Přednáška 3: Zobecněný lineární regresní model, regularizace volbou apriorna

Ondřej Tichý

Bayesovské metody v machine learningu (BML)

March 2, 2018

Obsah přednášky

- Zobecněný lineární model a k čemu je to dobré
- Nejmenší čtverce a v čem je problém
- Regularizace
- Formulace pomocí optimalizace
  - Tichonovova regularizace
  - LASSO
  - elastic net
- Bayesovská formulace problému
  - bayesovská regrese
  - ridge regression
  - sparse regression

Zobecněný lineární regresní model

Opakování

Předpoklad vysvětlení každé naměřené hodnoty:

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + e_i. \tag{1}$$

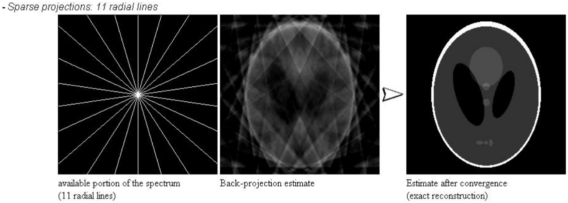
My budeme využívat maticový zápis:

$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix}}_{\mathbf{y}} = \underbrace{\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & & x_{2n} \\ \vdots & & \ddots & \vdots \\ x_{p1} & x_{p2} & \dots & x_{pn} \end{pmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}}_{\boldsymbol{\beta}} + \mathbf{e}. \tag{2}$$

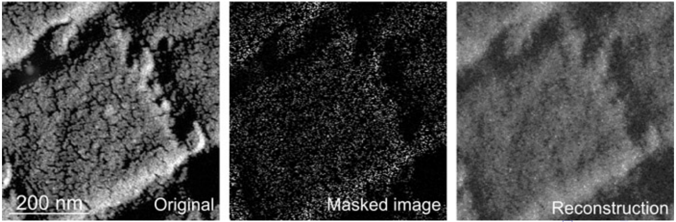
Motivace

K čemu je to dobré?

- kdekoliv kde vzniká špatně podmíněná soustava lineárních rovnic
- rekonstrukce obrazu v tomografii:



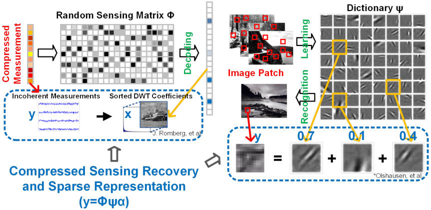
- elektronová mikroskopie:



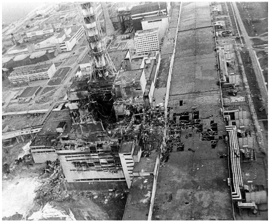
Motivace

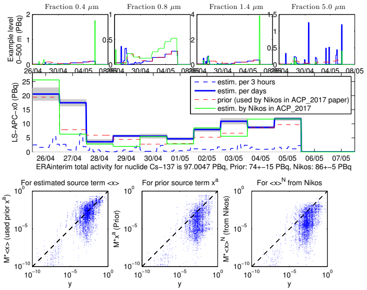
K čemu je to dobré?

- rozpoznání a řídká reprezentace



- určení průběhu úniku látek do prostředí





Zobecněný lineární regresní model

Nejmenší čtverce (OLS)

$$\mathbf{y} = X\boldsymbol{\beta} + \mathbf{e} \tag{3}$$

- minimalizujeme součet čtverců (kvadrátů) odchylek:

$$\sum_j e_j^2 = \mathbf{e}^T\mathbf{e} = (\mathbf{y} - X\boldsymbol{\beta})^T(\mathbf{y} - X\boldsymbol{\beta}) \tag{4}$$

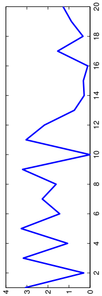
- derivace podle $\boldsymbol{\beta}$:

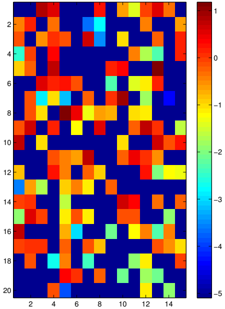
$$2X^T X\boldsymbol{\beta} - 2X^T\mathbf{y} = \mathbf{0} \tag{5}$$

