SD-TSIA204: Lasso

Ekhine Irurozki
Télécom Paris, IP Paris

Reminding the model

$$\mathbf{y} = X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon} \in \mathbb{R}^n$$

$$X = [\mathbf{x}_1, \dots, \mathbf{x}_p] = \begin{pmatrix} x_{1,1} & \dots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \dots & x_{n,p} \end{pmatrix} \in \mathbb{R}^{n \times p}, \boldsymbol{\theta}^\star \in \mathbb{R}^p$$

Motivation

In the presence of super-collinearity the OLS estimators can not be given.
Estimators $\hat{\boldsymbol{\theta}}$ with many zero coefficients are useful:

- for interpretation
- for computational efficiency if p is huge

Underlying idea: **variable selection**

Rem: also useful if $\boldsymbol{\theta}^{\star}$ has few non-zero coefficients

Variable selection overview

- **Screening**: remove the $\mathbf{x}_j$'s whose correlation with $\mathbf{y}$ is weak
  - pros: fast ($+++$), *i.e.*,one pass over data, intuitive ($+++$)
  - cons: neglect variables interactions $\mathbf{x}_j$, weak theory (- - -)
- **Greedy** methods aka stagewise / stepwise
  - pros: fast ($++$), intuitive ($++$)
  - cons: propagates wrong selection forward; weak theory (-)
- Sparsity enforcing **penalized** methods (*e.g.*,Lasso)
  - pros: better theory for convex cases ($++$)
  - cons: can be still slow (-)

The ℓ_0 pseudo-norm

The **support** of $\boldsymbol{\theta} \in \mathbb{R}^p$ is the set of indexes of non-zero coordinates:

$$\operatorname{supp}(\boldsymbol{\theta}) = \{j \in [\![1, p]\!], \theta_j \neq 0\}$$

The ℓ_0 **pseudo-norm** of a $\boldsymbol{\theta} \in \mathbb{R}^p$ is the number of non-zero coordinates:

$$\|\boldsymbol{\theta}\|_0 = \operatorname{card}\{j \in [\![1, p]\!], \theta_j \neq 0\}$$

Rem: $\|\cdot\|_0$ is not a norm, $\forall t \in \mathbb{R}^*, \|t\boldsymbol{\theta}\|_0 = \|\boldsymbol{\theta}\|_0$

Rem: $\|\cdot\|_0$ it is not even convex, $\boldsymbol{\theta}_1 = (1, 0, 1, \ldots, 0)$
$\boldsymbol{\theta}_2 = (0, 1, 1, \ldots, 0)$ and $3 = \|\frac{\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2}{2}\|_0 \geqslant \frac{\|\boldsymbol{\theta}_1\|_0 + \|\boldsymbol{\theta}_2\|_0}{2} = 2$

Regularization with the ℓ_0 penalty

First try to get a sparsity enforcing penalty: use ℓ_0 as a penalty (or regularization)

$$\boxed{\hat{\boldsymbol{\theta}}_\lambda = \mathop{\arg\min}_{\boldsymbol{\theta}\in\mathbb{R}^p} \left(\underbrace{\frac{1}{2}\|\mathbf{y} - X\boldsymbol{\theta}\|_2^2}_{\textbf{data fitting}} + \underbrace{\lambda\|\boldsymbol{\theta}\|_0}_{\textbf{regularization}} \right)}$$

Combinatorial problem!!!

Exact solution: require considering all sub-models, *i.e.*,computing OLS for all possible support; meaning one might need 2^p least squares computation!

Example :
$p = 10$ possible: $\approx 10^3$ least squares
$p = 30$ impossible: $\approx 10^{10}$ least squares

Rem: problem "NP-hard", can be solved for small problems by mixed integer programming.

Lasso: penalty point of view

Lasso: *Least Absolute Shrinkage and Selection Operator* Tibshirani (1996)

$$\boxed{\hat{\boldsymbol{\theta}}_\lambda^{\text{Lasso}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\arg\min} \left(\underbrace{\frac{1}{2}\|\mathbf{y} - X\boldsymbol{\theta}\|_2^2}_{\textbf{data fitting}} + \underbrace{\lambda\|\boldsymbol{\theta}\|_1}_{\textbf{regularization}} \right)}$$

or $\|\boldsymbol{\theta}\|_1 = \sum_{j=1}^{p} |\theta_j|$ sum of absolute values of the coefficients)

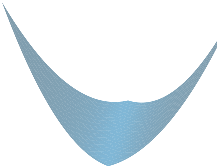
- We recover the limiting cases:

$$\lim_{\lambda\to 0} \hat{\boldsymbol{\theta}}_\lambda^{\text{Lasso}} = \hat{\boldsymbol{\theta}}^{\text{OLS}}$$

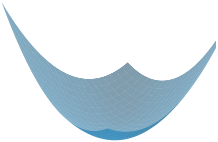
$$\lim_{\lambda\to +\infty} \hat{\boldsymbol{\theta}}_\lambda^{\text{Lasso}} = 0 \in \mathbb{R}^p$$

Exercise: the Lasso estimator is not always **unique** for a fixed λ (consider cases with two equals columns in X). However, the prediction is unique. Show these points.

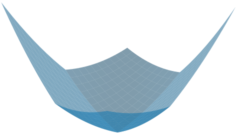
Optimization in $\mathbb{R}^d$



OLS

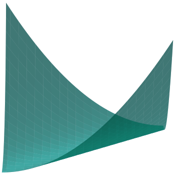


Ridge

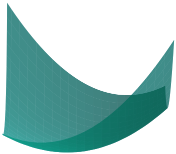


Lasso

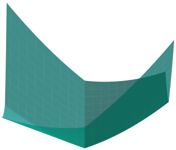
Optimization in $\mathbb{R}^d$



OLS

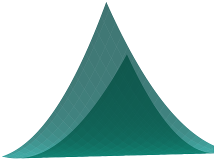


Ridge

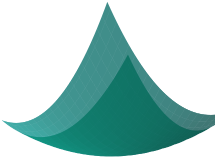


Lasso

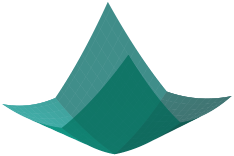
Optimization in $\mathbb{R}^d$



OLS

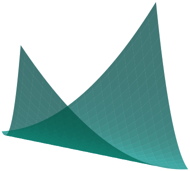


Ridge

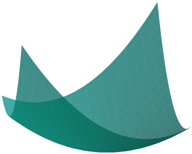


Lasso

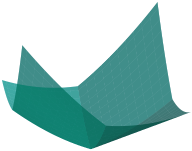
Optimization in $\mathbb{R}^d$



OLS

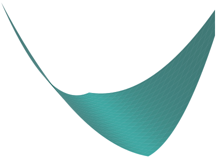


Ridge

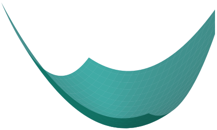


Lasso

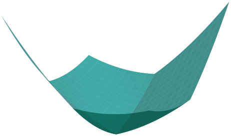
Optimization in $\mathbb{R}^d$



OLS



Ridge



Lasso

Constraint point of view

The following problem:

$$\boxed{\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{Lasso}} = \mathop{\arg\min}_{\boldsymbol{\theta}\in\mathbb{R}^p} \left(\underbrace{\frac{1}{2}\|\mathbf{y} - X\boldsymbol{\theta}\|_2^2}_{\textbf{data fitting}} + \underbrace{\lambda\|\boldsymbol{\theta}\|_1}_{\textbf{regularization}} \right)}$$

shares the same solutions as the constrained formulation:

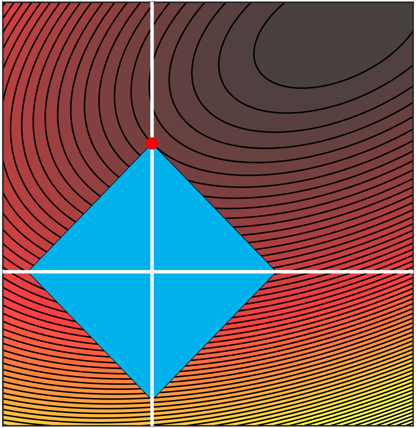
$$\begin{cases} \mathop{\arg\min}_{\boldsymbol{\theta}\in\mathbb{R}^p} \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2 \\ \text{s.t. } \|\boldsymbol{\theta}\|_1 \leqslant T \end{cases}$$

for some $T > 0$.

Rem: unfortunately the link $T \leftrightarrow \lambda$ is not explicit

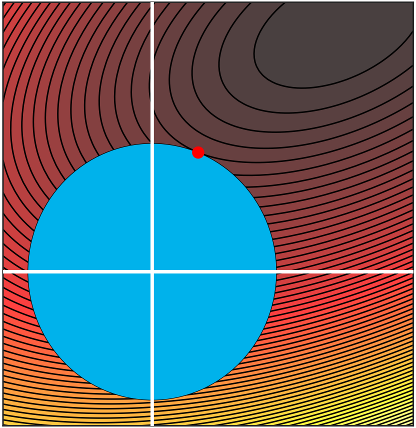
- If $T \to 0$ one recovers the null vector: $0 \in \mathbb{R}^p$
- If $T \to \infty$ one recovers $\hat{\boldsymbol{\theta}}^{\mathrm{OLS}}$ (unconstrained)

Zeroing coefficients



Optimization under ℓ_1 constraint : sparse solution

Zeroing coefficients



Optimization under ℓ_2 constraint : non sparse solution

Analitical solution

In general, there is no explicit solution

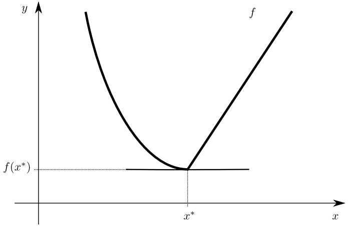
- Quadratic programming with constraints
- Iterative ridge
- Proximal gradient method (SD-TSIA 211)

Sub-gradients / sub-differential

For a convex function $f : \mathbb{R}^n \to \mathbb{R}$, $u \in \mathbb{R}^n$ is a **sub-gradient** of f at x^*, if for any $x \in \mathbb{R}^n$,

$$f(x) \geqslant f(x^*) + \langle u, x - x^* \rangle$$

The **sub-differential** is the set
$\partial f(x^*) = \{u \in \mathbb{R}^n : \forall x \in \mathbb{R}^n, f(x) \geqslant f(x^*) + \langle u, x - x^* \rangle\}$.
Rem: if the sub-gradient is unique, one recovers the standard gradient

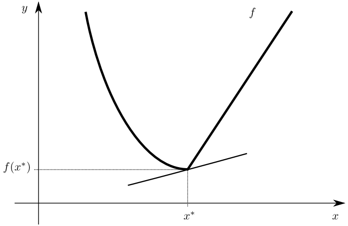


Sub-gradients / sub-differential

For a convex function $f : \mathbb{R}^n \to \mathbb{R}$, $u \in \mathbb{R}^n$ is a **sub-gradient** of f at x^*, if for any $x \in \mathbb{R}^n$,

$$f(x) \geqslant f(x^*) + \langle u, x - x^* \rangle$$

The **sub-differential** is the set
$\partial f(x^*) = \{u \in \mathbb{R}^n : \forall x \in \mathbb{R}^n, f(x) \geqslant f(x^*) + \langle u, x - x^* \rangle\}$.
Rem: if the sub-gradient is unique, one recovers the standard gradient

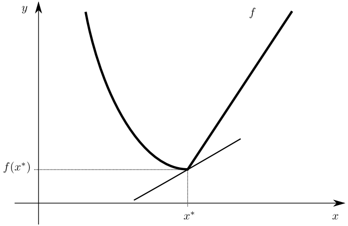


Sub-gradients / sub-differential

For a convex function $f : \mathbb{R}^n \to \mathbb{R}$, $u \in \mathbb{R}^n$ is a **sub-gradient** of f at x^*, if for any $x \in \mathbb{R}^n$,

$$f(x) \geqslant f(x^*) + \langle u, x - x^* \rangle$$

The **sub-differential** is the set
$\partial f(x^*) = \{u \in \mathbb{R}^n : \forall x \in \mathbb{R}^n, f(x) \geqslant f(x^*) + \langle u, x - x^* \rangle\}$.
Rem: if the sub-gradient is unique, one recovers the standard gradient

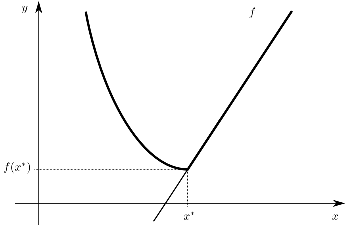


Sub-gradients / sub-differential

For a convex function $f : \mathbb{R}^n \to \mathbb{R}$, $u \in \mathbb{R}^n$ is a **sub-gradient** of f at x^*, if for any $x \in \mathbb{R}^n$,

$$f(x) \geqslant f(x^*) + \langle u, x - x^* \rangle$$

The **sub-differential** is the set
$\partial f(x^*) = \{u \in \mathbb{R}^n : \forall x \in \mathbb{R}^n, f(x) \geqslant f(x^*) + \langle u, x - x^* \rangle\}$.
Rem: if the sub-gradient is unique, one recovers the standard gradient



Fermat's Rule

Theorem A point x^* is a minimum of a convex function $f : \mathbb{R}^n \to \mathbb{R}$ if and only if $0 \in \partial f(x^*)$
Proof: use the sub-gradient definition:

- 0 is a sub-gradient of f at x^* if and only if $\forall x \in \mathbb{R}^n, f(x) \geqslant f(x^*) + \langle 0, x - x^* \rangle$

