Convex optimization, sparsity and regression in high dimension

CIMAT 2015

Joseph Salmon
http://josephsalmon.eu
LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay

Outline

Variable selection and sparsity

- Motivation and variable selection variants
- ℓ_0 and ℓ_1 penalties
- Sub-gradients / sub-differential

Lasso extensions and improvements

- LSLasso : Least-Square Lasso
- Lasso variants : Elastic Net
- Group structure
- Multivariate / Multi-task regression

Optimization for the Lasso

- Coordinate descent
- Proximal methods — Forward / Backward

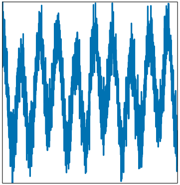
Theoretical results for the Lasso

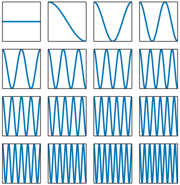
- Prediction error
- Estimation error

Sparsity of signals is all around

Signals can often be represented through a combination of a few **atoms** / **features** :

- Fourier decomposition for sounds

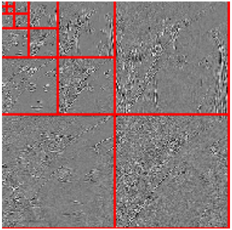




Sparsity of signals is all around

Signals can often be represented through a combination of a few **atoms** / **features** :

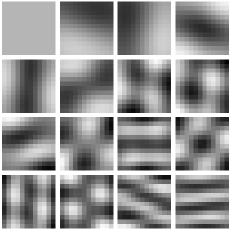
- Fourier decomposition for sounds
- Wavelet for images (1990's)



Sparsity of signals is all around

Signals can often be represented through a combination of a few **atoms** / **features** :

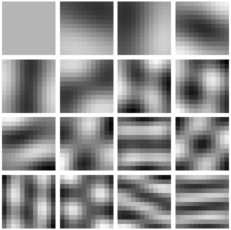
- Fourier decomposition for sounds
- Wavelet for images (1990's)
- Dictionary learning for images (late 2000's)



Sparsity of signals is all around

Signals can often be represented through a combination of a few **atoms** / **features** :

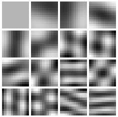
- Fourier decomposition for sounds
- Wavelet for images (1990's)
- Dictionary learning for images (late 2000's)
- etc.

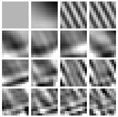


Sparse linear model

Let $y \in \mathbb{R}^n$ be a signal

Let $X = [\mathbf{x}_1, \ldots, \mathbf{x}_p] \in \mathbb{R}^{n \times p}$ be a collection of p atoms/features : corresponds to a **dictionary**





X is well suited if one can approximate the signal $y \approx X\beta^*$ with a **sparse** vector $\beta^* \in \mathbb{R}^p$

Objectives :

- Estimation β^*
- Prediction $X\beta^*$

Constraints : large p, n, sparse β^*

$$\underbrace{\begin{pmatrix} y \end{pmatrix}}_{y \in \mathbb{R}^n} \approx \underbrace{\begin{pmatrix} \mathbf{x}_1 & \Big| & \ldots & \Big| & \mathbf{x}_p \end{pmatrix}}_{X \in \mathbb{R}^{n \times p}} \cdot \underbrace{\begin{pmatrix} \beta_1^* \\ \vdots \\ \beta_p^* \end{pmatrix}}_{\beta^* \in \mathbb{R}^p}$$

Statistical model : linear regression

$$\boxed{y = X\beta^* + \varepsilon}$$

Observed signal : $y \in \mathbb{R}^n$

Noise : $\varepsilon \in \mathbb{R}^n$ (*e.g.*, $\mathcal{N}(0, \sigma^2 \operatorname{Id}_n)$)

Design matrix : $X = [\mathbf{x}_1, \ldots, \mathbf{x}_p] = \begin{pmatrix} x_{1,1} & \ldots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \ldots & x_{n,p} \end{pmatrix} \in \mathbb{R}^{n \times p}$

True (unknown) signal : $\beta^* \in \mathbb{R}^p$

Estimated signal : $\hat{\beta} \in \mathbb{R}^p$

Rem: from now on, we assume normalized atoms, *e.g.*, $\|\mathbf{x}_j\|^2 = 1, n$

Outline

Motivation for sparsity

Finding a **sparse** $\hat{\beta}$ (with only a few non-zero coefficients) :

- useful for interpretation (*e.g.*, genomics)
- useful for computational efficiency when p large. Can help either at training or at predicting (*e.g.*, on-line advertising)

Underlying goal/idea : **variable selection**

Successful applications :

- Dictionary learning, *e.g.*, image processing Mairal *et al.* (2010)
- bio-statistics Haury *et al.* (2012)
- medical imaging Lustig *et al.* (2007), Gramfort *et al.* (2012)
- etc.

Variable Selection : many variants

- **Screening** methods : correlation-screening
- **Greedy** methods : forward/stage-wise, forward-backward
- **Penalized** methods
  - convex (main focus for today and tomorrow !)
  - non-convex
- **Tree-based** methods Breiman(2001)
- **Approximate Message Passing** (AMP) methods Donoho *et al.* (2009)

Rem: last two points not developed here

Screening rules

Screening (aka correlation screening) : remove the $\mathbf{x}_j$'s weakly correlated with y (either w.r.t to a threshold or as a fixed proportion) Fan and Lv (2008)

Screening rules : « if $|\langle \mathbf{x}_j, y\rangle| = |\mathbf{x}_j^\top y| < \tau$, then remove $\mathbf{x}_j$ »

- pros :
  - fast $(+++)$
  - light computation : p inner products $(++)$
  - intuitive $(+++)$
- cons :
  - neglect variables interactions between $\mathbf{x}_j's$ $(---)$
  - weak theoretical results $(--)$

Rem: we will revisit screening rules tomorrow

Greedy methods

Many variants : Efroymson (1960), Mallat and Zhang (1993) :

- forward stage-wise = Matching Pursuit
- forward step-wise = Orthogonal Matching Pursuit

Initialize at zero : $\hat{\beta} = 0$
Iteratively select variable $\mathbf{x}_j$ most correlated with residual $\rho = y - X\hat{\beta}$, possibly perform least square on selected variables

---

- pros :
  - fast$(++)$
  - intuitive$(++)$
- cons :
  - errors propagated to next step$(-)$
  - weak theory$(-)$

Rem: competitive theory for forward-backward Zhang (2011)

Penalized (convex) regression

Penalized convex regression is the main object of the tutorial :

- pros :
  - good theoretical control $(++)$
  - guarantees for convex problems $(++)$
- cons :
  - still slow, even for convex $(-)$
  - need to tailor algorithms for specific data constraints like images, text $(-)$

Sorrow summary in Buhlmann and van de Geer (2011)

Outline

Pseudo-norm ℓ_0

Definition : support and pseudo-norm ℓ_0

The **support** of β is the set of non-zero indexes :

$$\mathrm{supp}(\beta) = \{j \in [\![1, p]\!], \beta_j \neq 0\}$$

The ℓ_0-**pseudo norm** of $\beta \in \mathbb{R}^p$ is the number of non-zeros coefficients :

$$\|\beta\|_0 = \mathrm{card}\,\{j \in [\![1, p]\!], \beta_j \neq 0\}$$

Rem: $\|\cdot\|_0$ not a norm, $\forall t \in \mathbb{R}^*, \|t\beta\|_0 = \|\beta\|_0$
Rem: $\|\cdot\|_0$ not even convex, $\beta_1 = (1, 0, 1, \cdots, 0)$
$\beta_2 = (0, 1, 1, \cdots, 0)$ and $3 = \|\frac{\beta_1+\beta_2}{2}\|_0 \geqslant \frac{\|\beta_1\|_0+\|\beta_2\|_0}{2} = 2$

ℓ_0 penalty : the dreamed target

First try to get sparsity enforcing penalty : use ℓ_0

$$\hat{\beta}^{(\lambda)} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left(\underbrace{\frac{1}{2}\|y - X\beta\|_2^2}_{\text{data fitting}} + \underbrace{\lambda\|\beta\|_0}_{\text{regularization}} \right)$$

BEWARE this is a combinatorial problem. Exact resolution requires considering all possible supports and computing least square estimators for all of them ; there are 2^p least square to perform ! ! !