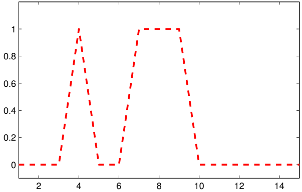
$$\widehat{\boldsymbol{\beta}}_{OLS} = (X^T X)^{-1} X^T \mathbf{y}. \tag{6}$$

Zobecněný lineární regresní model

Cvičný příklad - data

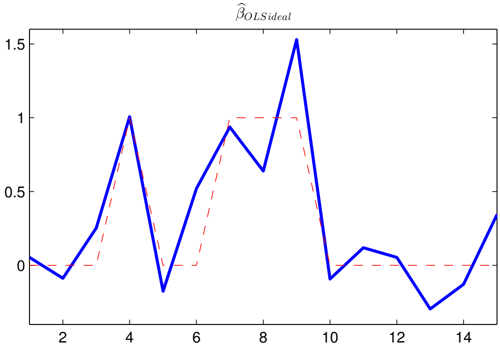






$$\mathbf{y} \quad = \quad X \quad \beta \quad + \text{šum}$$

Zobecněný lineární regresní model
Řešení OLS



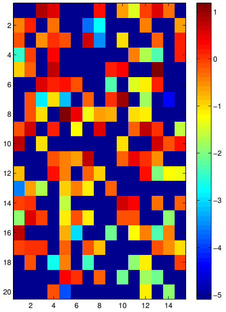
Potud je vše jednoduché a krásné, proč se s tím nespokojíme?

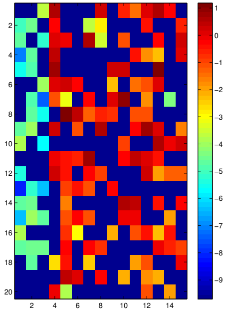
Inverze $(X^T X)^{-1}$.

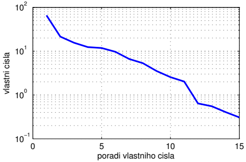
Zobecněný lineární regresní model

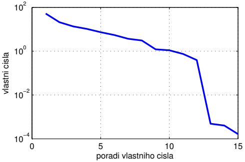
Cvičný příklad

Co když si problém trochu ztížíme (přiblížíme realitě)?



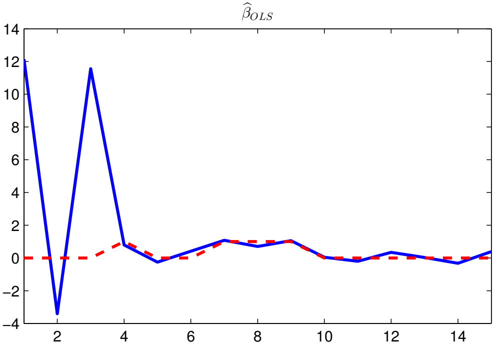






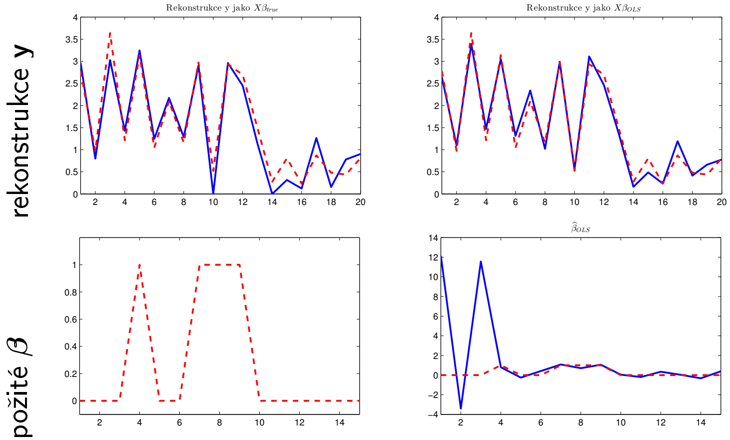
Zobecněný lineární regresní model

Řešení OLS



Regularizace
Principy

- chceme vložit dodatečnou informaci do problému,
- snaha vyhnout se přefitování:



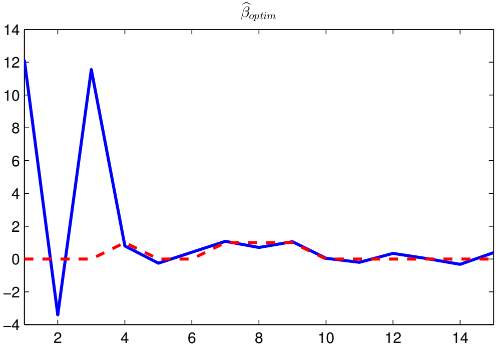
- dodatečná informace příliš slabá x silná,

Optimalizační přístup

Formulace OLS jako optimalizační úlohy

$$\widehat{\boldsymbol{\beta}}_{optim} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - X\boldsymbol{\beta}||_2^2 \tag{7}$$

- $||.||_2$ je Eukleidovská norma,
- fukce $||\mathbf{y} - X\boldsymbol{\beta}||_2^2$ je v $\boldsymbol{\beta}$ (naštěstí) konvexní,
- vede na identické řešení jako OLS



Optimalizace s Tichonovovou regularizací

Formulace

- též “Ridge regression”.
- obecný zápis:

$$\widehat{\boldsymbol{\beta}}_{Tikhonov} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - X\boldsymbol{\beta}||_2^2 + ||\Gamma\boldsymbol{\beta}||_2^2, \tag{8}$$

- ovšem velmi časté použití pro (α skalár, I jednotková matice):

$$\Gamma = \alpha I, \tag{9}$$

- a pozorný student snadno nahlédne řešení ve tvaru:

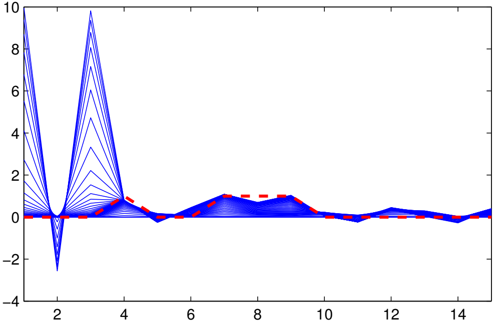
$$\widehat{\boldsymbol{\beta}}_{Tikhonov} = \left(X^T X + \alpha^2 I\right)^{-1} X^T \mathbf{y}. \tag{10}$$

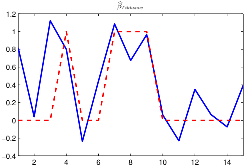
- co to vyřešilo za problém a v čem je potenciální nedostatek?

Optimalizace s Tichonovovou regularizací

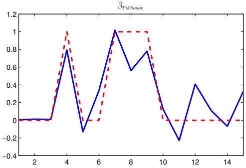
Řešení

Pro Tichonovovův faktor $\alpha = 10^{-3 \text{ až } +3}$:

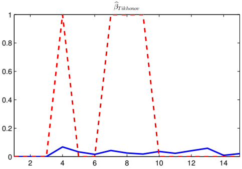




$\alpha = 10^{-2}$



$\alpha = 10^{0}$

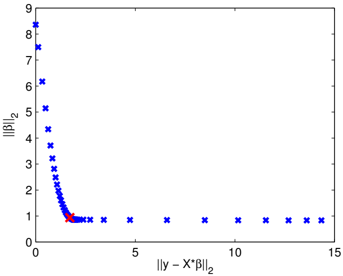


$\alpha = 10^{+2}$

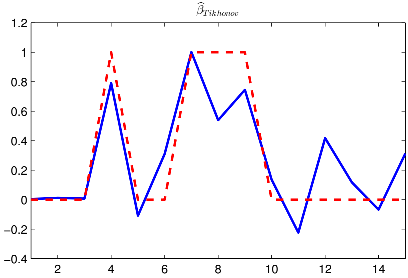
Optimalizace s Tichonovovou regularizací

Volba Tichonovova faktoru

- např. L-křivka



pro $\alpha = 1.2589$



- cross-validace
- ovšem velmi často metoda “kouknu a vidím”...