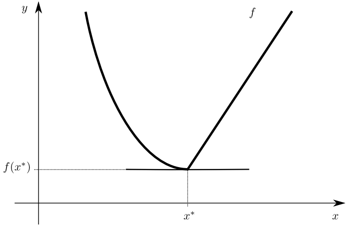
Fermat's Rule

Theorem A point x^* is a minimum of a convex function $f : \mathbb{R}^n \to \mathbb{R}$ if and only if $0 \in \partial f(x^*)$

Proof: use the sub-gradient definition:

- 0 is a sub-gradient of f at x^* if and only if $\forall x \in \mathbb{R}^n, f(x) \geqslant f(x^*) + \langle 0, x - x^* \rangle$

Rem:Visually, it corresponds to a horizontal tangent



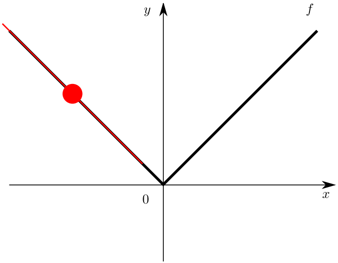
Absolute value sub-differential

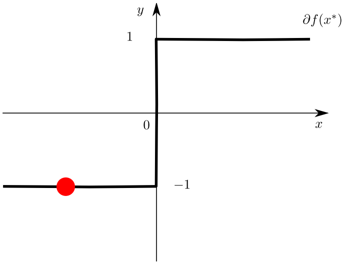
Function (abs):

$$f : \begin{cases} \mathbb{R} & \to \mathbb{R} \\ x & \mapsto |x| \end{cases}$$

Sub-differential (sign)

$$\partial f(x^*) = \begin{cases} \{-1\} & \text{if } x^* \in]-\infty, 0[\\ \{1\} & \text{if } x^* \in]0, \infty[\\ [-1, 1] & \text{if } x^* = 0 \end{cases}$$

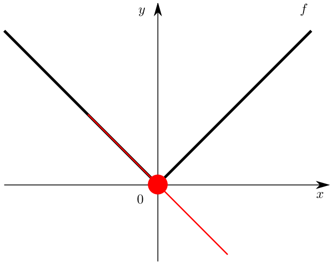




Absolute value sub-differential

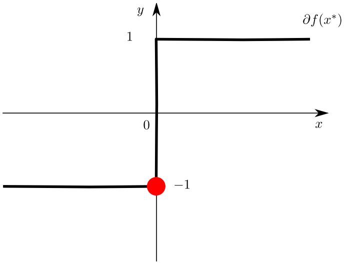
Function (abs):

$$f : \begin{cases} \mathbb{R} & \to \mathbb{R} \\ x & \mapsto |x| \end{cases}$$



Sub-differential (sign)

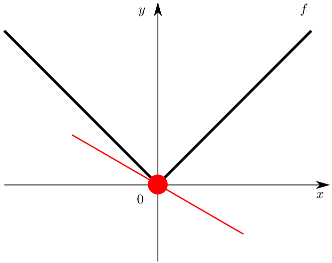
$$\partial f(x^*) = \begin{cases} \{-1\} & \text{if } x^* \in]-\infty, 0[\\ \{1\} & \text{if } x^* \in]0, \infty[\\ [-1, 1] & \text{if } x^* = 0 \end{cases}$$



Absolute value sub-differential

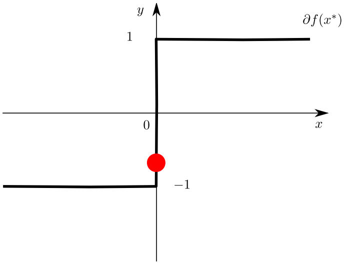
Function (abs):

$$f : \begin{cases} \mathbb{R} & \to \mathbb{R} \\ x & \mapsto |x| \end{cases}$$



Sub-differential (sign)

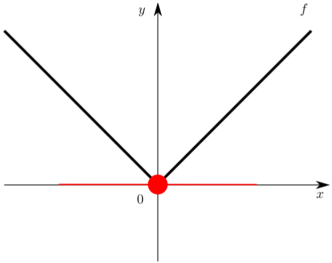
$$\partial f(x^*) = \begin{cases} \{-1\} & \text{if } x^* \in]-\infty, 0[\\ \{1\} & \text{if } x^* \in]0, \infty[\\ [-1, 1] & \text{if } x^* = 0 \end{cases}$$



Absolute value sub-differential

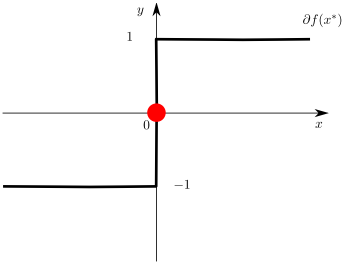
Function (abs):

$$f : \begin{cases} \mathbb{R} & \to \mathbb{R} \\ x & \mapsto |x| \end{cases}$$



Sub-differential (sign)

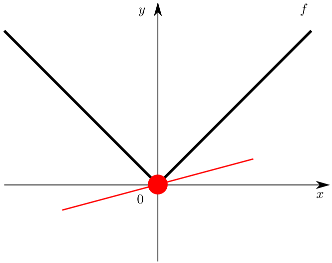
$$\partial f(x^*) = \begin{cases} \{-1\} & \text{if } x^* \in]-\infty, 0[\\ \{1\} & \text{if } x^* \in]0, \infty[\\ [-1, 1] & \text{if } x^* = 0 \end{cases}$$



Absolute value sub-differential

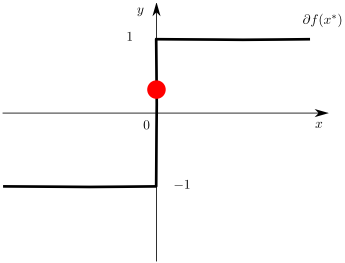
Function (abs):

$$f : \begin{cases} \mathbb{R} & \rightarrow \mathbb{R} \\ x & \mapsto |x| \end{cases}$$



Sub-differential (sign)

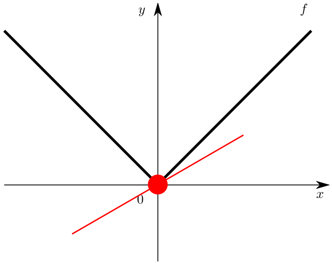
$$\partial f(x^*) = \begin{cases} \{-1\} & \text{if } x^* \in]-\infty, 0[\\ \{1\} & \text{if } x^* \in]0, \infty[\\ [-1, 1] & \text{if } x^* = 0 \end{cases}$$



Absolute value sub-differential

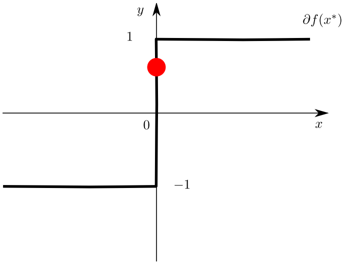
Function (abs):

$$f : \begin{cases} \mathbb{R} & \rightarrow \mathbb{R} \\ x & \mapsto |x| \end{cases}$$