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

Rem: this is a NP-Hard problem

The Lasso and variations

Vocabulary : the "Modern least square" Candès *et al.* (2008)

- Statistics : **Lasso** Tibshirani (1996)
- Signal processing variant : **Basis Pursuit** Chen *et al.* (1998)

$$\hat{\beta}^{(\lambda)} \in \mathop{\arg\min}_{\beta \in \mathbb{R}^p} \left(\underbrace{\frac{1}{2}\|y - X\beta\|^2}_{\textbf{data fitting term}} + \underbrace{\lambda\|\beta\|_1}_{\textbf{sparsity-inducing penalty}} \right)$$

where $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$

Rem: The regularization parameter $\lambda > 0$ controls the trade-off

Rem: Convex optimization problem, can be solved with guarantees

Le Lasso : penalized point of view

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

- Limiting cases :

$$\begin{aligned} \lim_{\lambda \to 0} \hat{\beta}^{(\lambda)} &= \hat{\beta}^{\mathrm{OLS}} \\ \lim_{\lambda \to +\infty} \hat{\beta}^{(\lambda)} &= 0 \in \mathbb{R}^p \end{aligned}$$

- **Beware** : Uniqueness is not automatic, see discussion in Tibshirani (2013) (*e.g.*, when two atoms are identical)

Constrained interpretation

$$\hat{\beta}^{(\lambda)} = \arg\min_{\beta \in \mathbb{R}^p} \left(\underbrace{\frac{1}{2}\|y - X\beta\|_2^2}_{\text{data fitting}} + \underbrace{\lambda\|\beta\|_1}_{\text{regularization}} \right)$$

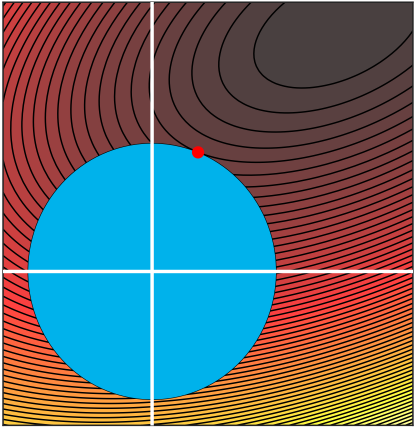
has the same solution(s) as a constrained version : for some $T > 0$

$$\begin{cases} \arg\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 \\ \text{s.t. } \|\beta\|_1 \leqslant T \end{cases}$$

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

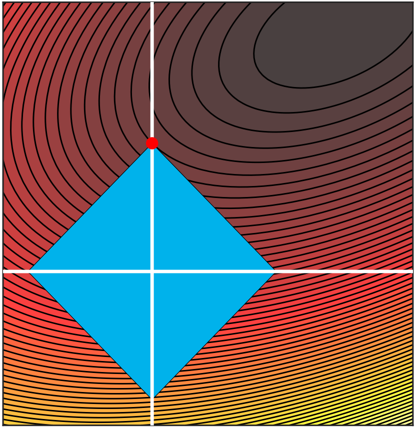
- If $T \to 0$ one finds the null-solution : $0 \in \mathbb{R}^p$
- If $T \to +\infty$ one gets $\hat{\beta}^{\mathrm{OLS}}$ (non-constrained least square)

Sparsity enforcing penalty



Ridge - ℓ_2 constraint : non-sparse solution

Sparsity enforcing penalty



Lasso - ℓ_1 constraint : sparse solution

Orthogonal case : Soft-Thresholding

Let us consider a simple **orthogonal** design : $X^\top X = \mathrm{Id}_p$

$$\|y - X\beta\|_2^2 = \|X^\top y - X^\top X\beta\|_2^2 = \|X^\top y - \beta\|_2^2$$

because X is isometric. The Lasso objectives becomes :

$$\frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 = \sum_{j=1}^{p} \left(\frac{1}{2}(\mathbf{x}_j^\top y - \beta_j)_2^2 + \lambda|\beta_j| \right)$$

Separable problem : minimize term by term the sum

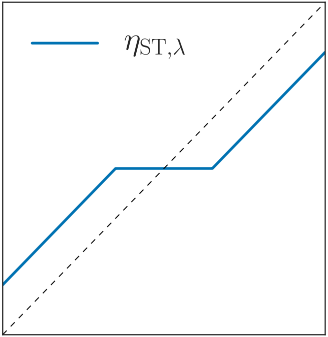
Need to solve : $\underset{x\in\mathbb{R}}{\arg\min}\, \frac{1}{2}(z - x)_2^2 + \lambda|x|$ for $z = \mathbf{x}_j^\top y$

Vocabulary : The previous solution is called the **proximal operator** at z of the function $x \mapsto \lambda|x|$ (*cf.* Parikh and Boyd (2013) or Bauschke and Combettes (2011), for more on proximal methods)

1D regularization

Problem solution : $\eta_\lambda(z) = \arg\min_{x\in\mathbb{R}} x \mapsto \frac{1}{2}(z-x)_2^2 + \lambda|x|$

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

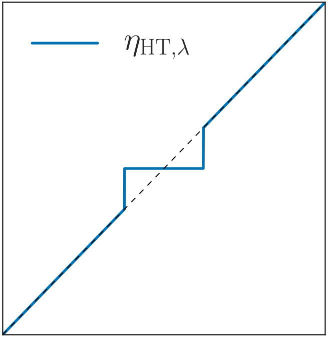


ℓ_1 : Soft Thresholding

1D regularization

Problem solution : $\eta_\lambda(z) = \arg\min_{x \in \mathbb{R}} x \mapsto \frac{1}{2}(z-x)_2^2 + \lambda \mathbb{1}_{x \neq 0}$

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



ℓ_0 : Hard Thresholding

Outline

Sub-gradients / sub-differential

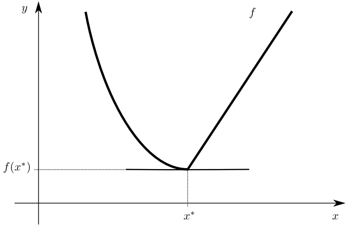
Definition : sub-gradient / sub-differential

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

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

The **sub-differential** is the set of all sub-gradients :
$\partial f(x^*) = \{u \in \mathbb{R}^d : \forall x \in \mathbb{R}^d, f(x) \geqslant f(x^*) + \langle u, x - x^* \rangle\}$.

Rem: when the sub-gradient is unique this is the standard gradient



Sub-gradients / sub-differential

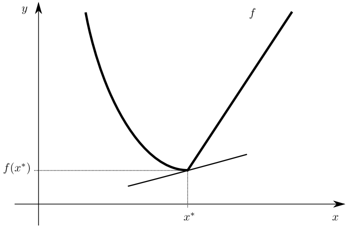
Definition : sub-gradient / sub-differential

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

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

The **sub-differential** is the set of all sub-gradients :
$\partial f(x^*) = \{u \in \mathbb{R}^d : \forall x \in \mathbb{R}^d, f(x) \geqslant f(x^*) + \langle u, x - x^* \rangle\}$.