Optimalizace s LASSO regularizací

Formulace

- LASSO (least absolute shrinkage and selection operator),
- provádí zároveň regularizaci a selekci proměnných,
- zápis problému jako

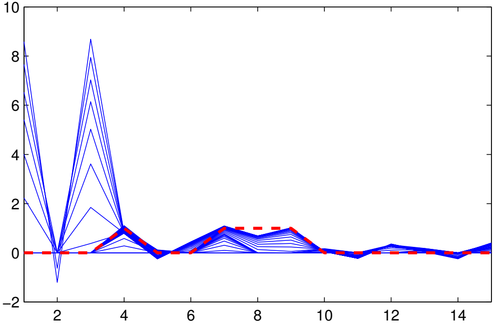
$$\widehat{\beta}_{LASSO} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - X\boldsymbol{\beta}||_2^2 + \alpha||\boldsymbol{\beta}||_1, \tag{11}$$

- kde $||\mathbf{a}||_1 = \sum_j |a_j|$.
- princip...

Optimalizace s LASSO regularizací

Řešení

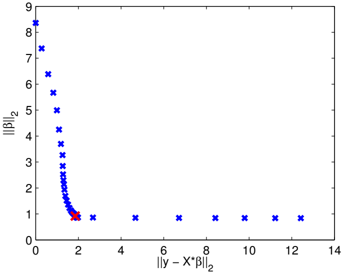
Pro parametr $\alpha = 10^{-3 \text{ až } +3}$:



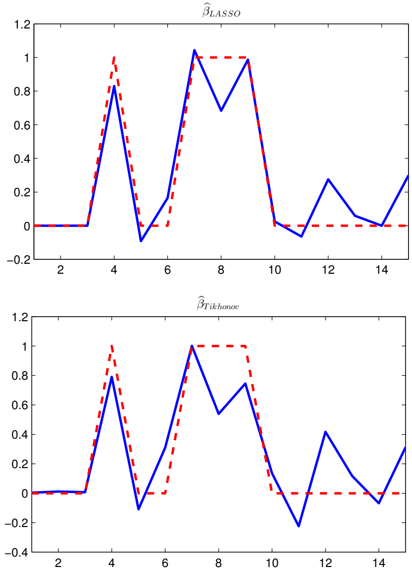
Optimalizace s LASSO regularizací

Řešení

- opět např. L-křivka



pro $\alpha = 0.2512$



Optimalizace a další možnosti

Elastic net

- proč řešení není ideální?
- nic nám nebrání v kombinaci:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - X\boldsymbol{\beta}||_2^2 + \alpha_1||\boldsymbol{\beta}||_2^2 + \alpha_2||\boldsymbol{\beta}||_1, \qquad (12)$$

Bayesovský přístup

Snadno a rychle...

- zvolíme model dat,
- zvolíme apriorní rozdělení parametrů modelu,
- spočítáme odhady aposteriorních rozdělení parametrů.

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

$$f(\theta_1, \theta_2, \ldots, \theta_q) = \prod_{i=1}^{q} f(\theta_i), \tag{13}$$

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

$$\tilde{f}(\theta_i|\mathbf{y}) \propto \exp\left[\mathrm{E}_{\tilde{f}(\theta_{\setminus i}|\mathbf{y})}\left(\ln f(\theta_1, \theta_2, \ldots, \theta_q, \mathbf{y})\right)\right], \quad \forall i. \tag{14}$$

VB: příklad skalární dekompozice

Skalární dekompozice

- skalární model:

$$d = ax + e, \;\; e \sim \mathcal{N}(0, r_e), \tag{15}$$

- tzn. model dat

$$f(d|a, x) = \mathcal{N}(ax, r_e), \tag{16}$$

- volba apriorních rozdělení pro parametry a a x:

$$f(a) = \mathcal{N}(0, r_a), \tag{17}$$

$$f(x) = \mathcal{N}(0, r_x). \tag{18}$$

- r_e, r_a, r_x jsou nyní konstanty (+ data d), $\widehat{a}$ a $\widehat{x}$ chceme odhadnout.