Sub-differential (sign)

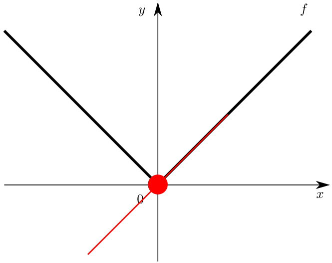
$$\partial f(x^*) = \begin{cases} \{-1\} & \text{if } x^* \in]-\infty, 0[\\ \{1\} & \text{if } x^* \in]0, \infty[\\ [-1, 1] & \text{if } x^* = 0 \end{cases}$$



Absolute value sub-differential

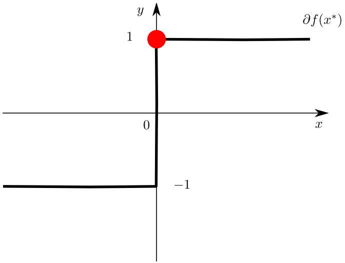
Function (abs):

$$f : \begin{cases} \mathbb{R} & \rightarrow \mathbb{R} \\ x & \mapsto |x| \end{cases}$$



Sub-differential (sign)

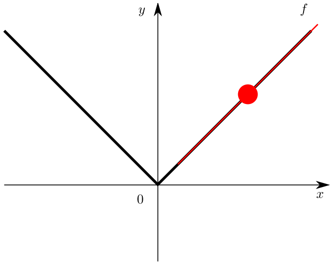
$$\partial f(x^*) = \begin{cases} \{-1\} & \text{if } x^* \in]-\infty, 0[\\ \{1\} & \text{if } x^* \in]0, \infty[\\ [-1, 1] & \text{if } x^* = 0 \end{cases}$$



Absolute value sub-differential

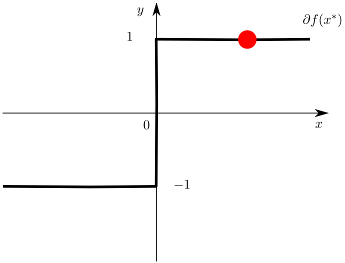
Function (abs):

$$f : \begin{cases} \mathbb{R} & \to \mathbb{R} \\ x & \mapsto |x| \end{cases}$$



Sub-differential (sign)

$$\partial f(x^*) = \begin{cases} \{-1\} & \text{if } x^* \in]-\infty, 0[\\ \{1\} & \text{if } x^* \in]0, \infty[\\ [-1, 1] & \text{if } x^* = 0 \end{cases}$$



Fermat's rule for the Lasso

$$\boxed{\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{Lasso}} = \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\mathbb{R}^p} \left(\underbrace{\frac{1}{2}\|\mathbf{y} - X\boldsymbol{\theta}\|_2^2}_{\textbf{data fitting}} + \underbrace{\lambda\|\boldsymbol{\theta}\|_1}_{\textbf{regularization}} \right)}$$

Necessary and sufficient optimality (Fermat):

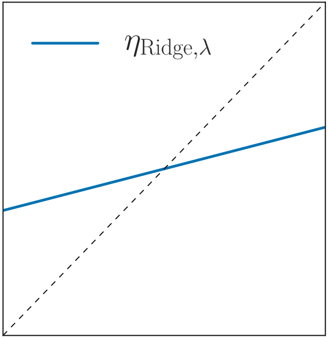
$$\boxed{\forall j \in [p],\ \mathbf{x}_j^\top \left(\frac{y - X\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{Lasso}}}{\lambda} \right) \in \begin{cases} \{\mathrm{sign}(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{Lasso}})_j\} & \text{if} \quad (\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{Lasso}})_j \neq 0, \\ [-1, 1] & \text{if} \quad (\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{Lasso}})_j = 0. \end{cases}}$$

Rem: If $\lambda > \lambda_{\max} := \max_{j\in[\![1,p]\!]} |\langle \mathbf{x}_j, \mathbf{y}\rangle|$, then $\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{Lasso}} = 0$

1D Regularization: Ridge

Solve: $\eta_\lambda(z) = \arg\min_{x \in \mathbb{R}} x \mapsto \frac{1}{2}(z-x)^2 + \frac{\lambda}{2}x^2$

$$\eta_\lambda(z) = \frac{z}{1+\lambda}$$

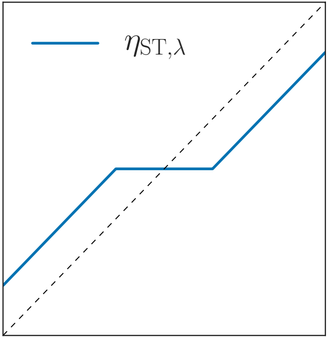


ℓ_2 shrinkage : Ridge

1D Regularization: Lasso

Solve: $\eta_\lambda(z) = \arg\min_{x\in\mathbb{R}} x \mapsto \frac{1}{2}(z-x)^2 + \lambda|x|$

$$\eta_\lambda(z) = \operatorname{sign}(z)(|z|-\lambda)_+$$

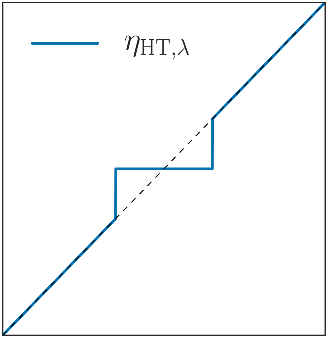


ℓ_1 shrinkage: soft thresholding

1D Regularization: ℓ_0

Solve: $\eta_\lambda(z) = \arg\min_{x \in \mathbb{R}} x \mapsto \frac{1}{2}(z-x)^2 + \lambda \mathbb{1}_{x \neq 0}$

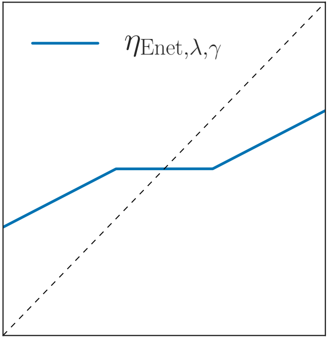
$$\eta_\lambda(z) = z \mathbb{1}_{|z| \geqslant \sqrt{2\lambda}}$$



ℓ_0 shrinkage: hard thresholding

1D Regularization: enet

Solve: $\eta_\lambda(z) = \arg\min_{x\in\mathbb{R}} x \mapsto \frac{1}{2}(z-x)^2 + \lambda(\gamma|x| + (1-\gamma)\frac{x^2}{2})$