Rem: when the sub-gradient is unique this is the standard gradient



Sub-gradients / sub-differential

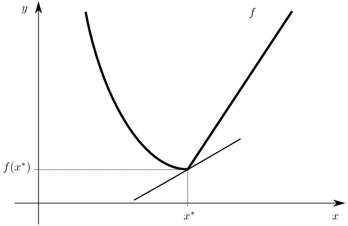
Definition : sub-gradient / sub-differential

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

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

The **sub-differential** is the set of all sub-gradients :
$\partial f(x^*) = \{u \in \mathbb{R}^d : \forall x \in \mathbb{R}^d, f(x) \geqslant f(x^*) + \langle u, x - x^* \rangle\}$.

Rem: when the sub-gradient is unique this is the standard gradient



Sub-gradients / sub-differential

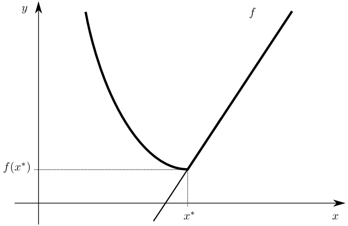
Definition : sub-gradient / sub-differential

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

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

The **sub-differential** is the set of all sub-gradients :
$\partial f(x^*) = \{u \in \mathbb{R}^d : \forall x \in \mathbb{R}^d, f(x) \geqslant f(x^*) + \langle u, x - x^* \rangle\}$.

Rem: when the sub-gradient is unique this is the standard gradient



Fermat's Rule

Theorem

A point x^* minimizes a convex function $f : \mathbb{R}^d \to \mathbb{R}$ iff $0 \in \partial f(x^*)$

Proof : use the sub-gradient definition :

- $0 \partial f(x^*)$ iff $\forall x \in \mathbb{R}^d, f(x) \geqslant f(x^*) + \langle 0, x - x^* \rangle = f(x^*)$

Fermat's Rule

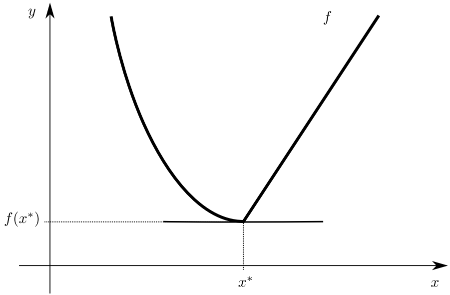
Theorem

A point x^* minimizes a convex function $f : \mathbb{R}^d \to \mathbb{R}$ iff $0 \in \partial f(x^*)$

Proof : use the sub-gradient definition :

- $0 \partial f(x^*)$ iff $\forall x \in \mathbb{R}^d, f(x) \geqslant f(x^*) + \langle 0, x - x^* \rangle = f(x^*)$

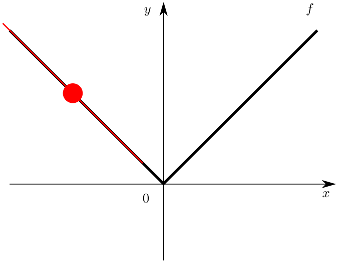
<u>Rem</u>: Visually this means a horizontal tangent is admissible



Sub-differential for the absolute value

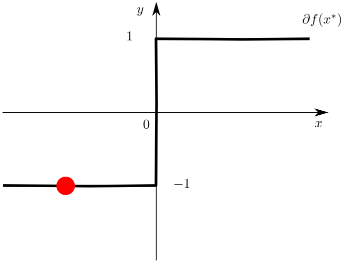
Function : abs

$$f : \begin{cases} \mathbb{R} & \longrightarrow \mathbb{R} \\ x & \longmapsto |x| \end{cases}$$



Sub-differential : sign

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



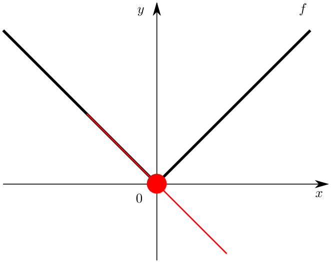
Sub-differential for the absolute value

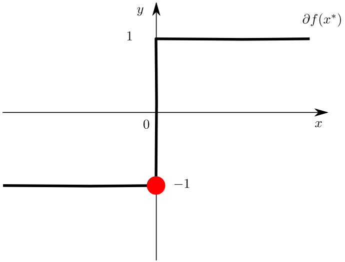
Function : abs

$$f : \begin{cases} \mathbb{R} & \longrightarrow \mathbb{R} \\ x & \longmapsto |x| \end{cases}$$

Sub-differential : sign

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

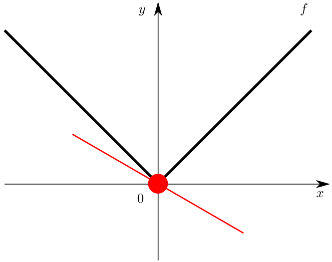




Sub-differential for the absolute value

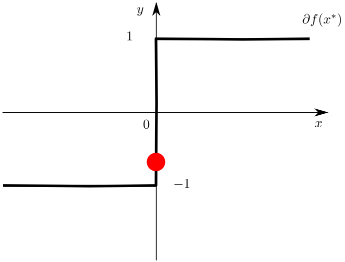
Function : abs

$$f : \begin{cases} \mathbb{R} & \longrightarrow \mathbb{R} \\ x & \longmapsto |x| \end{cases}$$



Sub-differential : sign

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



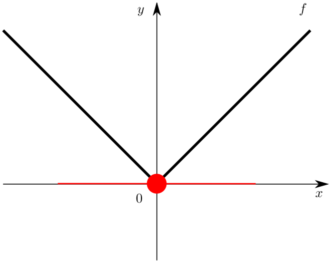
Sub-differential for the absolute value

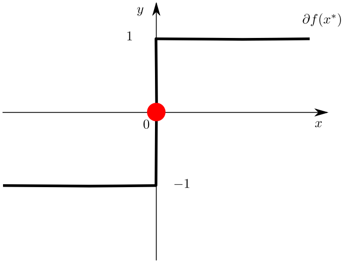
Function : abs

$$f : \begin{cases} \mathbb{R} & \longrightarrow \mathbb{R} \\ x & \longmapsto |x| \end{cases}$$

Sub-differential : sign

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

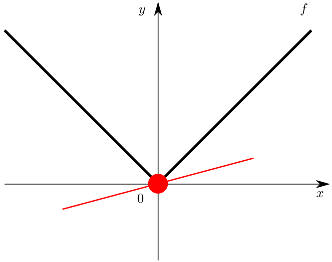




Sub-differential for the absolute value

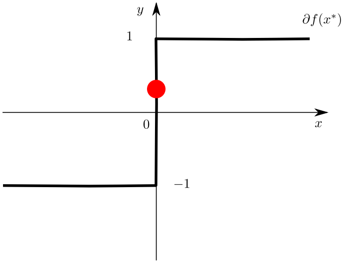
Function : abs

$$f : \begin{cases} \mathbb{R} & \longrightarrow \mathbb{R} \\ x & \longmapsto |x| \end{cases}$$



Sub-differential : sign

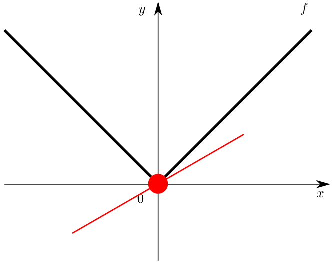
$$\partial f(x^*) = \begin{cases} \{-1\} & \text{if } x^* \in]-\infty, 0[\\ \{1\} & \text{if } x^* \in]0, +\infty[\\ [-1, 1] & \text{if } x^* = 0 \end{cases}$$



