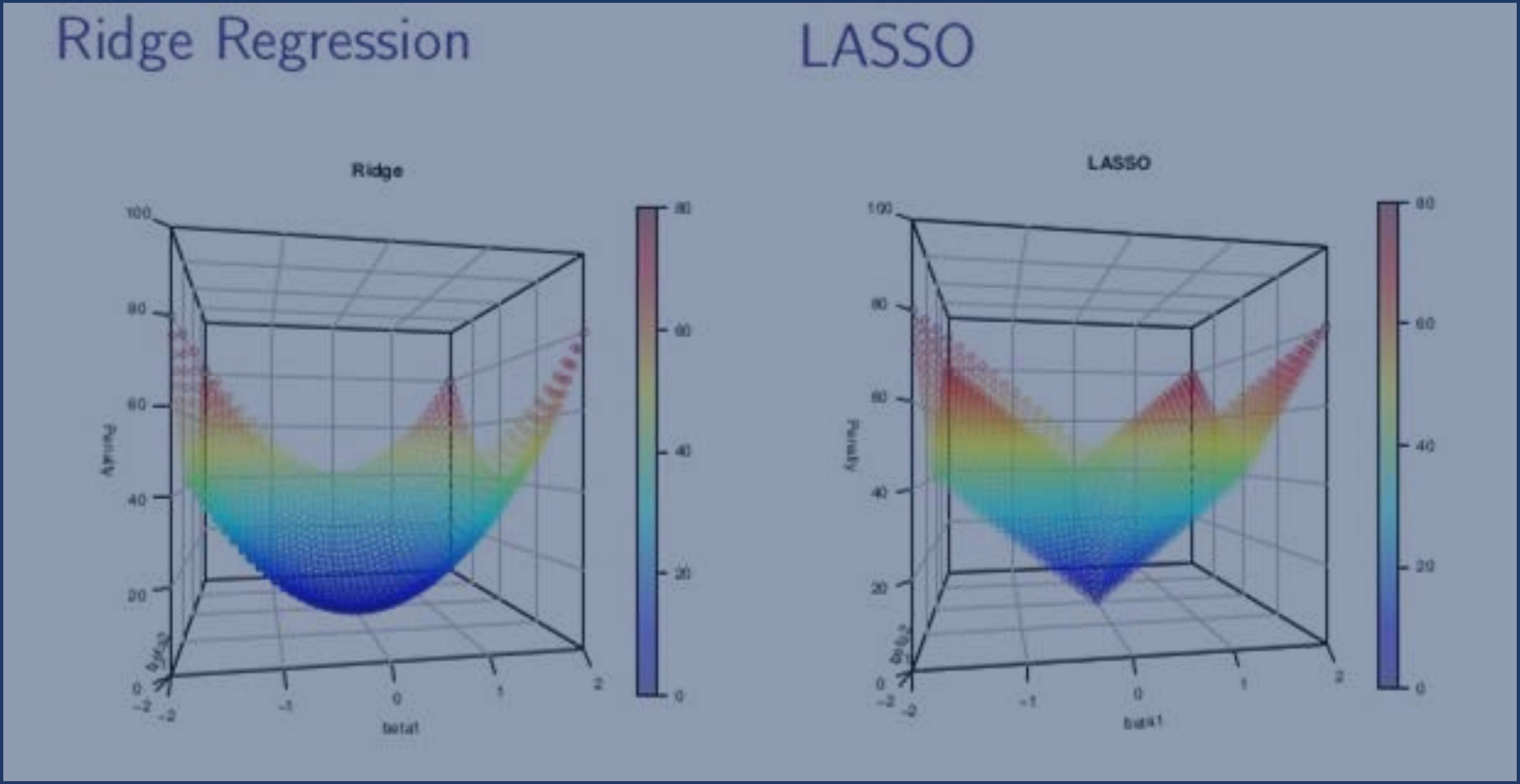
Lasso regularizacija



Motivacija: selekcija obeležja

- Do sada smo videli da možemo uvrstiti mnogo različitih obeležja u naš model
  - sanitarije, vakcinacija, BDP, pristup lekovima,...

- Kako odabrati obeležja koja su važna za model?
  - Najvažnija obeležja treba da uz sebe imaju najveće koeficijente θ
  - Obeleža koja nisu u korelaciji sa y treba da se može sa 0

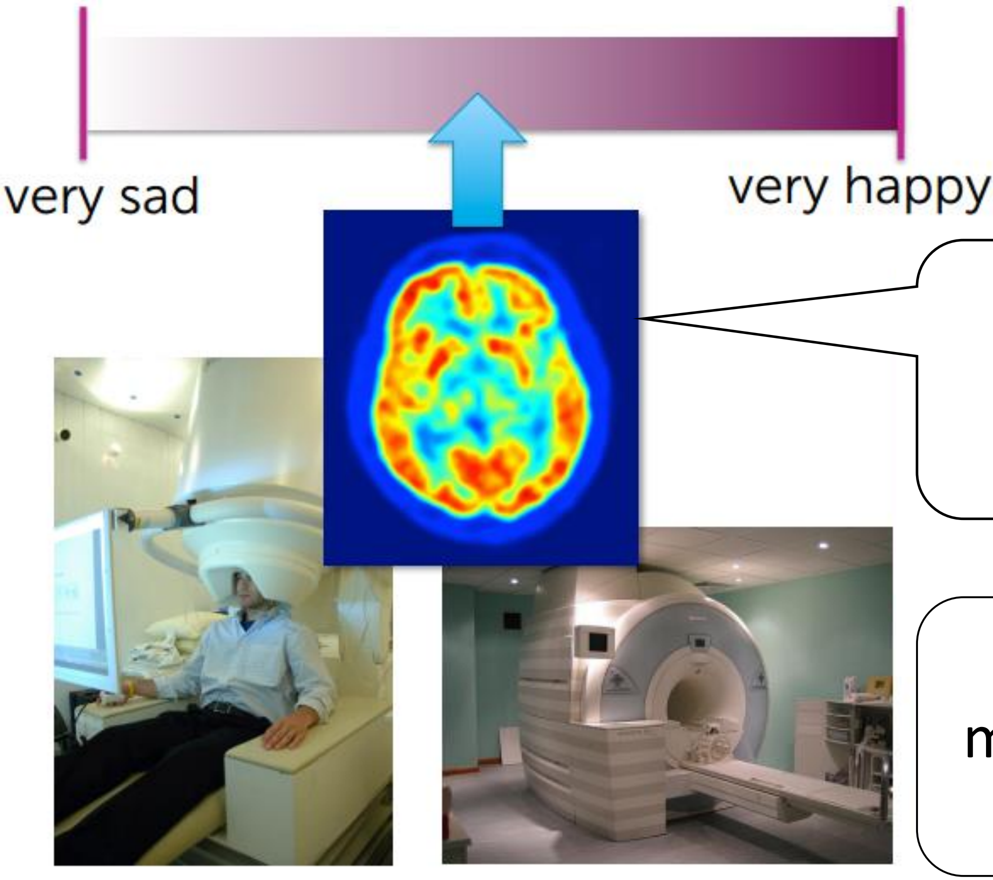
Zašto vršiti selekciju obeležja?

- Efikasnost
  - Ako imamo puno obeležja (npr. 10^{12}), predikcija je veoma računarski zahtevna (imamo mnogo operacija množenja)
  - Ako je θ *sparse* (sadrži puno nula) onda ovo ne mora biti veliki problem

- Interpretabilnost: koja obeležja su relevantna za predikciju?

