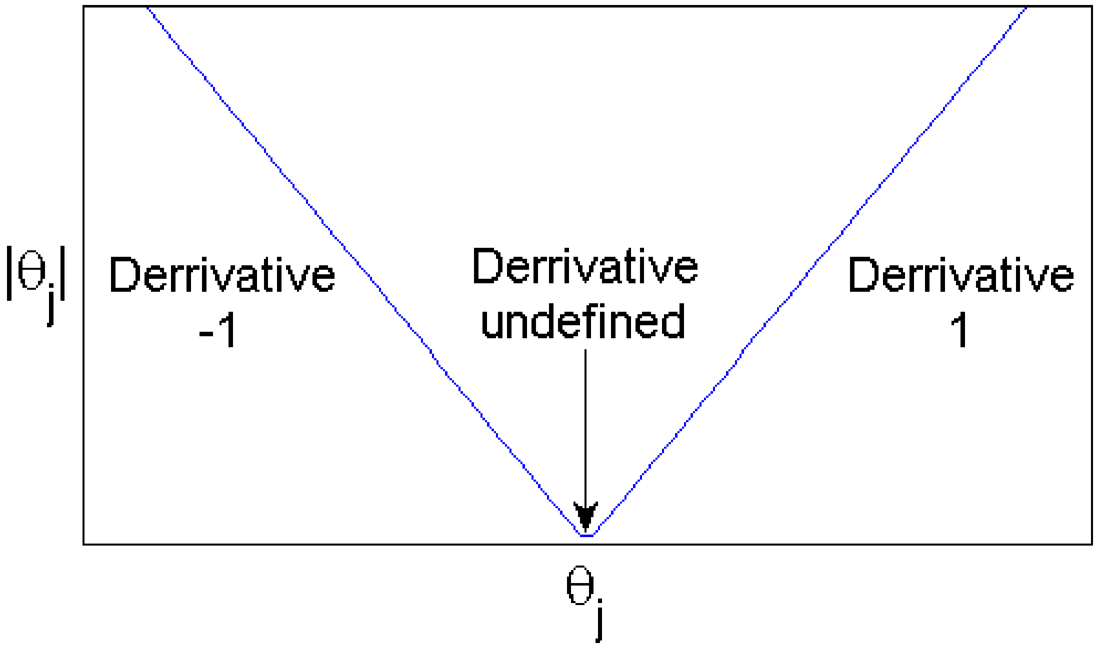
Lasso regression gradient descent

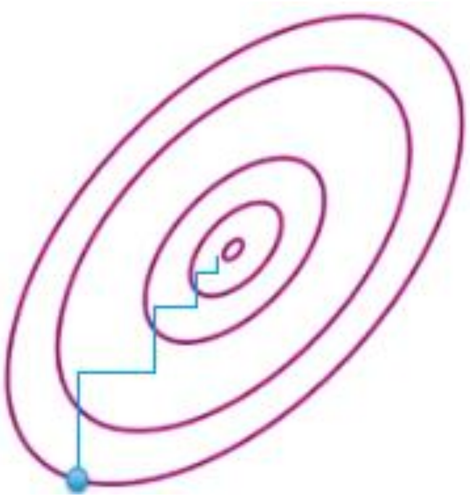
- Nedostatak lasso: ciljna funkcija nije diferencijabilna



- Closed-form solution ne postoji

Coordinate descent

- Cilj: minimizacija funkcije $J(\theta) = J(\theta_0, \theta_1, \dots, \theta_d)$
- Intuicija:
  - često je teško pronaći minimum za sve koordinate istovremeno
  - ali je jednostavno pronaći minimum za pojedinačnu koordinatu ako držimo sve ostale koordinate fiksirane



Coordinate descent

Ulaz	• $J(\theta)$ – funkcija koja se optimizuje • $\theta_{\mathbf{0}}$ – početno rešenje • *maxIters* – maksimalan broj iteracija
Postupak	for *t* = 1, 2,…, *maxIters*: Odabrati koordinatu *j* $\theta_j^{(t+1)} = \min_{\theta} g(\theta_0^{(t)}, \dots, \theta_{j-1}^{(t)}, \theta, \theta_{j+1}^{(t)}, \dots, \theta_d^{(t)})$
Izlaz	θ (tačka u kojoj funkcija $J(\theta)$ ima minimum)

Coordinate descent

- Normalizovaćemo obeležja

$$f_j\left(x^{(k)}\right) \rightarrow \frac{f_j\left(x^{(k)}\right)}{\sqrt{\sum_{i=1}^{N} f_j(x^{(k)})^2}}$$

- Kako da odaberemo sledeću koordinatu?
  - Na slučajan način (*random/stohastic gradient descent*)
  - Naizmenično (*round robin*)
  - ...