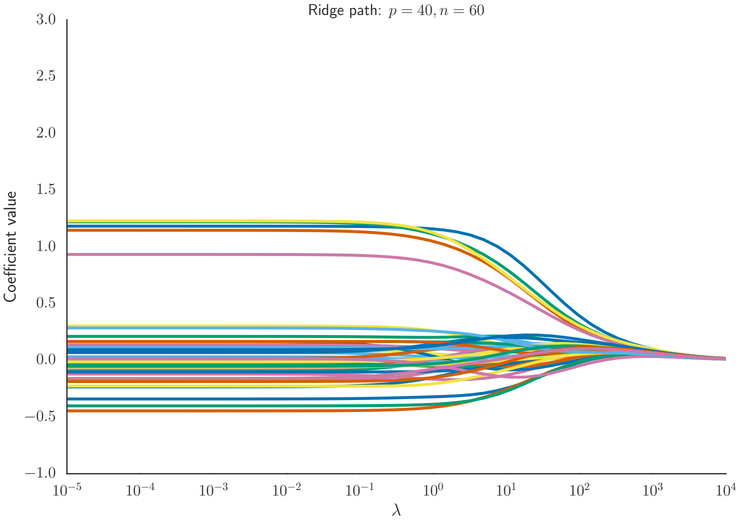
ℓ_1/ℓ_2

Numerical example on simulated data

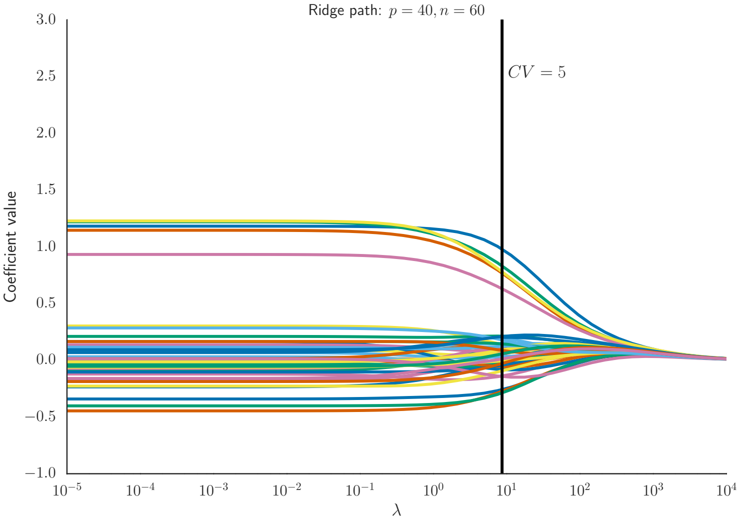
- $\boldsymbol{\theta}^\star = (1, 1, 1, 1, 1, 0, \dots, 0) \in \mathbb{R}^p$ (5 non-zero coefficients)
- $X \in \mathbb{R}^{n \times p}$ has columns drawn according to a Gaussian distribution
- $y = X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon} \in \mathbb{R}^n$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \operatorname{Id}_n)$
- We use a grid of 50 λ values

For this example : $n = 60, p = 40, \sigma = 1$

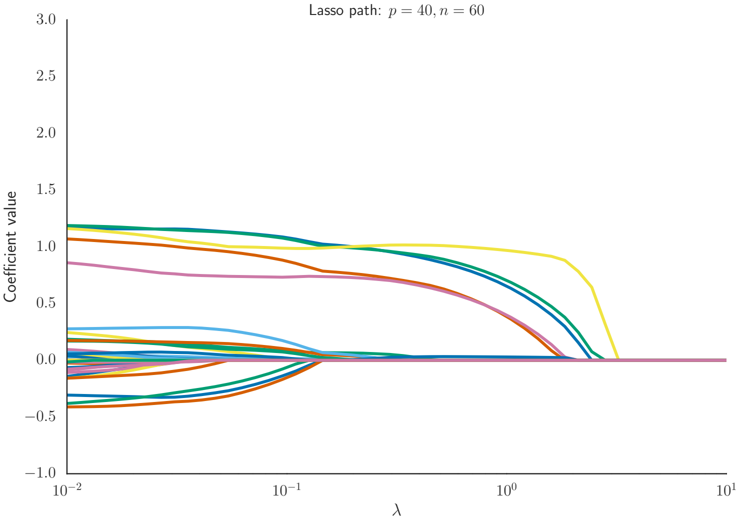
Lasso vs Ridge



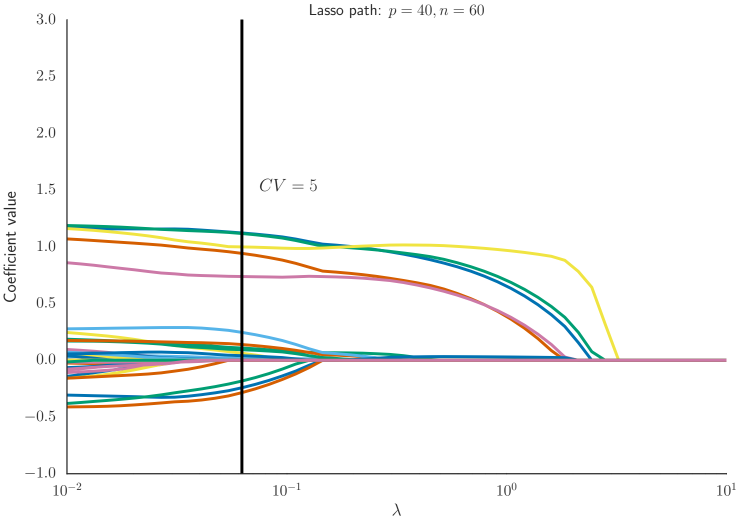
Lasso vs Ridge



Lasso vs Ridge



Lasso vs Ridge



Lasso properties

- Solutions is not necessarily unique
- The analytic form does not necessarily exist
- Numerical aspect: the Lasso is a **convex** problem
- Variable selection / sparse solutions: $\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{Lasso}}$ has potentially many zeroed coefficients. The λ parameter controls the sparsity level: if λ is large, solutions are very sparse.

  Example : We got 17 non-zero coefficients for `LassoCV` in the previous simulated example

  Rem: RidgeCV has no zero coefficients

Lasso analysis

Theory : more involved for the Lasso than for least squares / Ridge
Recent reference : Bühlmann and van de Geer (2011)

In a nutshell: add bias to the standard least squares to perform variance reduction

Elastic-net : ℓ_1/ℓ_2 regularization

The Elastic-Net, introduced by Zou and Hastie (2005) is the (unique) solution of

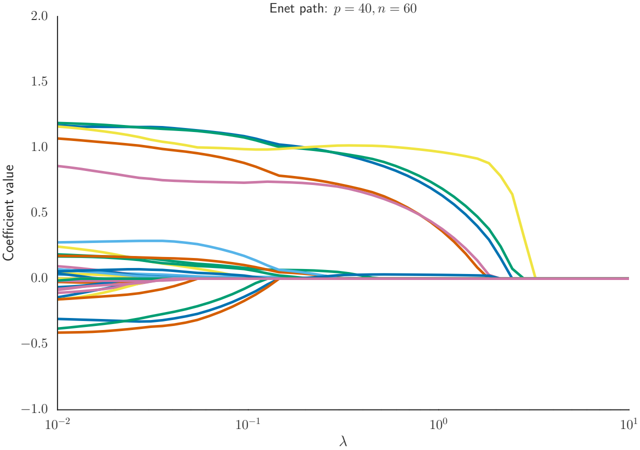
$$\hat{\boldsymbol{\theta}}_\lambda = \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\arg\min}\left[\frac{1}{2}\|\mathbf{y} - X\boldsymbol{\theta}\|_2^2 + \lambda\left(\gamma\|\boldsymbol{\theta}\|_1 + (1-\gamma)\frac{\|\boldsymbol{\theta}\|_2^2}{2}\right)\right]$$