Primer – čitanje misli

Model koji predviđa da li je osoba tužna ili srećna tako što prikazujemo reči ili slike i očitavamo aktivnost mozga



Aktivnost mozga možemo predstaviti kao sliku: Intenziteti piksela

Želimo da nađemo regione u mozgu koji su relevantni za ovaj zadatak - interpretabilnost

Kako selektovati obeležja?

- Opcija: za svaku kombinaciju obeležja izračunati performanse rezultujućeg modela i odbrati najbolji
  - Kompleksnost: 2^{D+1}
  - Postoje efikasnije alternative

- Ovde ćemo koristiti drugi pristup: selekciju obeležja ćemo izvršiti pomoću regularizacije

Možemo li koristiti *ridge*?

$$J(\theta) = \frac{1}{2N}\sum_{i=1}^{N}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \frac{\lambda}{2}\sum_{i=1}^{d}\theta_j^2$$

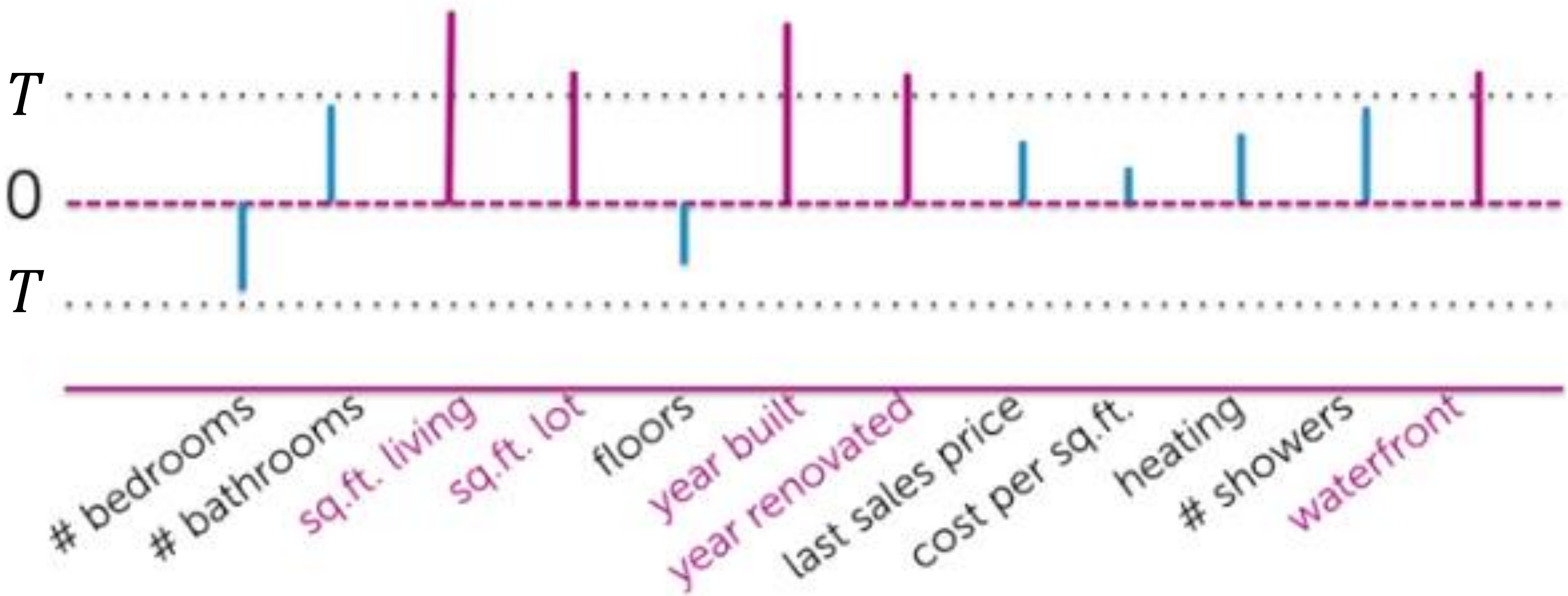
- Za $\lambda \to \infty$ svi parametri θ teže 0, ali nećemo imati situaciju da su neki od njih tačno 0 dok su drugi različiti od 0

$$\theta_j^{(t+1)} = \theta_j^{(t)}(1 - \alpha\lambda) - \frac{\alpha}{N}\sum_{i=1}^{N}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right){x_0}^{(i)}$$

- Međutim, u kontekstu selekcije obeležja, mi želimo da nerelevantnim obeležjima dodelimo tačno 0 (da ih izbacimo iz modela)

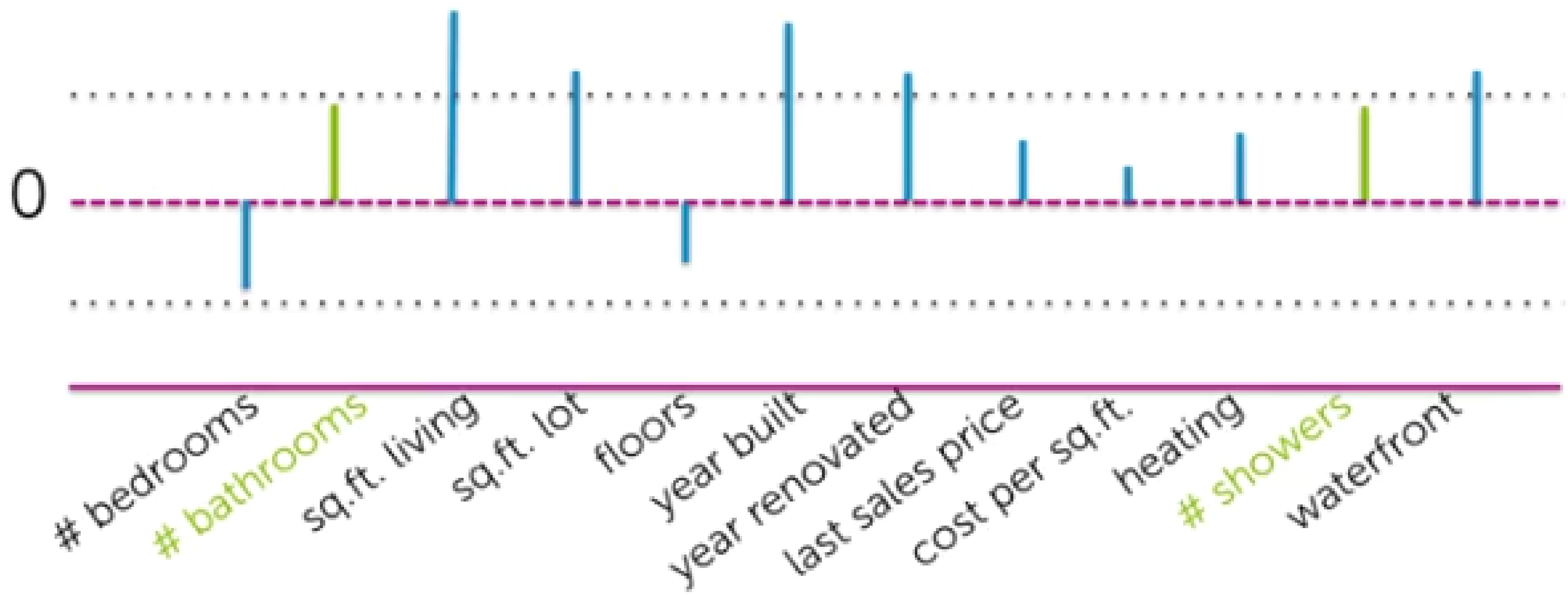
Možemo li koristiti *ridge*?