VB: příklad skalární dekompozice

Skalární dekompozice

- vyjádříme si logaritmus sdružené pravděpodobnosti $\ln f(\theta_1, \ldots, \theta_q, \mathbf{y})$

$$\ln f(a, x, d) \propto -\frac{1}{2}\left(\frac{(ax-d)^2}{r_e} + \frac{a^2}{r_a} + \frac{x^2}{r_x}\right), \qquad (19)$$

- a prostým dosazením do $\tilde{f}(\theta_i|\mathbf{y}) \propto \exp\left[\mathrm{E}_{\tilde{f}(\theta_{\setminus i}|\mathbf{y})}(\ln f(\theta_1, \theta_2, \ldots, \theta_q, \mathbf{y}))\right]$ dostáváme

$$\tilde{f}(a|d) \propto \exp\left[-\frac{1}{2}a^2\left(\widehat{x^2}r_e^{-1} + r_a^{-1}\right) - a\left(d\widehat{x}r_e^{-1}\right)\right], \qquad (20)$$

$$\tilde{f}(x|d) \propto \exp\left[-\frac{1}{2}x^2\left(\widehat{a^2}r_e^{-1} + r_x^{-1}\right) - x\left(d\widehat{a}r_e^{-1}\right)\right], \qquad (21)$$

- a všimneme si, že to je opět tvar normálního rozdělení:

$$\tilde{f}(a|d) = \mathcal{N}_a(\mu_a, \sigma_a), \qquad \tilde{f}(x|d) = \mathcal{N}_x(\mu_x, \sigma_x). \qquad (22)$$

VB: příklad skalární dekompozice

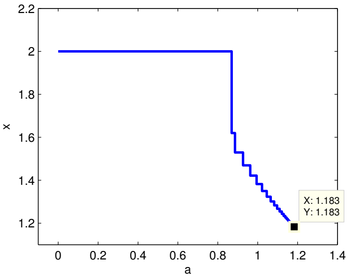
Skalární dekompozice

- dopočítáme tzv. tvarovací parametry $\mu_a, \sigma_a, \mu_x, \sigma_x$:

$$\sigma_a = \left(\widehat{x^2} r_e^{-1} + r_a^{-1}\right)^{-1}, \qquad \mu_a = \sigma_a d \widehat{x} r_e^{-1}, \tag{23}$$

$$\sigma_x = \left(\widehat{a^2} r_e^{-1} + r_x^{-1}\right)^{-1}, \qquad \mu_x = \sigma_x d \widehat{a} r_e^{-1}. \tag{24}$$

- a to je (skoro) vše...



- připomenutí: máme rovnici

$$\mathbf{y} = X\boldsymbol{\beta} + \mathbf{e}, \tag{25}$$

- předpokládáme rozdělení chyb ve tvaru

$$e_i \sim \mathcal{N}\left(0, \omega^{-1}\right), \ \forall i, \tag{26}$$

- tím dostáváme přirozeně model dat jako

$$f(\mathbf{y}|\boldsymbol{\beta}, \omega) = \mathcal{N}\left(X\boldsymbol{\beta}, \omega^{-1} I_p\right), \tag{27}$$

- zajímá nás: $\boldsymbol{\beta}, \omega$.

Bayesovská formulace lineární regrese

Apriorno pro parametry

- bayesovský pohled: $\boldsymbol{\beta}$ a ω jsou pro nás neznámé parametry a budeme je odhadovat,
- volba apriorních rozdělení pro $\boldsymbol{\beta}$ a ω,
- my si pro začátek vystačíme s nejjednodušší možností:

$$f(\boldsymbol{\beta}) = \mathcal{N}(\mathbf{0}, I_n) = \underbrace{(2\pi)^{-\frac{n}{2}}}_{coef} \underbrace{|I_n|^{-\frac{1}{2}}}_{coef} \exp\left(-\frac{1}{2}\boldsymbol{\beta}' I_n^{-1} \boldsymbol{\beta}\right), \quad (28)$$

$$f(\omega) = \mathcal{G}(c_0, d_0) = \underbrace{\frac{c_0^{d_0}}{\Gamma(c_0)}}_{coef} \omega^{c_0 - 1} \exp(-d_0 \omega). \quad (29)$$

- naším cílem je určit $\widehat{\boldsymbol{\beta}} = \mathrm{E}[\boldsymbol{\beta}]$ a $\widehat{\omega} = \mathrm{E}[\omega]$.

