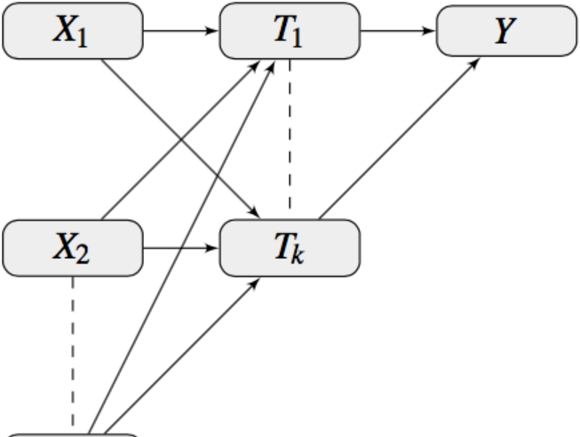
Obr. 7: Zdroj: `https://github.com/gbuesing/pca/tree/master/examples`

Partial Least Squares

Našou úlohou je nájsť také ortogonálne kombinácie $T_1, ..., T_k$ prediktorov $X_1, ..., X_p$, že predikcia

$$\hat{y} = \beta_1 T_1 + ... + \beta_k T_k,$$

je čo najlešia možná. Na odhadnute T_i sú rôzne algoritmy a ich počet sa vyberá pomocou krížovej validácie.



Ridge regression

Ako zostabilniť odhady parametrov? Tak, že im "zakážeme"príliš veľké hodnoty. Toto môžeme urobiť viacerými spôsobmi. Jedným z nich je hrebeňová regresia. Namiesto minimalizácie štvorcov, minimalizujeme

$$(y - X\beta)^T(y - X\beta) + \lambda \sum_j \beta_j^2$$

čo je ekvivalentné s

$$(y - X\beta)^T(y - X\beta) \quad \text{subject to} \quad \sum_j \beta_j^2 \leq t^2.$$

Odhad je $\hat{\beta} = (X^TX + \lambda I)^{-1}X^Ty$. Toto je príklad penalizovanej regresie.
Pre $\lambda \to 0$ dostávame $\hat{\beta} \to \hat{\beta}_{LS}$ a pre $\lambda \to \infty$ dostávame $\hat{\beta} \to 0$.
Parameter λ nastavíme pomocou krížovej validácie.
Odhad parametrov je vychýlený ale to je cena, ktorú platíme za stabilnejší, teda menej variabilný odhad.

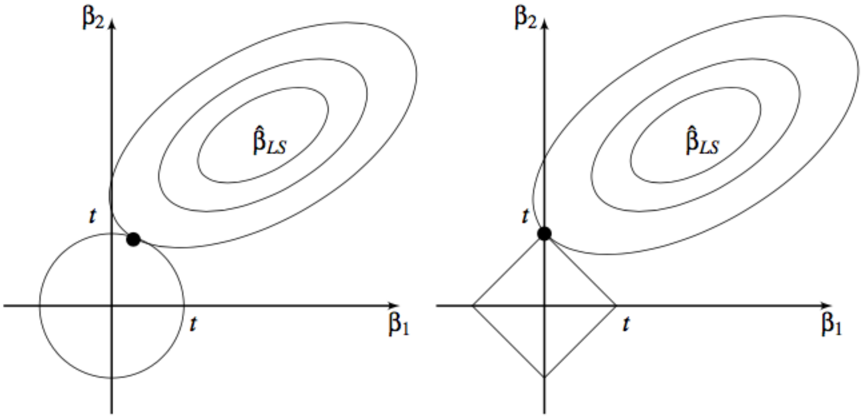
LASSO

$$(y - X\beta)^T(y - X\beta) + \lambda \sum_j |\beta_j|$$

čo je ekvivalentné s

$$(y - X\beta)^T(y - X\beta) \quad \text{subject to} \quad \sum_j |\beta_j| \leq t$$

Výhodou Lassa je, že vďaka tvaru penalty niektoré parametre priamo vynuluje. Preto akoby robil výber modelu aj odhadovanie modelu naraz! Lasso je rozumné používať ak sa domnievame, že existuje niekoľko silných efektov a veľa iných prediktorov nemá na odozvu žiaden vplyv.



Obr. 9: Stratová funkcia má minimum v $\hat{\beta}_{LS}$, tam sa dosahuje najlepší fit. Penalta λ však posúva optimum bližie k nule (scvrkáva parametre), a to na kružnicu (vľavo ridge: $\sum_{j=1}^{p} \beta_j^2 = t^2$) alebo štvorec (vpravo LASSO : $\sum_{j=1}^{p} |\beta_j| = t$). Zdroj: Faraway (2014)

Ďakujem za pozornosť.