Možemo li zadati prag T i eliminisati sva obeležja sa $\theta < T$?



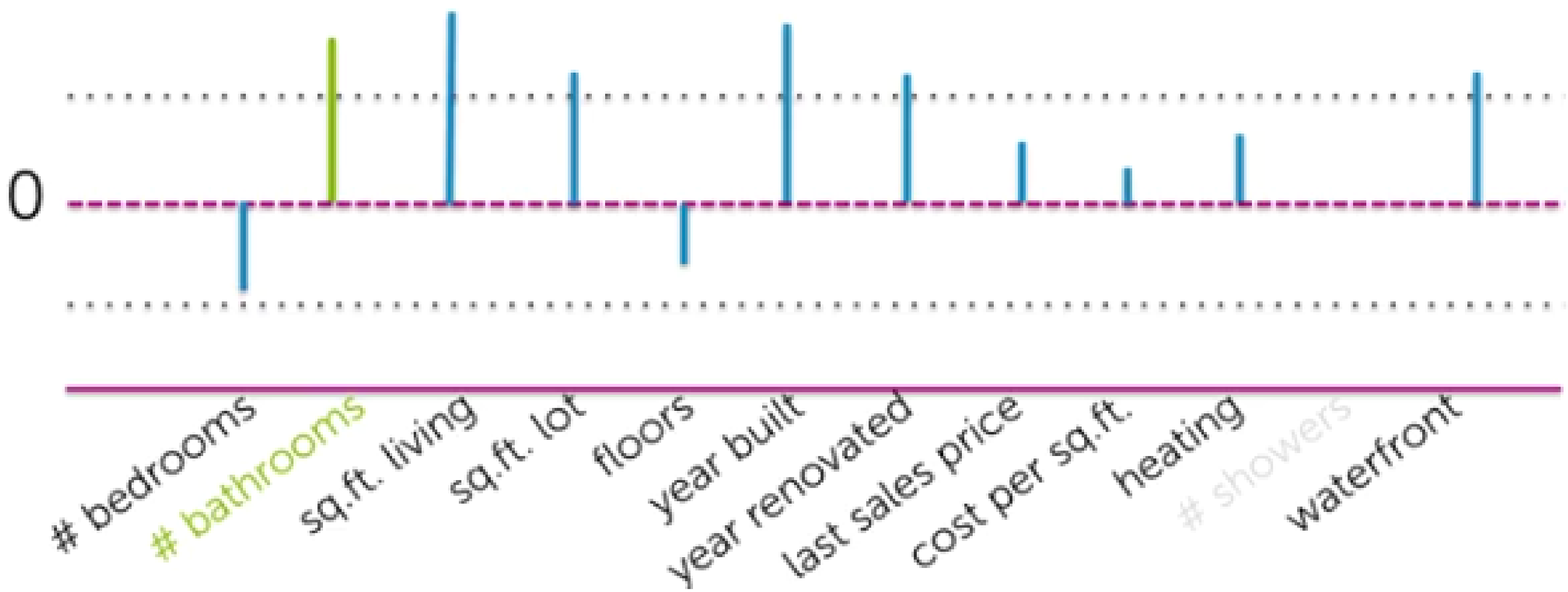
Možemo li koristiti *ridge*?

Možemo li zadati prag T i eliminisati sva obeležja sa $\theta < T$?



Možemo li koristiti *ridge*?

Možemo li zadati prag T i eliminisati sva obeležja sa $\theta < T$?



Možemo li koristiti *ridge*?

- Recimo da je naš model:

$$\theta_1 x_1 + \theta_2 x_2$$

- Recimo da je $x_1 = x_2$
- Model je u tom slučaju

$$(\theta_1 + \theta_2) x_1$$

- Odnosno, da smo inicijalno izbacili x_2 iz modela, težina ispred x_1 bi bila duplo veća i ovo obeležje bi bilo zadržano u modelu

Možemo li koristiti *ridge*?

- Recimo da imamo skup koreliranih obeležja

- Uzmimo u obzir dva modela:
  1) Jednako rasporediti (male) težine na sva korelirana obeležja
  2) Jednom obeležju dodeliti (veliku) težinu, a težine ostalih postaviti na 0

- Iako su ova dva modela slična (u smislu konačne funkcije koju dobijamo), *ridge* će preferirati rešenje 1)

- U funkciji greške figuriše θ^2 – jedan veliki θ koeficijent bi jako uvećao regularizacioni deo greške

Možemo li koristiti *ridge*?

- Ako bismo koristili prag da odsečemo neka obeležja eliminisali bismo sva korelirana obeležja

- Iako je jedno od njih (ili ceo skup zajedno) relevantno za zadatak predikcije

- Dakle, *rigde* ne možemo koristiti za selekciju obeležja

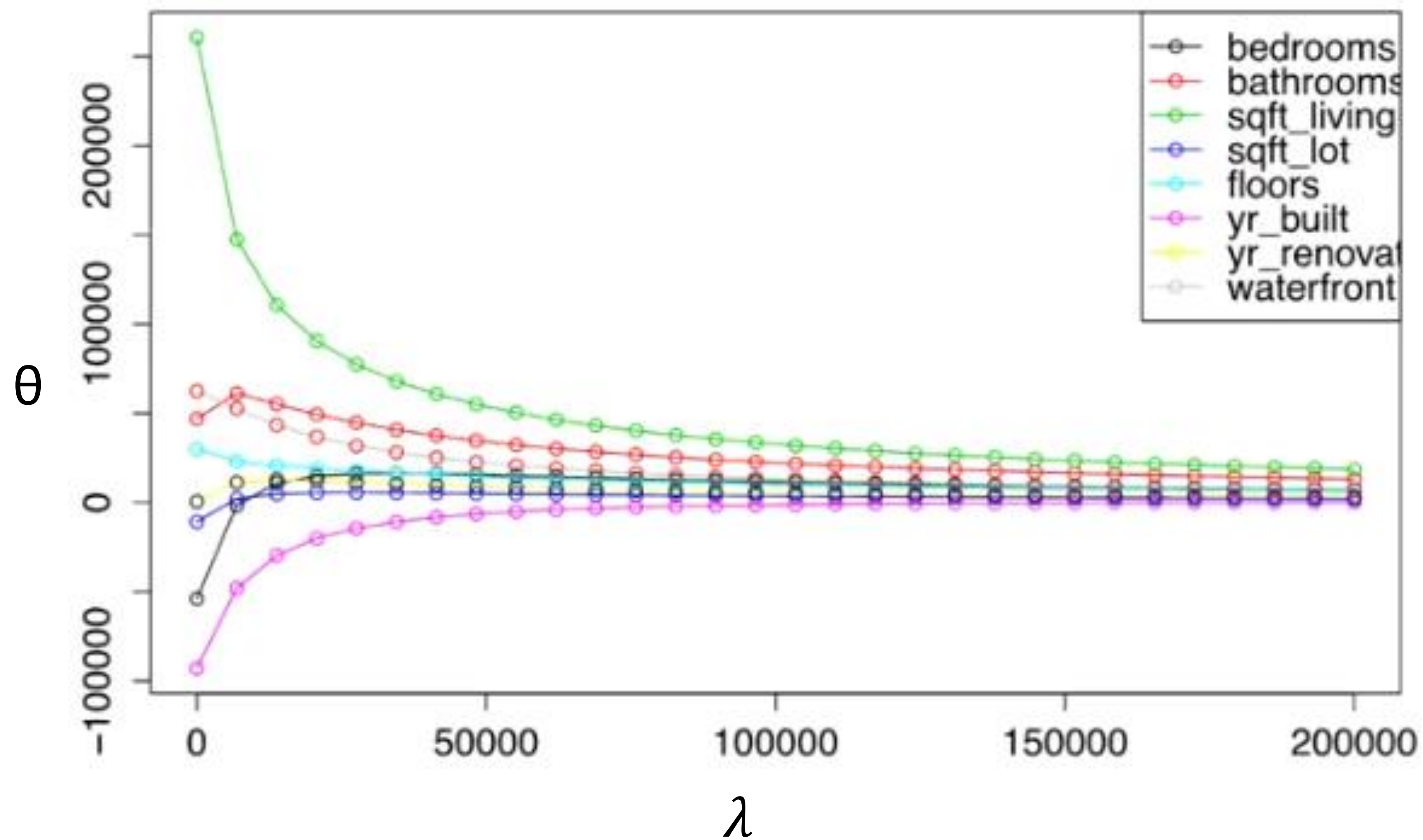
Lasso regresija (L$_1$ regularizacija)