Bayesovská formulace lineární regrese

VB aproximace

- aposteriorno ve tvaru $\tilde{f}(\boldsymbol{\beta}) = \mathcal{N}(\mu_{\boldsymbol{\beta}}, \Sigma_{\boldsymbol{\beta}})$ a $\tilde{f}(\omega) = \mathcal{G}(c, d)$ s tvarovacími parametry:

$$\Sigma_{\boldsymbol{\beta}} = \left(\widehat{\omega} X^T X + I_n\right)^{-1} \tag{30}$$

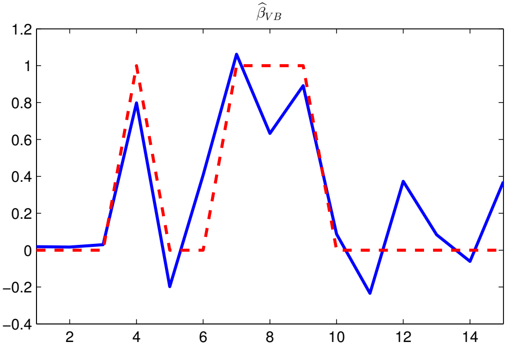
$$\mu_{\boldsymbol{\beta}} = \Sigma_{\boldsymbol{\beta}} \widehat{\omega} X^T \mathbf{y} = \left(\widehat{\omega} X^T X + I_n\right)^{-1} \widehat{\omega} X^T \mathbf{y} \tag{31}$$

$$c = c_0 + \frac{p}{2} \tag{32}$$

$$d = d_0 + \frac{1}{2}\left(\mathbf{y}\mathbf{y}^T - 2\mathbf{y}^T X \widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\beta}^T X^T X \boldsymbol{\beta}}\right) \tag{33}$$

Bayesovská formulace lineární regrese

Řešení



Bayesovská ridge regression

Formulace

- model dat a šumu zachováme,

$$f(\mathbf{y}|\beta,\omega) = \mathcal{N}\left(X\beta,\omega^{-1}I_p\right), \tag{34}$$
$$f(\omega) = \mathcal{G}\left(c_0, d_0\right). \tag{35}$$

- do modelu β zavedeme hyperparametr $v \in \mathbf{R}$ následovně:

$$f(\beta) = \mathcal{N}\left(\mathbf{0}, I_n\right) \quad \dashrightarrow \quad f(\beta|v) = \mathcal{N}\left(\mathbf{0}, v^{-1}I_n\right) \tag{36}$$

- v se stává parametrem modelu, zvolíme jeho apriorní rozdělení

$$f(v) = \mathcal{G}\left(a_0, b_0\right). \tag{37}$$

Bayesovská ridge regression

VB aproximace

- aposteriorno $\tilde{f}(\omega)$ zůstane stejné jako v předchozím případě.
- aposteriorno $\tilde{f}(\boldsymbol{\beta}|v) = \mathcal{N}(\mu_{\boldsymbol{\beta}}, \Sigma_{\boldsymbol{\beta}})$ a $\tilde{f}(v) = \mathcal{G}(a, b)$ má (po odvození) následující tvarovací parametry:

$$\Sigma_{\boldsymbol{\beta}} = \left(\widehat{\omega} X^T X + \widehat{v} I_n\right)^{-1} \tag{38}$$

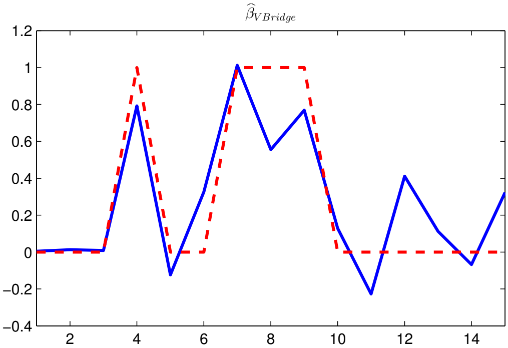
$$\mu_{\boldsymbol{\beta}} = \Sigma_{\boldsymbol{\beta}} \widehat{\omega} X^T \mathbf{y} = \left(\widehat{\omega} X^T X + \widehat{v} I_n\right)^{-1} \widehat{\omega} X^T \mathbf{y} \tag{39}$$

$$a = a_0 + \frac{n}{2} \tag{40}$$

$$b = b_0 + \frac{1}{2} \operatorname{trace}\left(\widehat{\boldsymbol{\beta}\boldsymbol{\beta}'}\right) \tag{41}$$