Motivation: help selecting all relevant but correlated variable (not only one as for the Lasso)

Rem: two parameters needed, one for global regularization, one trading-off Ridge vs. Lasso

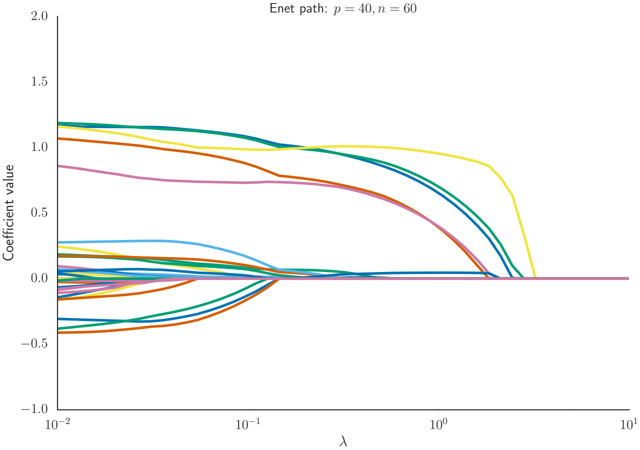
Rem: the solution is unique and the size of the Elastic-Net support is smaller than $\min(n, p)$

Elastic-Net: $\gamma\|\boldsymbol{\theta}\|_1 + (1-\gamma)\|\boldsymbol{\theta}\|_2^2/2$



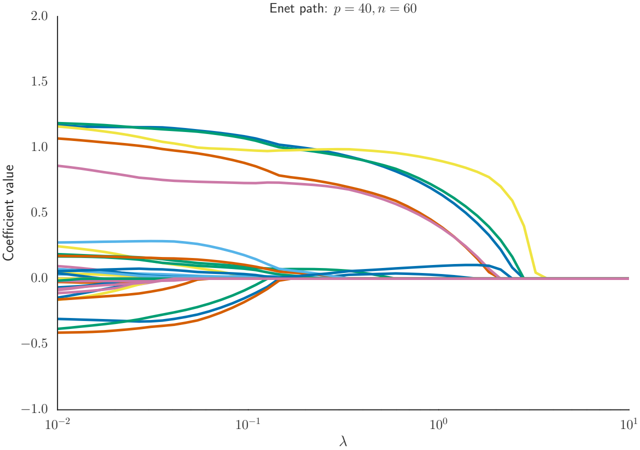
$$\gamma = 1.00$$

Elastic-Net: $\gamma\|\boldsymbol{\theta}\|_1 + (1-\gamma)\|\boldsymbol{\theta}\|_2^2/2$



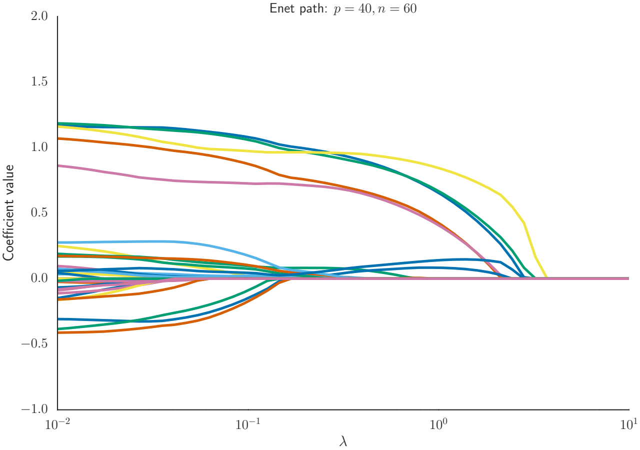
$$\gamma = 0.99$$

Elastic-Net: $\gamma\|\boldsymbol{\theta}\|_1 + (1-\gamma)\|\boldsymbol{\theta}\|_2^2/2$



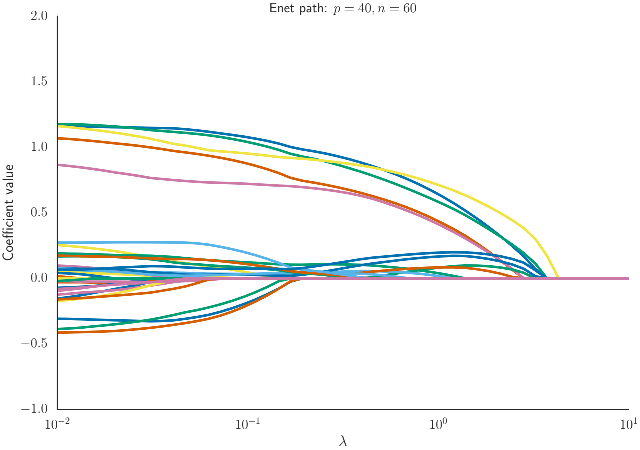
$$\gamma = 0.95$$

Elastic-Net: $\gamma\|\boldsymbol{\theta}\|_1 + (1-\gamma)\|\boldsymbol{\theta}\|_2^2/2$



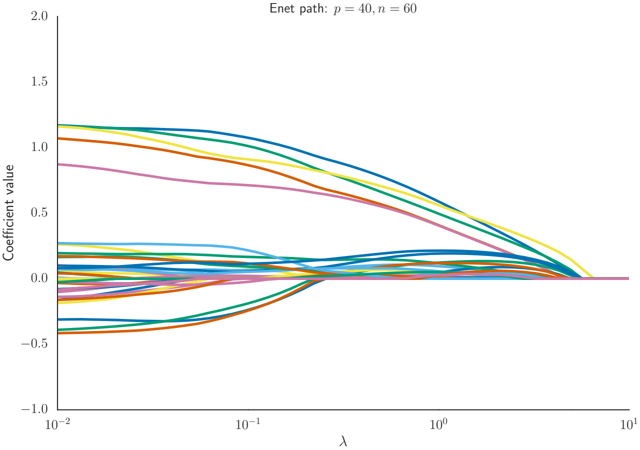
$$\gamma = 0.90$$

Elastic-Net: $\gamma\|\boldsymbol{\theta}\|_1 + (1-\gamma)\|\boldsymbol{\theta}\|_2^2/2$