$$J(\theta) = \frac{1}{2N}\sum_{i=1}^{N}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \lambda\sum_{i=1}^{d}\left|\theta_j\right|$$

- Ovo će da dovede do *sparse* modela (neki θ koeficijenti biće postavljeni tačno na 0)
- λ: koliko favorizujemo *sparsity* naspram prilagođavanja podacima
- $\lambda = 0$, vraćamo se na prethodni OLS (bez regularizacije)
- za $\lambda = \infty$, svi koeficijenti θ će težiti 0

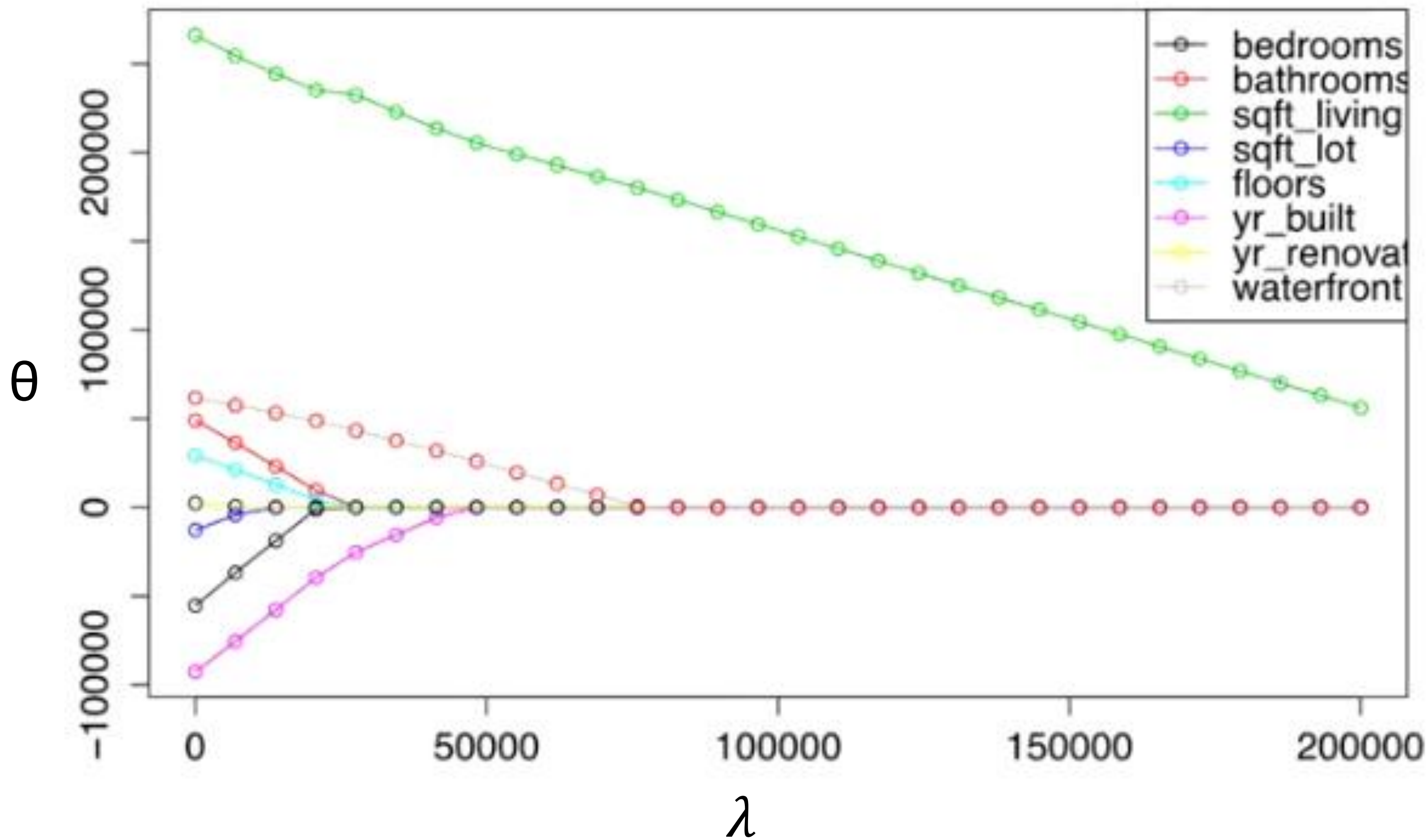
Lasso regresija je predložena u radu: *Tibshirani* (1996), “*Regression Shrinkage and Selection via the Lasso*”