Sub-differential for the absolute value

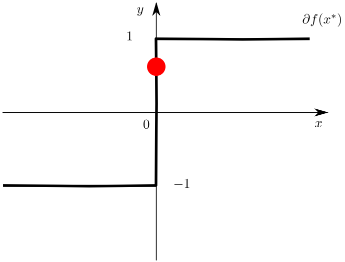
Function : abs

$$f : \begin{cases} \mathbb{R} & \longrightarrow \mathbb{R} \\ x & \longmapsto |x| \end{cases}$$



Sub-differential : sign

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



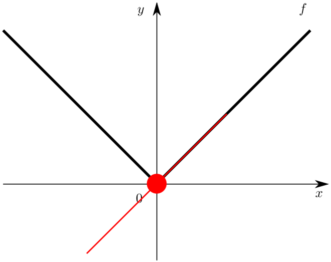
Sub-differential for the absolute value

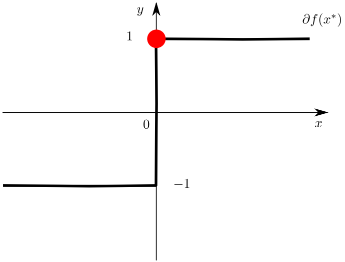
Function : abs

$$f : \begin{cases} \mathbb{R} & \longrightarrow \mathbb{R} \\ x & \longmapsto |x| \end{cases}$$

Sub-differential : sign

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





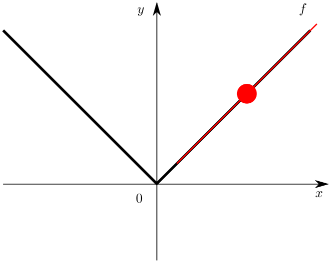
Sub-differential for the absolute value

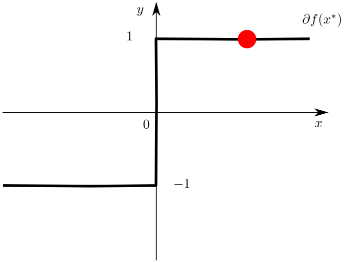
Function : abs

$$f : \begin{cases} \mathbb{R} & \longrightarrow \mathbb{R} \\ x & \longmapsto |x| \end{cases}$$

Sub-differential : sign

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





Soft thresholding through sub-gradients

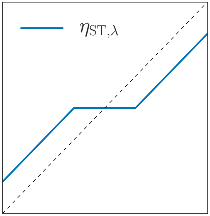
$$x^* \in \underset{x\in\mathbb{R}}{\arg\min}\, f_{\lambda,z}(x) \Leftrightarrow 0 \in \partial f_{\lambda,z}(x^*) \text{ for } f_{\lambda,z}(x) = \frac{1}{2}(z-x)_2^2 + \lambda|x|.$$

$$0 \in \partial f_{\lambda,z}(x^*) = z - x^* + \lambda\partial|\cdot|(x^*)$$
$$0 \in \partial f_{\lambda,z}(x^*) = z - x^* + \lambda\,\mathrm{sign}(x^*)$$

$$\text{So} \quad 0 \in \partial f_{\lambda,z}(x*) \Leftrightarrow x^* \in z + \lambda\,\mathrm{sign}(x)$$

Considering the cases $x^* > 0, x^* = 0, x^* < 0$ leads to :

$$\eta_{\mathrm{ST},\lambda}(z) = x^* = \begin{cases} 0 & \text{si } |z| \leqslant \lambda \\ z - \lambda & \text{si } \ z \ \geqslant \lambda \\ z + \lambda & \text{si } \ z \ \leqslant -\lambda \end{cases}$$



Fermat's Rule for the Lasso

$$\hat{\beta}^{(\lambda)} \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \quad \left(\frac{1}{2} \|y - X\beta\|_2^2 \quad + \lambda \|\beta\|_1 \right)$$

Necessary and sufficient optimality conditions (Fermat's Rule) :

$$\boxed{\forall j \in [\![1, p]\!],\ \mathbf{x}_j^\top \left(\frac{y - X\hat{\beta}^{(\lambda)}}{\lambda} \right) \in \begin{cases} \{\operatorname{sign}(\hat{\beta}^{(\lambda)})_j\} & \text{si } (\hat{\beta}^{(\lambda)})_j \neq 0, \\ [-1, 1] & \text{si } (\hat{\beta}^{(\lambda)})_j = 0. \end{cases}}$$

Rem: for OLS the **normal equation** are $\mathbf{x}_j^\top \left(y - X\hat{\beta}^{(\lambda)} \right) = 0$

Rem: There exists a **critical** value $\lambda_{\max} = \max_{j \in [\![1,p]\!]} |\langle \mathbf{x}_j, y \rangle|$ s.t.

$$\forall \lambda > \lambda_{\max},\ \hat{\beta}^{(\lambda)} = 0$$

Equi-correlation set and path properties

The set

$$E_\lambda = \{j \in [\![1, p]\!] : |\mathbf{x}_j^\top (y - X\hat{\beta}^{(\lambda)})| = \lambda\}$$

is called the **Equi-correlation** set Tibshirani (2013)

Proposition Mairal and Yu (2012)

Assume that X_{E_λ} is full rank for all $\lambda \in [\lambda_{\min}, \lambda_{\max}]$, then the Lasso solution $\hat{\beta}^{(\lambda)}$ is unique and

$$\begin{cases} [\lambda_{\min}, \lambda_{\max}] & \to \mathbb{R}^p \\ \lambda & \mapsto \hat{\beta}^{(\lambda)} \end{cases}$$

is a piecewise affine function (as a function of λ)

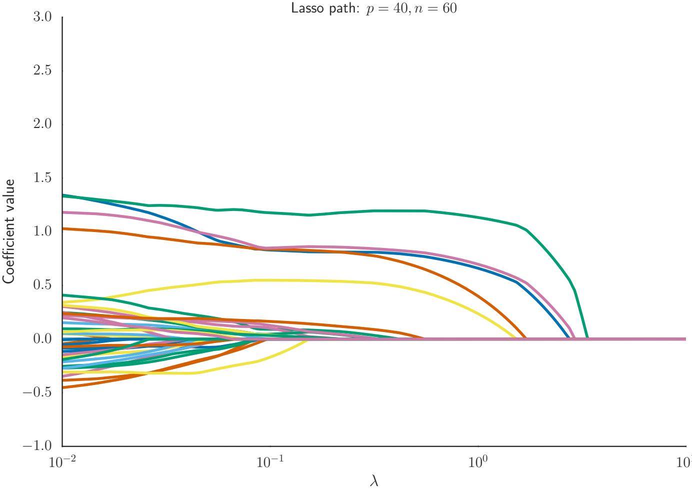
Rem: this will lead to special algorithm for solving the lasso and goes back to Osborne *et al.* (2000) and Efron *et al.* (2004)

Numerical example : simulation

Experiment settings :