- Nema koraka α koji bismo morali podešavati

- Konvergira do globalnog optimuma za lasso regresiju

Coordinate descent za RSS

$$RSS(\theta) = \sum_{i=1}^{N} \left(y^{(i)} - \sum_{j=0}^{d} \theta_j f_j\left(x^{(i)}\right) \right)^2$$

- Minimizovaćemo RSS po θ_j:

$$\frac{\partial}{\partial \theta_j} RSS(\theta) = -2 \sum_{i=1}^{N} f_j\left(x^{(i)}\right) \left(y^{(i)} - \sum_{j=0}^{d} \theta_j f_j\left(x^{(i)}\right) \right)$$

$$= -2 \sum_{i=1}^{N} f_j\left(x^{(i)}\right) \left(y^{(i)} - \sum_{k \neq j} \theta_k f_k\left(x^{(i)}\right) - \theta_j f_j\left(x^{(i)}\right) \right)$$

$$= -2 \sum_{i=1}^{N} f_j\left(x^{(i)}\right) \left(y^{(i)} - \sum_{k \neq j} \theta_k f_k\left(x^{(i)}\right) \right) + 2\theta_j \sum_{i=1}^{N} f_j\left(x^{(i)}\right)^2$$

Coordinate descent za RSS

$$\frac{\partial}{\partial \theta_j} RSS(\theta) =$$

Predikcija *i*-te opservacije ukoliko bi isključili *j*-to obeležje iz modela

$$-2\underbrace{\boxed{\sum_{i=1}^{N} f_j\left(x^{(i)}\right)\left(y^{(i)} - \sum_{k \neq j} \theta_k f_k\left(x^{(i)}\right)\right)}}_{\rho_j} + 2\theta_j \sum_{i=1}^{N} f_j\left(x^{(i)}\right)^2$$

= 1

(normalizovali smo obeležja

$$f_j\left(x^{(k)}\right) \to \frac{f_j\left(x^{(k)}\right)}{\sqrt{\sum_{i=1}^{N} f_j\left(x^{(k)}\right)^2}})$$

$$\frac{\partial}{\partial \theta_j} RSS(\theta) = -2\rho_j + 2\theta_j$$

Coordinate descent za RSS

$$\frac{\partial}{\partial \theta_j} RSS(\theta) = -2\rho_j + 2\theta_j = 0 \Longrightarrow \theta_j = \rho_j$$

Predikcija bez obeležja *j*

$$\rho_j = \sum_{i=1}^{N} f_j\left(x^{(i)}\right) \boxed{\left(y^{(i)} - \sum_{k \neq j} \theta_k f_k\left(x^{(i)}\right)\right)}$$

Rezidual bez obeležja *j*

korelacija *j*-tog obeležja i reziduala bez obeležja *j*

Ako su obeležje *j* i predikcija bez *j*-tog obeležja u korelaciji ρ_j je veliko, a stoga i θ_j (obeležje j je važno za model)

Coordinate descent za RSS

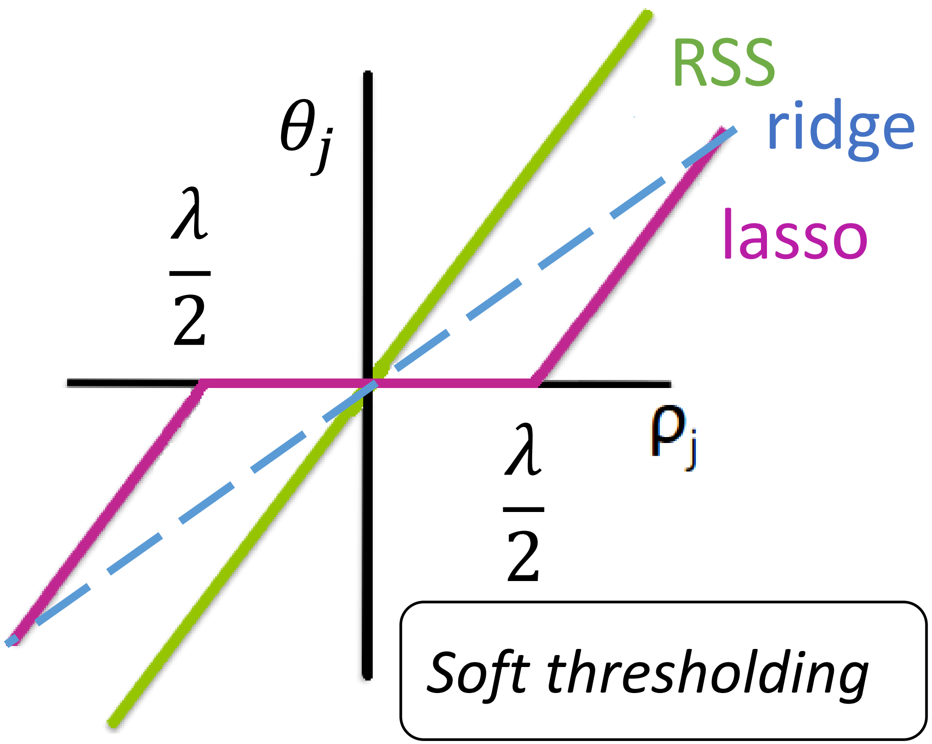
Ulaz	• $g(\theta)$ – funkcija koja se optimizuje • $\theta_{\mathbf{0}}$ – početno rešenje • *maxIters* – maksimalan broj iteracija
Postupak	for *t* = 1, 2,…, maxIters: for *j*=0,1,…,*d* (round robin) $\theta_j^{(t+1)} = \rho_j$
Izlaz	θ – tačka u kojoj funkcija $J(\theta)$ ima minimum