Lasso naspram ridge



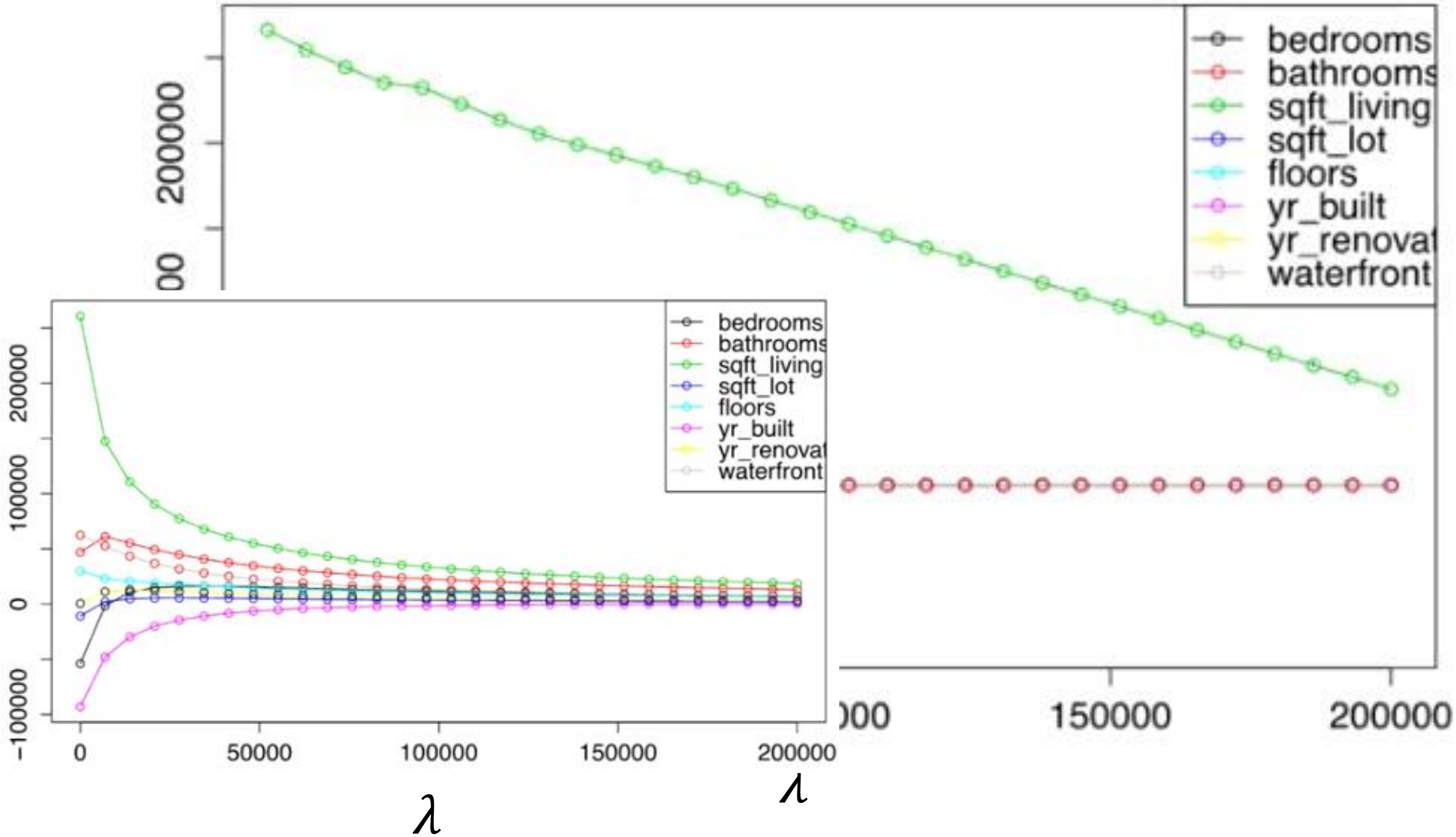
Ridge: sa povećanjem λ svi koeficijenti θ teže 0

Lasso naspram ridge



Lasso: za određene vrednosti λ, određeni koeficijenti θ postaju 0

Lasso naspram ridge



Lasso: za određene vrednosti λ, određeni koeficijenti θ postaju 0

Primer – sintetički podaci

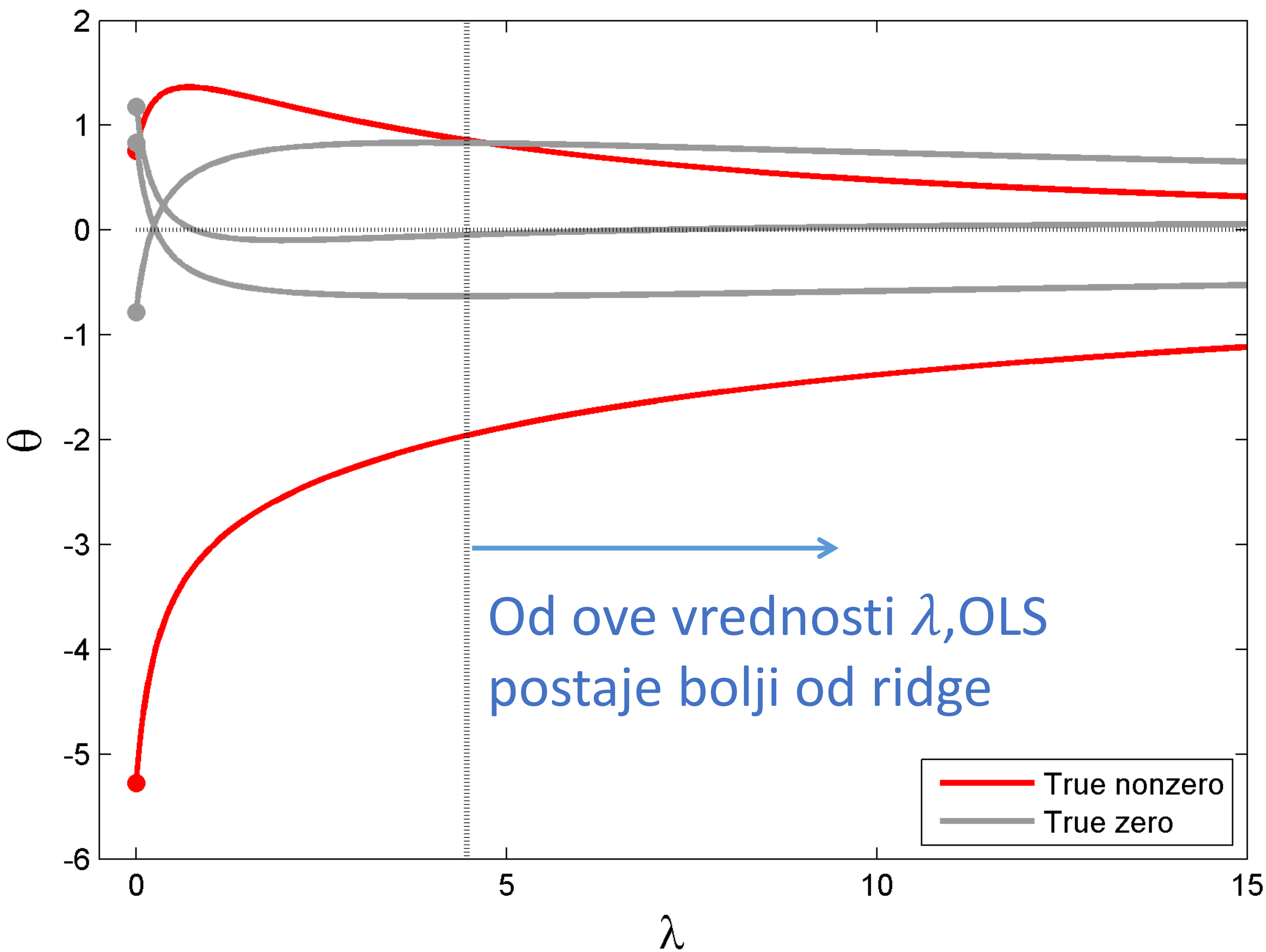
- 10 instanci, 5 prediktora $x_1, \dots, x_5$
  - Za svaku instancu $x_1, \dots, x_5$ su slučajno generisan brojevi iz standardne normalne raspodale
  - Napravljeno je ciljno obeležje

$$y = 2 \cdot x_1 - 3 \cdot x_2$$

  - Na y je dodat šum (slučan generisan broja iz normalne raspodele sa srednjom vrednošću 0 i standardnom devijacijom 0.9)