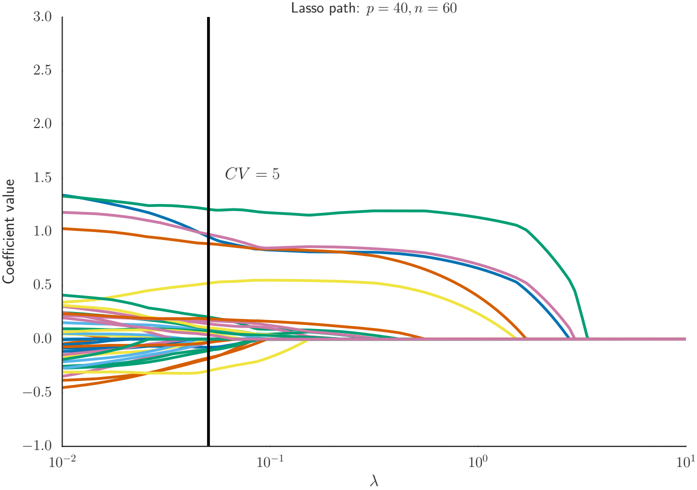
- Sizes are : $n = 60, p = 40$
- $\beta^* = (1, 1, 1, 1, 1, 0, \dots, 0) \in \mathbb{R}^p$ (5 non-zero coefficients)
- $X \in \mathbb{R}^{n \times p}$ with atoms being drawn according to a standard Gaussian distribution
- $y = X\beta^* + \varepsilon \in \mathbb{R}^n$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2 \operatorname{Id}_n)$, with $\sigma = 1$
- Using a grid of 500 values for λ

Lasso path w/o Cross-Validation



Code : `lasso_path` in `sklearn`

Lasso path w/o Cross-Validation



Code : `lasso_path` and `LassoCV` in `sklearn`

Practical interest for the Lasso

- Numerical property : the Lasso is a **convex** problem
- Variable selection / sparsity : $\hat{\beta}^{(\lambda)}$ has potentially many coefficients set to zero
- λ controls the sparsity level : if λ is large solutions are sparser (though monotonicity is sometimes not satisfied)

  Example: We obtained 25 non-zero coefficients for `LassoCV` for the previous example

Outline

The Lasso bias

Lasso bias : large coefficients shrunk toward 0 (soft-thresholding)

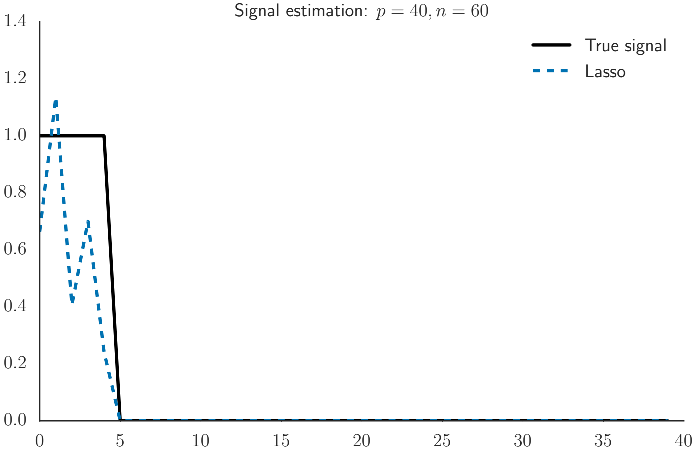


Illustration on the previous example

The Lasso bias

Lasso bias : large coefficients shrunk toward 0 (soft-thresholding)

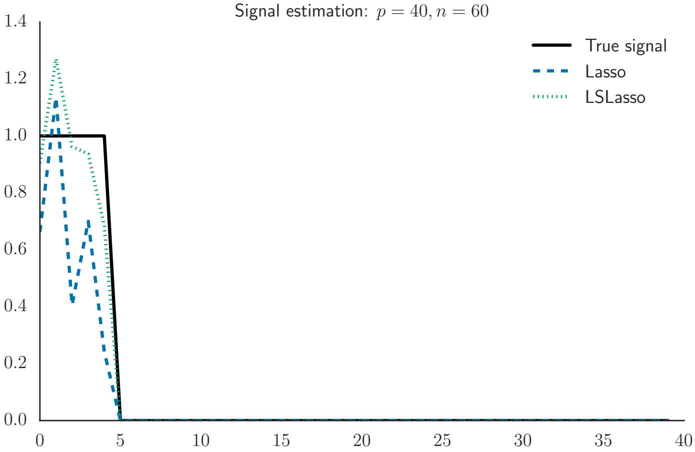


Illustration on the previous example

The Lasso bias : a simple remedy

A two-step strategy :

LSLasso (Least Square Lasso)

1. Lasso : get $\hat{\beta}^{(\lambda)}$ and its support $\mathrm{supp}(\hat{\beta}^{(\lambda)})$
2. Perform least square on the estimated support $\mathrm{supp}(\hat{\beta}^{(\lambda)})$

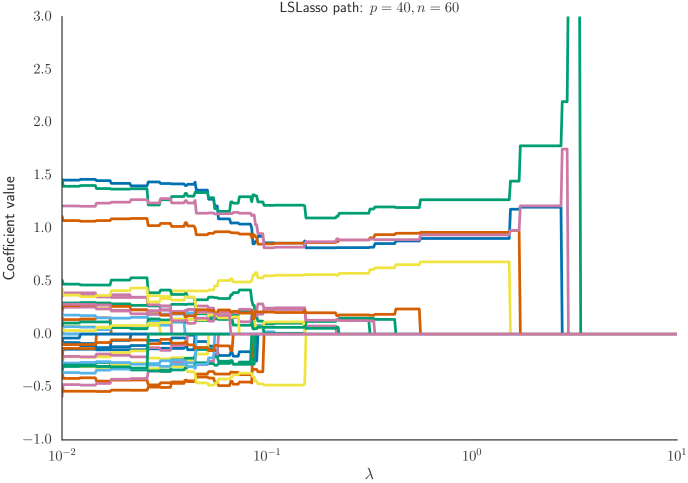
$$\hat{\beta}^{(\lambda)}_{\mathrm{LSLasso}} = \mathop{\arg\min}_{\substack{\beta \in \mathbb{R}^p \\ \mathrm{supp}(\beta) = \mathrm{supp}(\hat{\beta}^{(\lambda)})}} \frac{1}{2} \|y - X\beta\|_2^2$$

Rem: Use CV for the whole procedure ; choosing λ by CV over the Lasso and then performing least-square keeps too many variables

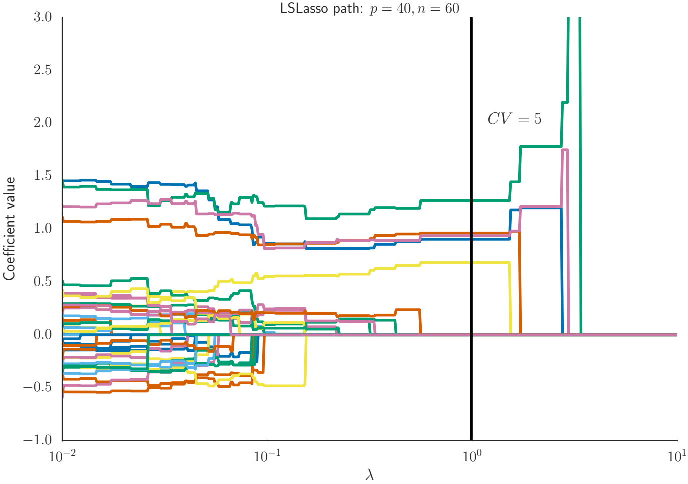
Rem: Many names : Gauss-Lasso, debiased-Lasso, LSLasso, etc.