Coordinate descent za lasso

$$\theta_j^{(t+1)} = \begin{cases} \rho_j + \lambda/2 & ako\ je\ \rho_j < \lambda/2 \\ 0 & ako\ \rho_j \epsilon [-\lambda/2, \lambda/2] \\ \rho_j - \lambda/2 & ako\ je\ \rho_j > \lambda/2 \end{cases}$$

Mala korelacija:
postavi težinu na 0

Velika (+ ili -) korelacija:
Ostavi obeležje u modelu, ali mu smanji težinu za $\lambda/2$



Soft thresholding

λ definiše šta znači jako/slabo

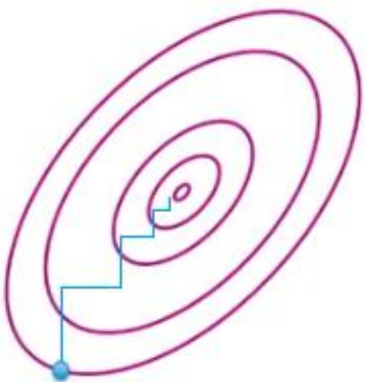
Normalizovana obeležja

Coordinate descent konvergencija

- Kada da stanemo (kako da znamo da je rešenje konvergiralo)?

- Za konveksne probleme, koraci će da budu sve manji i manji

- Merimo veličinu koraka (za sve koordinate) i stanemo kada je maksimalan korak $< \varepsilon$



Drugi lasso solveri

- Klasično: Least Angle Regression and Shrinkage (LARS)
- Kasnije: Coordinate descent
- Danas:
  - Parallel Coordinate Descent
  - Parallel stochastic gradient descent (SGD) [Niu et al. ‘11]
  - Parallel independent solutions then averaging [Zhang et al. ‘12]
  - Alternating directions method of multipliers (ADMM) [Boyd et al. ‘11]

Lasso: odabir λ

- Isto kao i kod *ridge* regresije (training/validation/test ili unakrsna validacija)

- Na ovaj način dobićemo λ koji nam daje najbolju prediktivnu tačnost

- Međutim, ovo znači da će λ biti malo manje nego što je optimalno za selekciju modela

- Postoje drugi načini za odabir λ koji rešavaju ovaj problem
  Machine Learning: A Probabilistic Perspective', Murphy, 2012