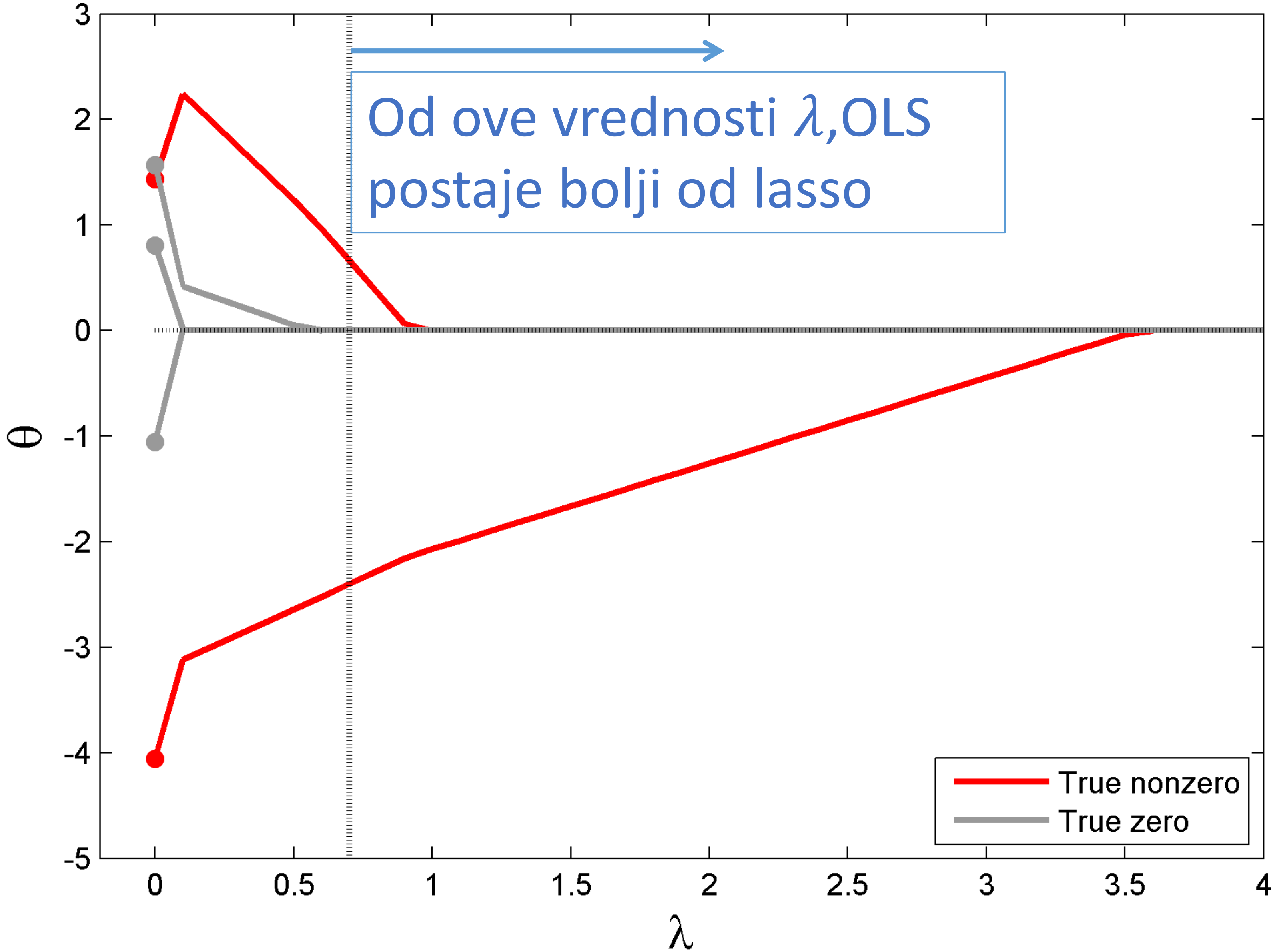
- Trebali bismo imati samo dva ne-nula koeficijenta (obeležja x_1 i x_2)

Primer – sintetički podaci *Ridge*



Od ove vrednosti λ,OLS postaje bolji od ridge

Primer – sintetički podaci *Lasso*



Ridge

- Dakle, *ridge* ne može vršiti selekciju obeležja, a *lasso* može
- *Ridge* će raditi dobro kada su stvarni ne-0 koeficijenti mali
- Ali neće raditi dobro ako su stvarni ne-0 koeficijenti veliki. Međutim, i dalje će raditi bolje od OLS

Regularizacija – dve formulacije

- Sledeće dve formulacije su ekvivalentne
- Formulacija bez ograničenja:

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \lambda \|\theta\|_1$$

- Formulacija sa ograničenjem:

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 \text{, pod uslovom } \|\theta\|_1 \leq t$$

- Regularizacija forsira rešenje da leži na nekom geometrijskom obliku centriranom oko (0, 0)

Zašto *lasso* postavlja koeficijente na 0?

Slika je preuzeta iz knjige: Friedman, J., Hastie, T. and Tibshirani, R., 2001. *The elements of statistical learning*

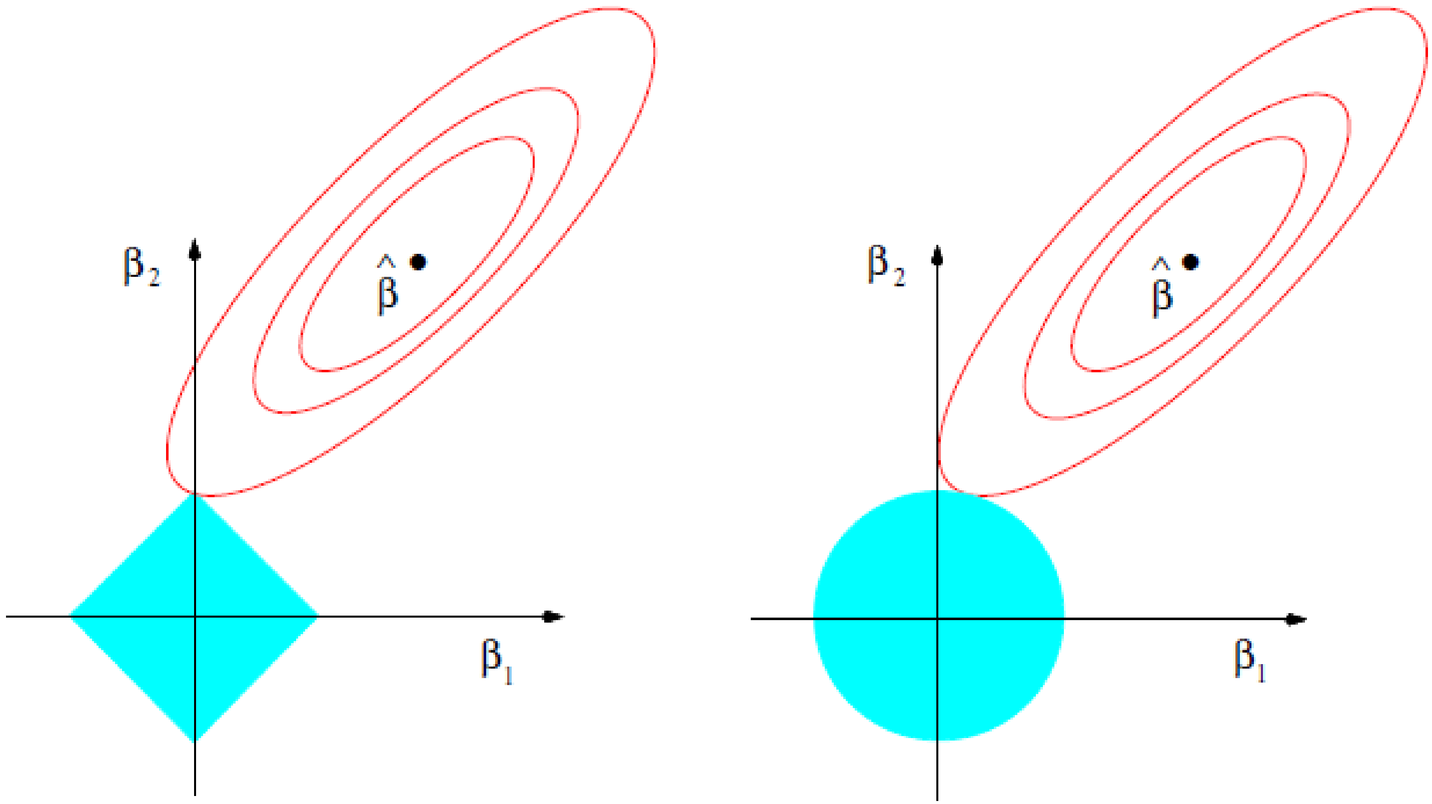
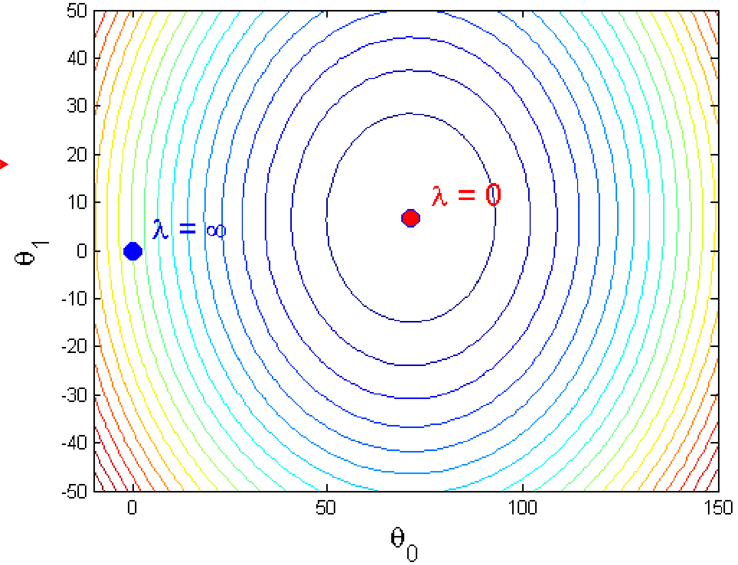