Rem: LSLasso not usually coded in standard packages

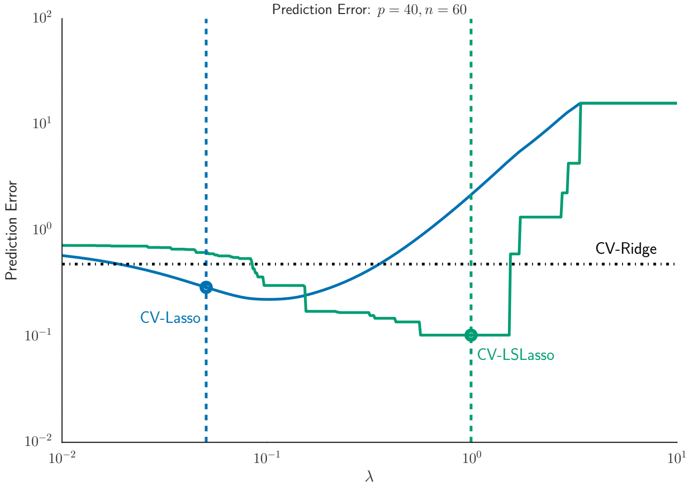
LSLasso path



LSLasso path



Prediction : Lasso vs. LSLasso



LSLasso properties

Advantages

- Large coefficients less shrunk
- Improved interpretablility : fewer "parasites" variables
  e.g., on the previous example LSLassoCV identifies correctly the 5 "true" non-zero variables

LSLasso : useful for estimation

Limitations

- In terms of prediction the difference can be small
- Need more computation : re-compute as many least squares as number of λ's considered (though with smaller sizes/supports)

Rem: procedures to perform debiasing on the fly Deledalle *et al.* (2015)

Outline

Elastic-net

Motivation : for correlated variables, the Lasso picks only one, though sharing the weights among them could be better

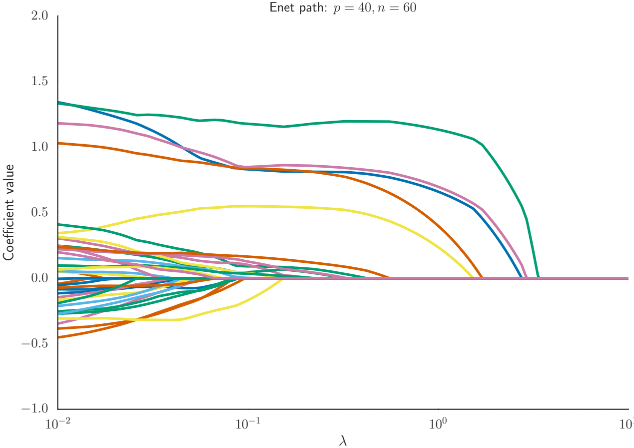
Elastic-Net Zou et Hastie (2005) is the unique solution of

$$\hat{\beta}^{(\lambda)}_{\mathrm{EN}} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left(\frac{1}{2} \|y - X\beta\|_2^2 + \lambda \left(\alpha \|\beta\|_1 + (1-\alpha) \|\beta\|_2^2 / 2 \right) \right)$$

Rem: requires two parameters — one for the global regularization, one for the trade-off between Ridge (aka Tikhonov) vs. Lasso
Rem: The Elastic-Net solution is unique

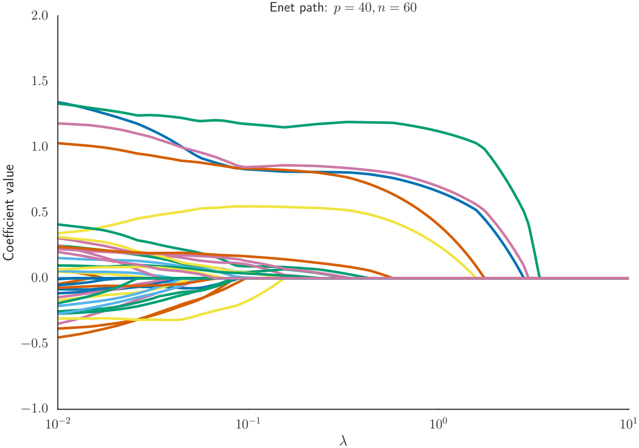
Example: Consider (normalized) $y = \mathbf{x}_1 = \mathbf{x}_2$
Lasso solutions : β with β_1 and β_2 s.t. $\beta_1 + \beta_2 = 1 - \lambda$ (for $\lambda < 1$)
Elastic- Net solution : β with $\beta_1 = \beta_2 = (1 - \lambda\alpha)/(2 + \lambda(1-\alpha))$