Bayesovská ridge regression

Řešení



- přílišná jednoduchost modelu

Bayesovská sparse regression

Formulace

- též “Sparse bayesian learning” nebo “Relevance vector machine”.
- rozptyl prvků vektoru $\boldsymbol{\beta}$ nebudeme modelovat pouze jedním parametrem, ale přiřadíme parametr každému prvku $\boldsymbol{\beta}$:

$$f(\boldsymbol{\beta}|\mathbf{v}) = \mathcal{N}\left(\mathbf{0}, \mathrm{diag}\left(\mathbf{v}\right)^{-1}\right) = \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} v_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & v_n \end{pmatrix}^{-1}\right), \tag{42}$$

- model pro prvky v_j volíme obdobně jako u ridge regression:

$$f(v_j) = \mathcal{G}\left(a_0, b_0\right), \ \forall j. \tag{43}$$

Bayesovská sparse regression

VB aproximace

- aposteriorno $\tilde{f}(\beta|v) = \mathcal{N}(\mu_{\boldsymbol{\beta}}, \Sigma_{\boldsymbol{\beta}})$ a $\tilde{f}(v_j) = \mathcal{G}(a_j, b_j), \ \forall j$, má (po odvození) následující tvarovací parametry:

$$\Sigma_{\boldsymbol{\beta}} = \left(\widehat{\omega} X^T X + \operatorname{diag}(\widehat{\mathbf{v}})\right)^{-1} \tag{44}$$

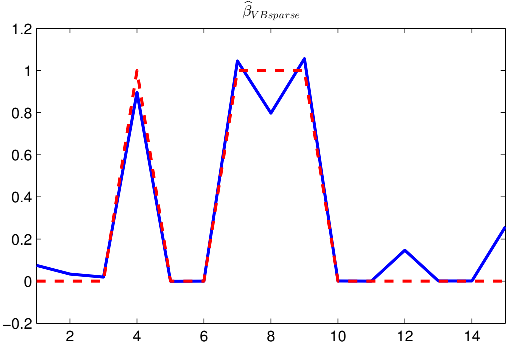
$$\mu_{\boldsymbol{\beta}} = \Sigma_{\boldsymbol{\beta}} \widehat{\omega} X^T \mathbf{y} = \left(\widehat{\omega} X^T X + \operatorname{diag}(\widehat{\mathbf{v}})\right)^{-1} \widehat{\omega} X^T \mathbf{y} \tag{45}$$

$$a_j = a_0 + \frac{1}{2} \tag{46}$$

$$b_j = b_0 + \frac{1}{2} \left(\widehat{\boldsymbol{\beta}\boldsymbol{\beta}'}\right)_{j,j} \tag{47}$$

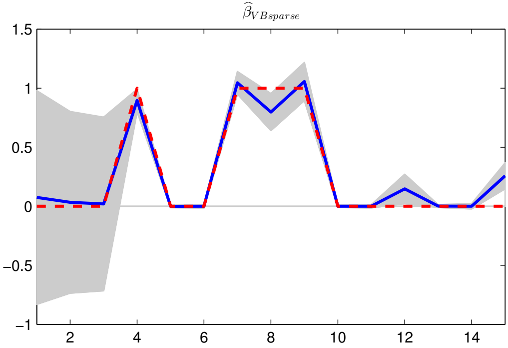
Bayesovská sparse regression

Řešení



Výhody bayesovského přístupu

- kvantifikace neurčitosti



- potlačení "ladicích" knoflíků, i když...
- nepřeberné modelovací možnosti

Dodatek

Co jsme neřešili a možná bychom měli...

- Pracovali jsme pouze s nagenerovaným příkladem bez fyzikální podstaty
- Co jsme vynechali?
  - pozitivita
  - vztah jednotlivých prvků $\boldsymbol{\beta}$
  - informace o měřeních $\mathbf{y}$
  - ...

Co nás čeká příště

Aplikace zobecněných lineárních modelů