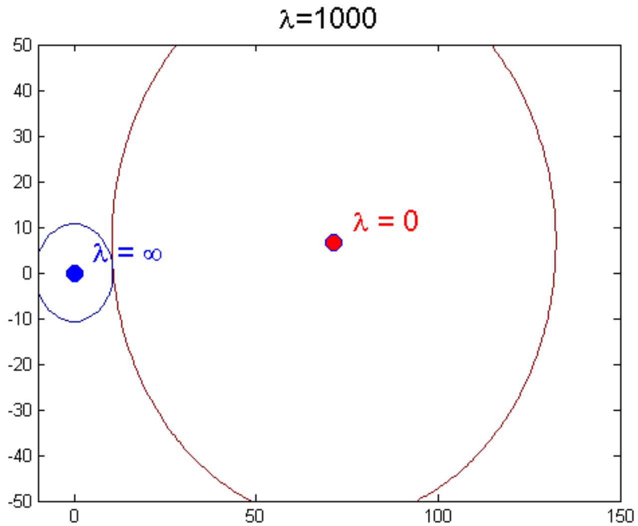
FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

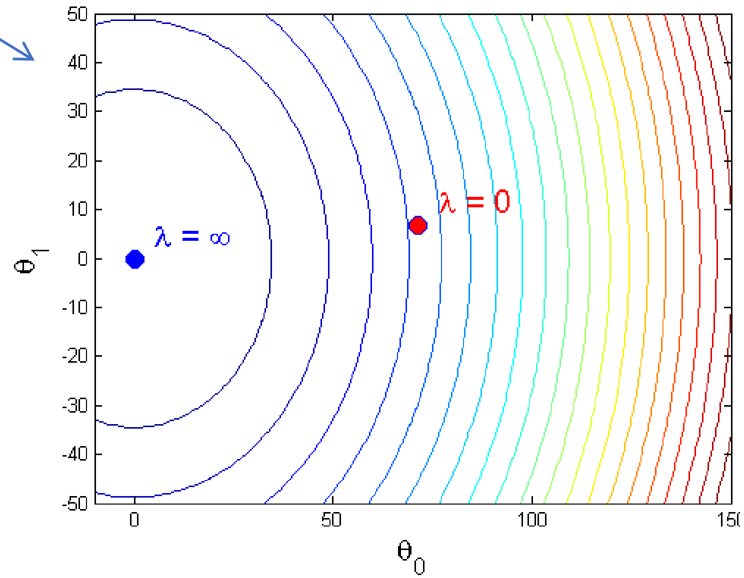
Ridge regression vizuelizacija

$$J(\theta) = \frac{1}{2N}\sum_{i=1}^{N}\left(\theta_0 + \theta_1 x^{(i)} - y^{(i)}\right)^2$$

$$+\lambda\left({\theta_0}^2 + {\theta_1}^2\right)$$





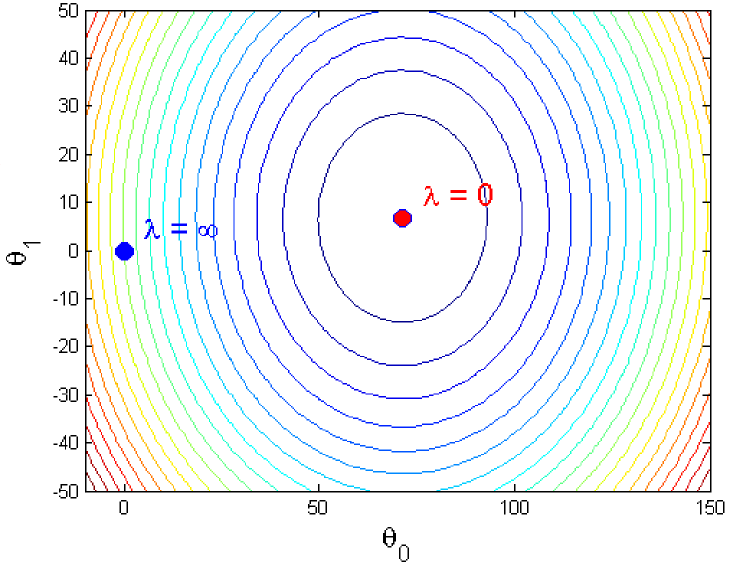


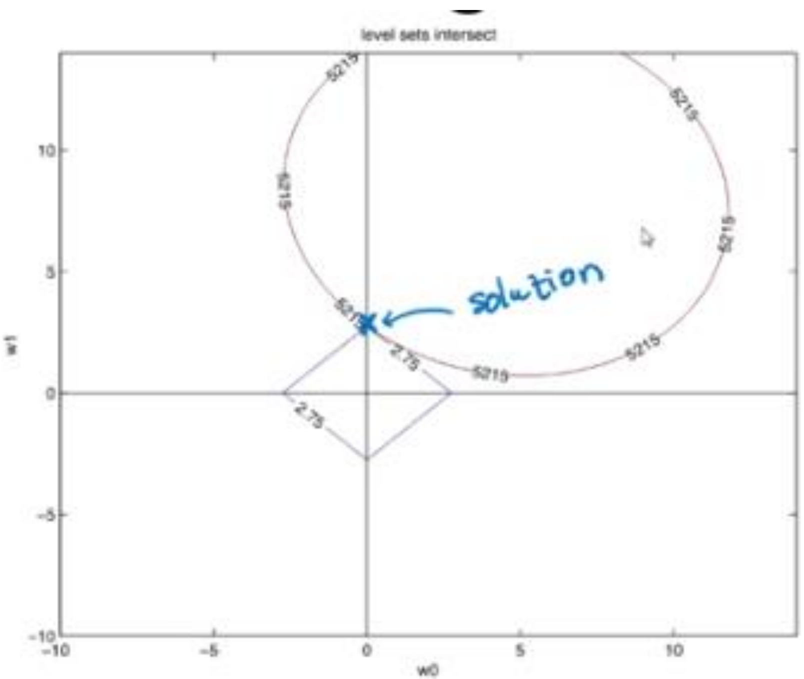
Za neku specifičnu vrednost
λ imamo nagodbu