Elastic-Net : $\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2/2$



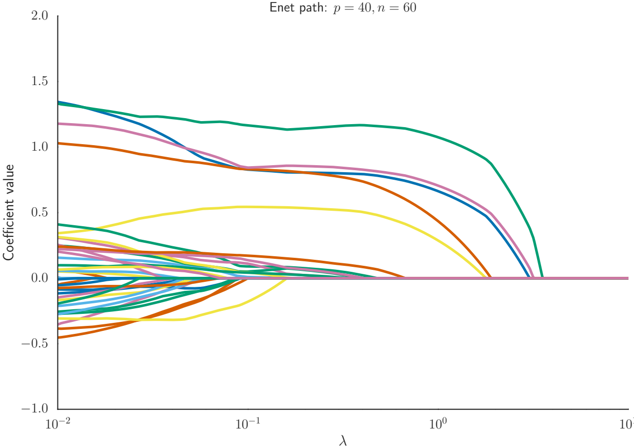
$$\alpha = 1.00$$

Elastic-Net : $\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2/2$



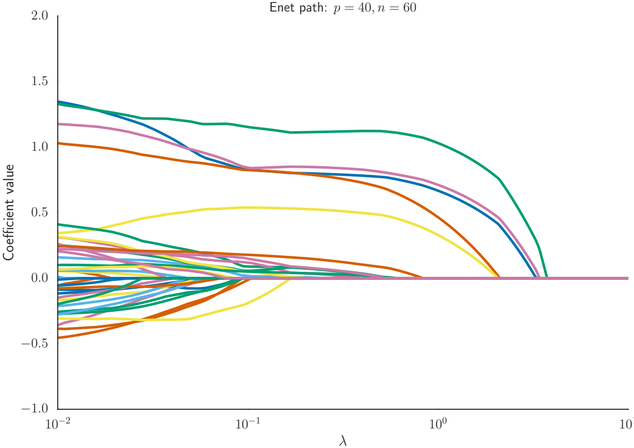
$$\alpha = 0.99$$

Elastic-Net : $\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2/2$



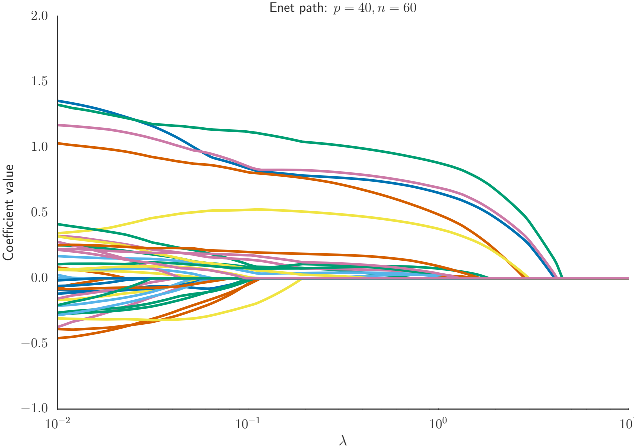
$$\alpha = 0.95$$

Elastic-Net : $\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2/2$



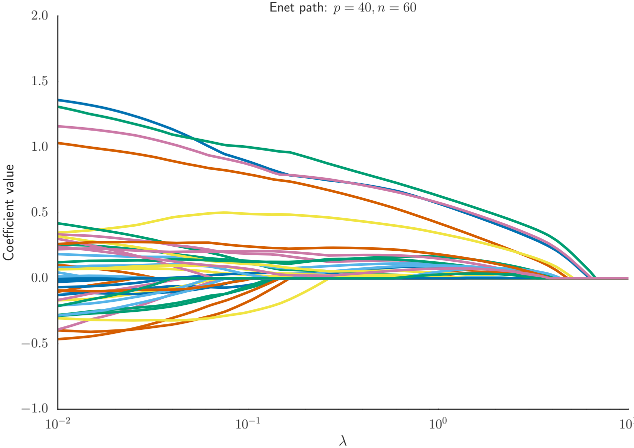
$$\alpha = 0.90$$

Elastic-Net : $\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2/2$



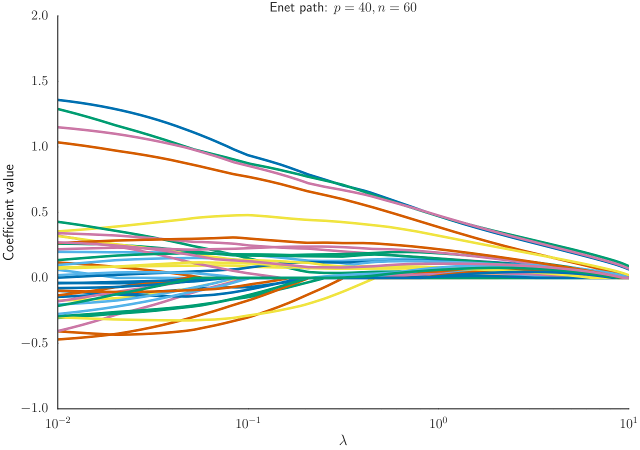
$$\alpha = 0.75$$

Elastic-Net : $\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2/2$



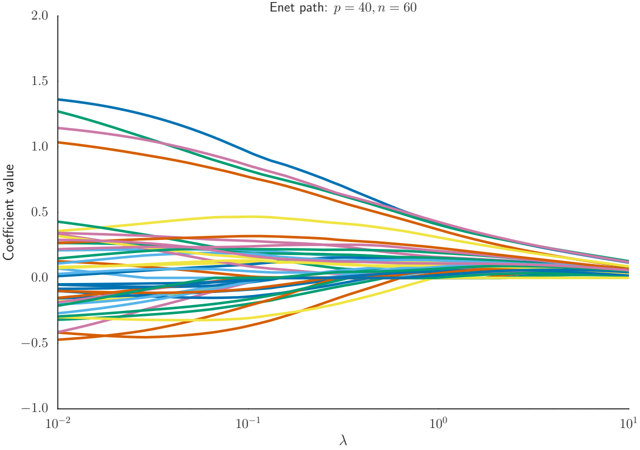
$\alpha = 0.50$

Elastic-Net : $\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2/2$



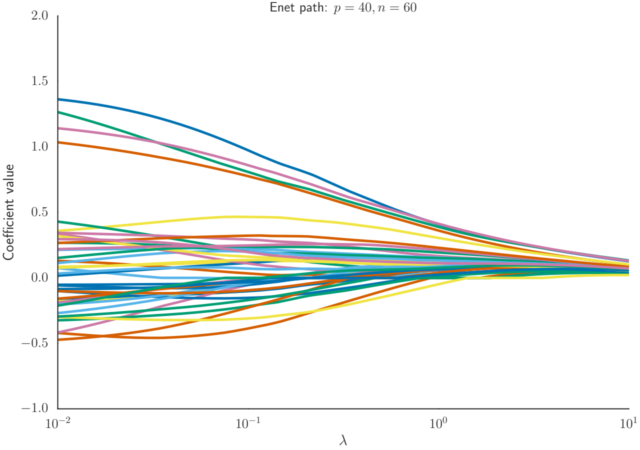
$$\alpha = 0.25$$

Elastic-Net : $\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2/2$



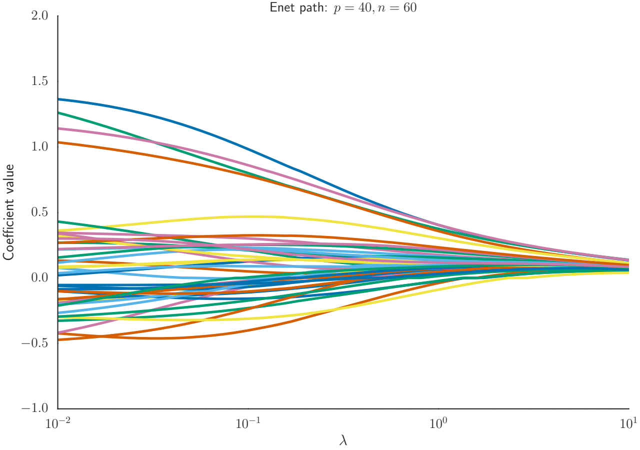
$$\alpha = 0.1$$

Elastic-Net : $\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2/2$



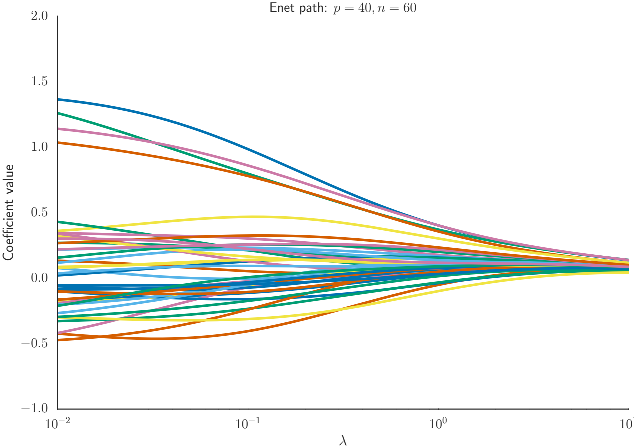
$\alpha = 0.05$

Elastic-Net : $\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2/2$



$$\alpha = 0.01$$

Elastic-Net : $\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2/2$



$$\alpha = 0.00$$

Outline

Group-Lasso

The ℓ_1 penalty ensures that few coefficients are active, but no structure on the support is enforced

We may be interested in specific sparsity patterns :

- Groups/blocks structure : Group-Lasso Yuan et Lin (2006)
- Groups/blocks + individual structure : Sparse-Group Lasso Simon *et al.* (2012)
- Hierarchical structure (*e.g.*, for higher order interactions of variables : $\mathbf{x}_j \cdot \mathbf{x}_k$) Bien *et al.* (2013)
- etc.

Sparsity patterns

We assume here that a group structure is known over the variables we investigate : $[\![1, p]\!] = \bigcup_{g \in \mathcal{G}} g$

Vector and active coefficients (in orange) :

Sparsity pattern : no structure

Penalty considered : Lasso

$\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$

Sparsity patterns

We assume here that a group structure is known over the variables we investigate : $[\![1, p]\!] = \bigcup_{g \in \mathcal{G}} g$

Vector and active coefficients (in orange) :

Sparsity pattern : groups

Penalty considered : Group-Lasso

$\|\beta\|_{2,1} = \sum_{g \in G} \|\beta_g\|_2$

Sparsity patterns

We assume here that a group structure is known over the variables we investigate : $[\![1, p]\!] = \bigcup_{g \in \mathcal{G}} g$

Vector and active coefficients (in orange) :

Sparsity pattern : groups + sub-groups

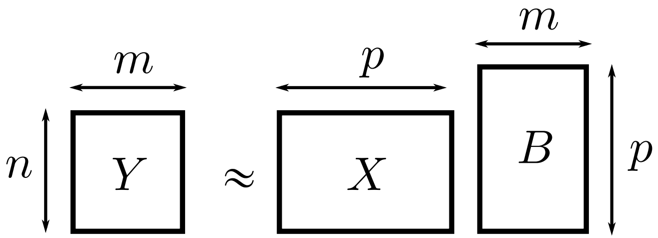
Penalty considered : Sparse-Group Lasso

$$\alpha \|\beta\|_1 + (1-\alpha)\|\beta\|_{2,1} = \alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{g \in G} \|\beta_g\|_2$$