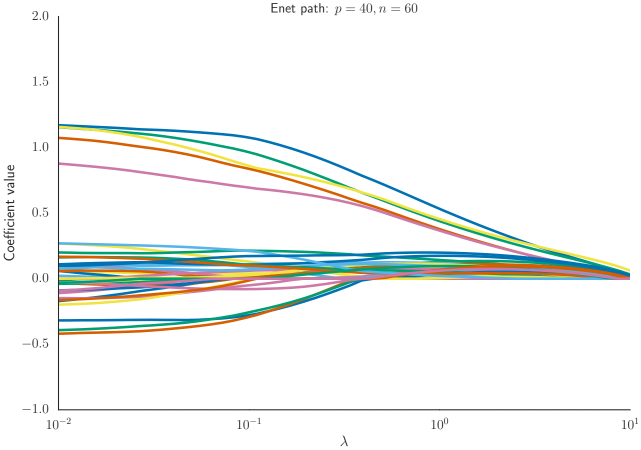
$$\gamma = 0.75$$

Elastic-Net: $\gamma\|\boldsymbol{\theta}\|_1 + (1-\gamma)\|\boldsymbol{\theta}\|_2^2/2$



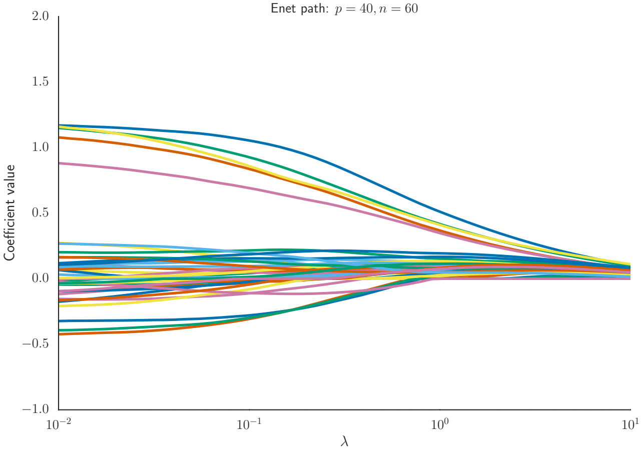
$$\gamma = 0.50$$

Elastic-Net: $\gamma\|\boldsymbol{\theta}\|_1 + (1-\gamma)\|\boldsymbol{\theta}\|_2^2/2$



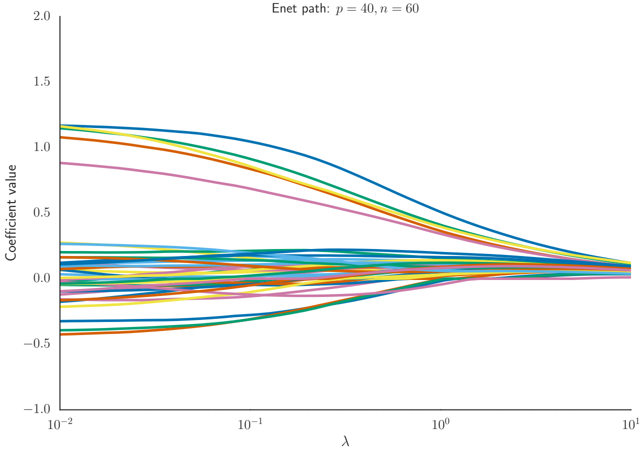
$$\gamma = 0.25$$

Elastic-Net: $\gamma\|\boldsymbol{\theta}\|_1 + (1-\gamma)\|\boldsymbol{\theta}\|_2^2/2$



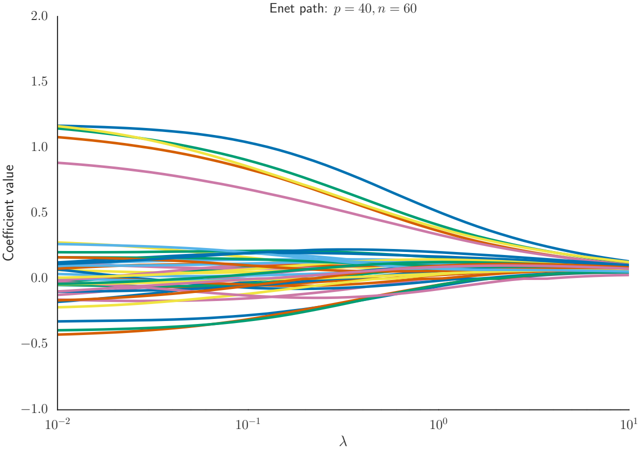
$$\gamma = 0.1$$

Elastic-Net: $\gamma\|\boldsymbol{\theta}\|_1 + (1-\gamma)\|\boldsymbol{\theta}\|_2^2/2$



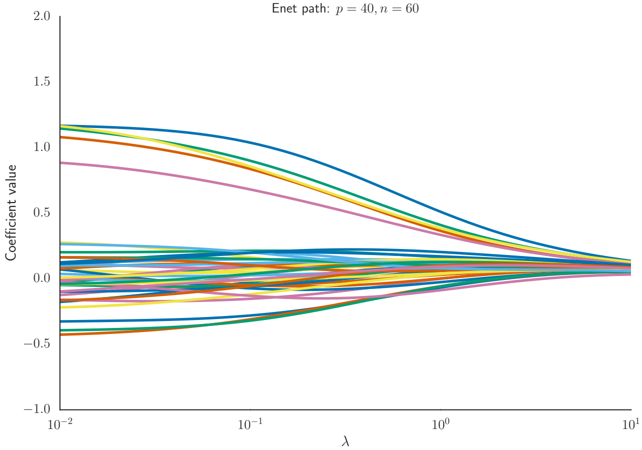
$$\gamma = 0.05$$

Elastic-Net: $\gamma\|\boldsymbol{\theta}\|_1 + (1-\gamma)\|\boldsymbol{\theta}\|_2^2/2$



$$\gamma = 0.01$$

Elastic-Net: $\gamma\|\boldsymbol{\theta}\|_1 + (1-\gamma)\|\boldsymbol{\theta}\|_2^2/2$