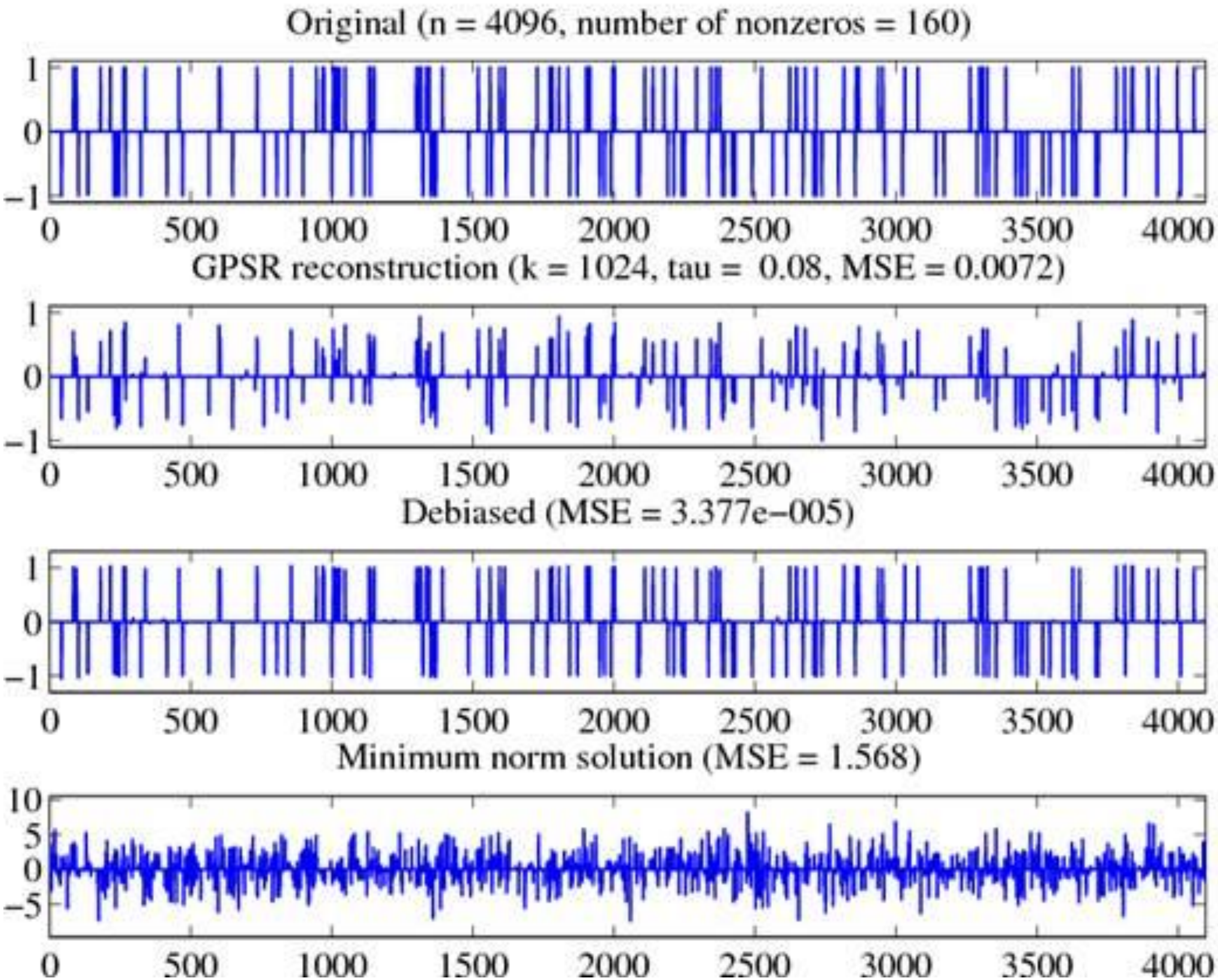
Debiasing lasso

- *Lasso* smanjuje koeficijente u odnosu na OLS
  - Dobijamo model sa većim sistematskim odstupanjem i manjom varijansom

- Sistematsko odstupanje možemo smanjiti na sledeći način (*debiasing the lasso solution*):
  1. Iskoristiti *lasso* za selekciju obeležja
  2. Primeniti OLS koristeći isključivo selektovanim obeležjima

Debiasing lasso

http://www.lx.it.pt/~mtf/GPSR/



Lasso – stabilnost modela

- Ako imamo grupu jako koreliranih obeležja *lasso* će odabrati jedno na proizvoljan način

- Preciznije, *lasso* će odabrati obeležje koje je više korelirano sa y

- Ali, koje obeležje je više korelirano sa y može da zavisi od šuma

Primer: Lasso – stabilnost modela

$$x_1 = x_2 = x_3$$

- Dodajmo šum na svako obeležje (obeležja nisu baš identična ali snažno korelirana)

$$y = x_1 + x_2 + x_3$$

- Pokretaćemo primer više puta, pri čemu će se svaki put obučavajući skup razlikovati zbog šuma dodatog na obeležja x_1, x_2 i x_3

Primer je preuzet sa: http://blog.datadive.net/selecting-good-features-part-ii-linear-models-and-regularization/

Primer: Lasso – stabilnost modela

```
Random seed 3
No regularization: 2.677 * X0 + 0.4 * X1 + -0.146 * X2
Ridge model: 0.999 * X0 + 0.9 * X1 + 0.911 * X2
Lasso model: 2.751 * X0 + 0.0 * X1 + 0.0 * X2

Random seed 10
No regularization: 0.497 * X0 + 2.596 * X1 + -0.204 * X2
Ridge model: 0.901 * X0 + 0.956 * X1 + 0.861 * X2
Lasso model: 0.475 * X0 + 2.226 * X1 + 0.0 * X2

Random seed 11
No regularization: -1.428 * X0 + 0.625 * X1 + 3.638 * X2
Ridge model: 0.771 * X0 + 0.884 * X1 + 1.054 * X2
Lasso model: 0.0 * X0 + 0.0 * X1 + 2.667 * X2
```

Selekcija obeležja

- Kod selekcije obeležja budite jako pažljivi oko interpretacije selektovanih obeležja

- Obeležja koja smo selektovali su uvek samo u kontekstu inicijalnih obeležja koja su postojala

- Selekcija je osetljiva na korelirana obeležja

- Skup selektovanih obeležja zavisi od algoritma koji je korišćen