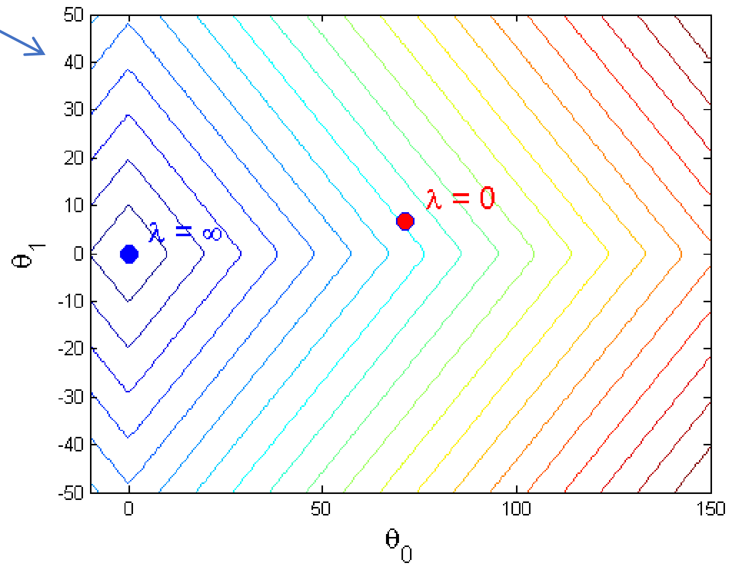
Lasso vizuelizacija

$$J(\theta) = \frac{1}{2N}\sum_{i=1}^{N}\left(\theta_0 + \theta_1 x^{(i)} - y^{(i)}\right)^2$$

$$+\lambda(|\theta_0| + |\theta_1|)$$





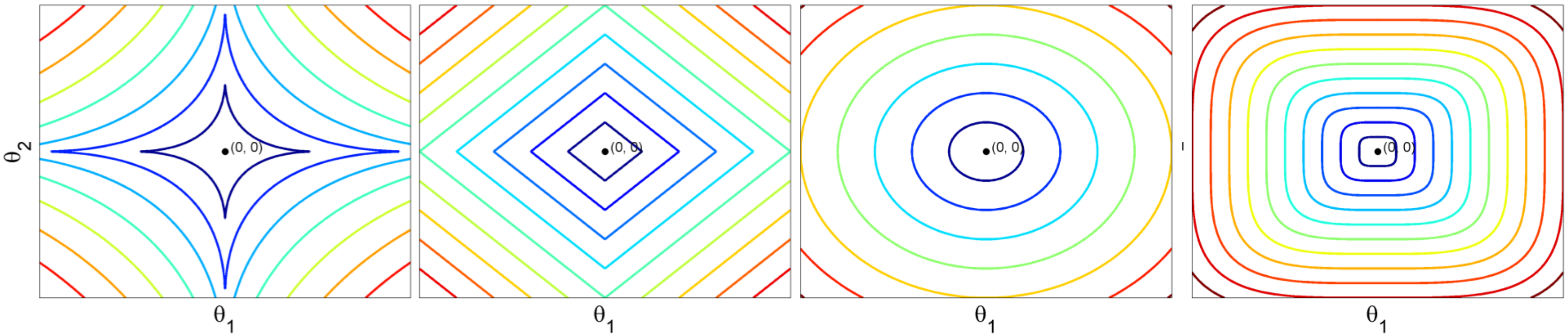


Izbor regularizacionog izraza Ω

- Lasso (L_1) i Ridge (L_2) su specijalni slučajevi L_p regularizacionog izraza:

$$\left(\sum_{d} |\theta_d|^p\right)^{\frac{1}{p}}$$

$p = 0.5$ | $p = 1$ (Lasso) | $p = 2$ (Ridge) | $p = 4$

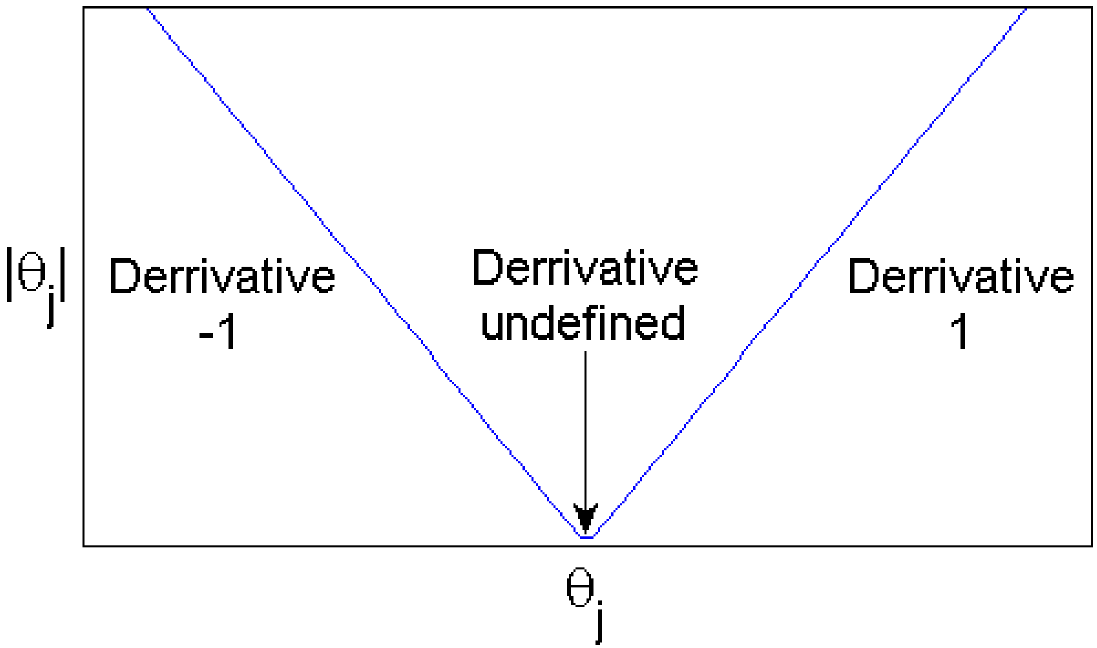


Prednost lasso nad ridge: interpretacija

- Za interpretaciju modela *Lasso* ima veliku prednost nad *ridge*
- Pošto lasso postavlja koeficijente *tačno* na 0, vrši selekciju obeležja i rezultuje *sparse* modelima

Lasso regression gradient descent

- Nedostatak lasso: ciljna funkcija nije diferencijabilna



- *Closed-form solution* ne postoji
- Umesto gradijenta se koriste podgradijenti (*subgradients*)