$$\gamma = 0.00$$

The Lasso bias

The Lasso is biased: it shrinks large coefficients towards 0

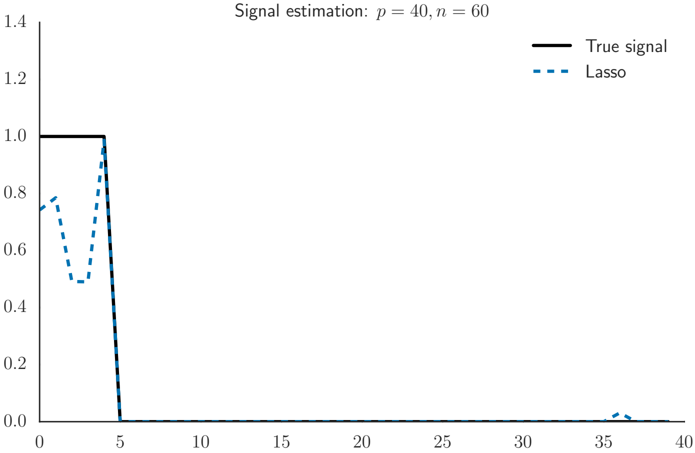


Illustration over the previous example

The Lasso bias

The Lasso is biased: it shrinks large coefficients towards 0

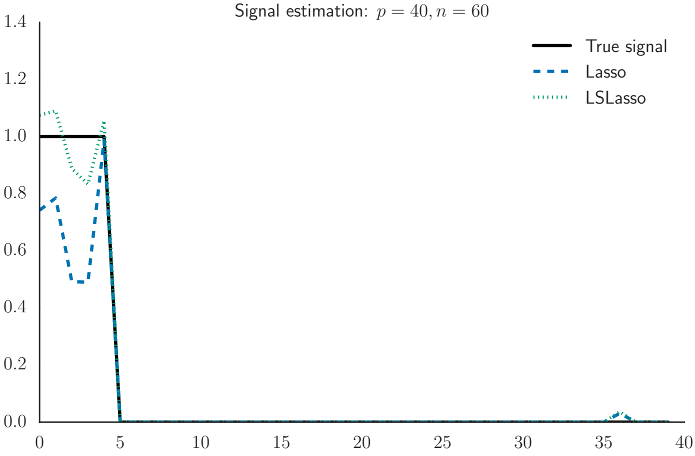


Illustration over the previous example

The Lasso bias: a simple remedy

How to rescale shrunk coefficients?

LSLasso (Least Square Lasso)

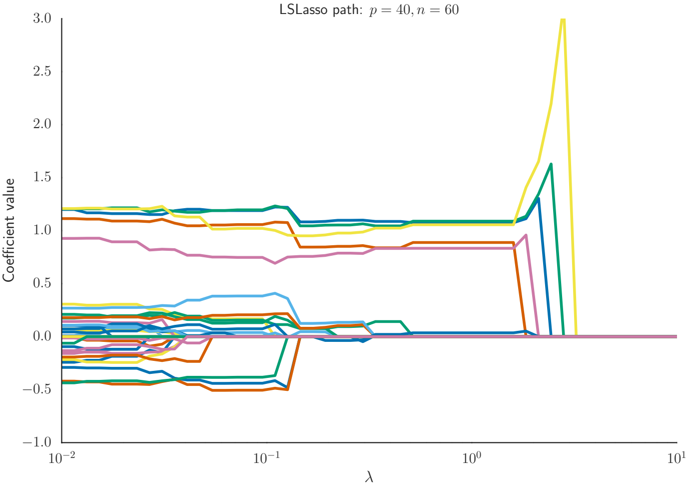
1. Lasso : compute $\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{Lasso}}$
2. Perform least squares over selected variables: $\mathrm{supp}(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{Lasso}})$

$$\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{LSLasso}} = \mathop{\arg\min}_{\substack{\boldsymbol{\theta}\in\mathbb{R}^p \\ \mathrm{supp}(\boldsymbol{\theta})=\mathrm{supp}(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{Lasso}})}} \frac{1}{2}\|\mathbf{y} - X\boldsymbol{\theta}\|_2^2$$

<u>Rem</u>: perform CV for the double step procedure; choosing λ by LassoCV and then performing OLS keeps too many variables

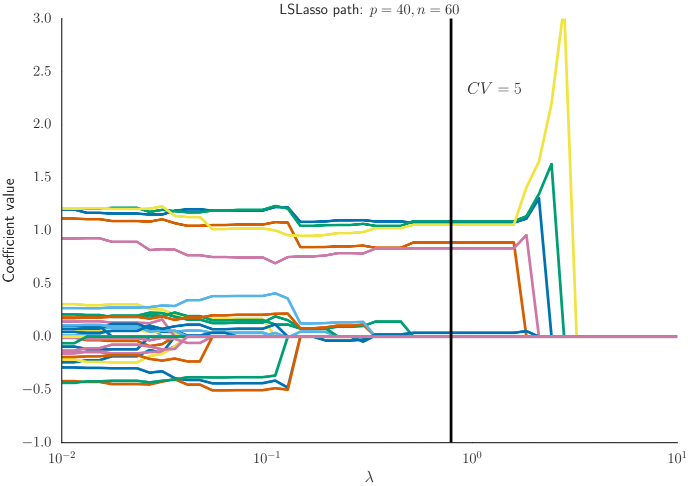
<u>Rem</u>: LSLasso is not coded in standard packages

De-biasing



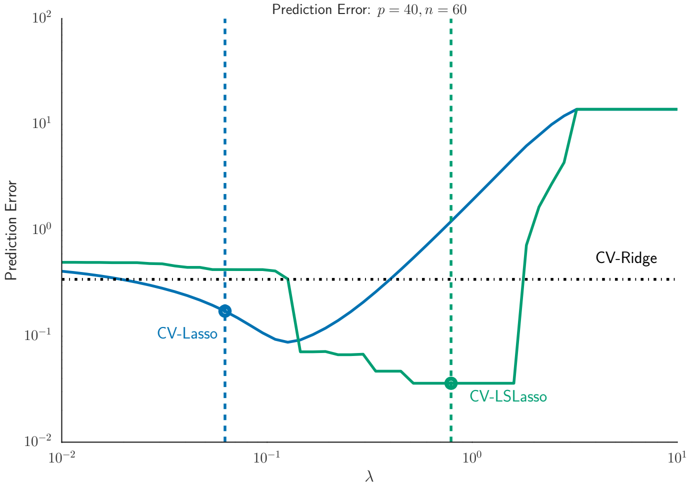
LSLasso path: $p = 40, n = 60$
Coefficient value
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
10^{-2}
10^{-1}
10^{0}
10^{1}
λ

De-biasing



LSLasso path: $p = 40, n = 60$
$CV = 5$
Coefficient value
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
10^{-2}
10^{-1}
10^{0}
10^{1}
λ

Prediction: Lasso vs. LSLasso



LSLasso evaluation

Pros

- the “true” large coefficients are less shrunk
- CV recovers less “parasite” variables (improve interpretability) *e.g.*,in the previous example the LSLassoCV recovers exactly the 5 “true” non zero variables, up to a single false positive

LSLasso: especially useful for <u>estimation</u>

Cons

- the difference in term of prediction is not always striking
- requires (slightly) more computation: needs to compute as many OLS as λ's

References I

- P. Bühlmann and S. van de Geer.
  Statistics for high-dimensional data.
  Springer Series in Statistics. Springer, Heidelberg, 2011.
  Methods, theory and applications.
- R. Tibshirani.
  Regression shrinkage and selection via the lasso.
  J. Roy. Statist. Soc. Ser. B, 58(1):267–288, 1996.
- H. Zou and T. Hastie.
  Regularization and variable selection via the elastic net.
  J. Roy. Statist. Soc. Ser. B, 67(2):301–320, 2005.