Outline

Multivariate / Multi-task regression

Aim : solving m (tasks) linear regression jointly : $Y \approx XB$



- $Y \in \mathbb{R}^{n \times m}$: observations matrix
- $X \in \mathbb{R}^{n \times p}$: design matrix (shared)
- $B \in \mathbb{R}^{p \times m}$: coefficients matrix

Example: several signals are observed during a time slot, *e.g.*, various sensors for the same phenomenon

Rem: *cf.* `MultiTaskLasso` in `sklearn`

Penalized least-square for multi-task regression

For multi-task one can regularize the least square :

$$\hat{B}_\lambda = \operatorname*{arg\,min}_{B \in \mathbb{R}^{p\times m}} \left(\underbrace{\frac{1}{2}\| Y - XB\|_F^2}_{\text{data fitting}} + \underbrace{\lambda\Omega(B)}_{\text{regularization}} \right)$$

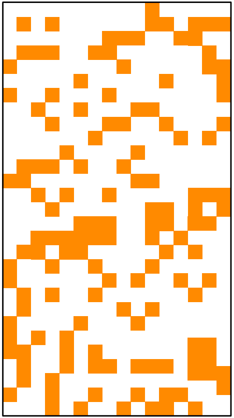
Ω is a penalty term to be specified (to enforce sparsity)

Rem: the Frobenius norm $\|\cdot\|_F$ is defined for any matrix $A \in \mathbb{R}^{n_1 \times n_2}$:

$$\|A\|_F^2 = \sum_{j_1=1}^{n_1} \sum_{j_2=1}^{n_2} A_{j_1,j_2}^2$$

Multi-task penalties

Vector penalties need to be adapted for matrices :



B Parameter

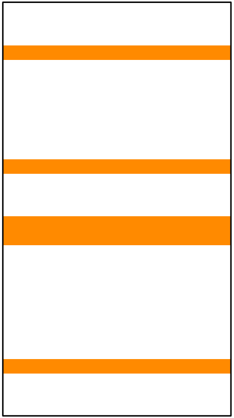
Sparse matrix :
unstructured

Lasso :

$$\|B\|_1 = \sum_{j=1}^{p} \sum_{k=1}^{m} |B_{j,k}|$$

Multi-task penalties

Vector penalties need to be adapted for matrices :



B Parameter

Sparse matrix :
groups

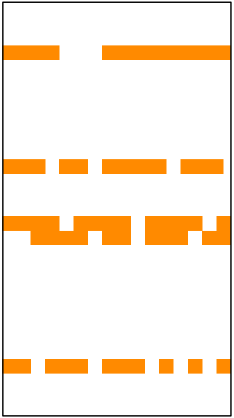
Group-Lasso :

$$\|B\|_{2,1} = \sum_{j=1}^{p} \|B_{j:}\|_2$$

Rem: $B_{j,:}$ is the j^{th} line of B

Multi-task penalties

Vector penalties need to be adapted for matrices :



B Parameter

Sparse matrix :
groups + sub-groups

Sparse-Group Lasso :

$$\alpha \|B\|_1 \ + \ (1 \ - \ \alpha)\|B\|_{2,1}$$

Logistic regression - Generalized Linear Model

Other data-fitting terms : Generalized Linear Model (GLM)
Motivation : other noise like Poisson, Laplace, etc. or different problem like classification

Logistic regression (binary case)

One observes for each $i \in [\![1, n]\!]$, a class label $c_i \in \{1, 2\}$, so the observations can be recast as $y_i = \mathbb{1}_{\{c_i=1\}}$. Then, the data-fitting term considered is

$$f(\beta) = \sum_{i=1}^{n} \left(-y_i X_{i,:}\beta + \log\left(1 + \exp\left(X_{i,:}\beta\right)\right)\right),$$

instead of the least square term $f(\beta) = \|y - X\beta\|_2^2/2$, see for instance Buhlmann and van de Geer (2011), Ch. 3

Outline

Coordinate descent description

$$\text{Objective : solve } \arg\min_{\beta \in \mathbb{R}^p} f(\beta)$$

Initialization : $\beta^{(0)}$
While not converged

Coordinate descent description

Objective : solve $\arg\min_{\beta \in \mathbb{R}^p} f(\beta)$

Initialization : $\beta^{(0)}$

While not converged

$$\beta_1^{(k)} \in \arg\min_{\beta_1 \in \mathbb{R}} f(\beta_1, \beta_2^{(k-1)}, \beta_3^{(k-1)} \dots, \beta_p^{(k-1)})$$

Coordinate descent description

Objective : solve $\arg\min_{\beta \in \mathbb{R}^p} f(\beta)$

Initialization : $\beta^{(0)}$
While not converged

$$\beta_1^{(k)} \in \arg\min_{\beta_1 \in \mathbb{R}} f(\beta_1, \beta_2^{(k-1)}, \beta_3^{(k-1)} \dots, \beta_p^{(k-1)})$$

$$\beta_2^{(k)} \in \arg\min_{\beta_2 \in \mathbb{R}} f(\beta_1^{(k)}, \beta_2, \beta_3^{(k-1)}, \dots, \beta_p^{(k-1)})$$

Coordinate descent description

Objective : solve $\arg\min_{\beta \in \mathbb{R}^p} f(\beta)$

Initialization : $\beta^{(0)}$
While not converged

$$\beta_1^{(k)} \in \arg\min_{\beta_1 \in \mathbb{R}} f(\beta_1, \beta_2^{(k-1)}, \beta_3^{(k-1)} \dots, \beta_p^{(k-1)})$$

$$\beta_2^{(k)} \in \arg\min_{\beta_2 \in \mathbb{R}} f(\beta_1^{(k)}, \beta_2, \beta_3^{(k-1)}, \dots, \beta_p^{(k-1)})$$

$$\beta_3^{(k)} \in \arg\min_{\beta_3 \in \mathbb{R}} f(\beta^{(k)}, \beta_2^{(k)}, \beta_3, \dots, \beta_p^{(k-1)})$$

Coordinate descent description

Objective : solve $\displaystyle\arg\min_{\beta \in \mathbb{R}^p} f(\beta)$

Initialization : $\beta^{(0)}$

While not converged

$$\beta_1^{(k)} \in \arg\min_{\beta_1 \in \mathbb{R}} f(\beta_1, \beta_2^{(k-1)}, \beta_3^{(k-1)} \dots, \beta_p^{(k-1)})$$

$$\beta_2^{(k)} \in \arg\min_{\beta_2 \in \mathbb{R}} f(\beta_1^{(k)}, \beta_2, \beta_3^{(k-1)}, \dots, \beta_p^{(k-1)})$$

$$\beta_3^{(k)} \in \arg\min_{\beta_3 \in \mathbb{R}} f(\beta^{(k)}, \beta_2^{(k)}, \beta_3, \dots, \beta_p^{(k-1)})$$

$$\vdots$$

$$\beta_p^{(k)} \in \arg\min_{\beta_p \in \mathbb{R}} f(\beta_1^{(k)}, \beta_2^{(k)}, \beta_3^{(k)}, \dots, \beta_p)$$

Coordinate descent description

Objective : solve $\underset{\beta \in \mathbb{R}^p}{\arg\min} f(\beta)$

Initialization : $\beta^{(0)}$

While not converged

$$\beta_1^{(k)} \in \underset{\beta_1 \in \mathbb{R}}{\arg\min} f(\beta_1, \beta_2^{(k-1)}, \beta_3^{(k-1)} \ldots, \beta_p^{(k-1)})$$

$$\beta_2^{(k)} \in \underset{\beta_2 \in \mathbb{R}}{\arg\min} f(\beta_1^{(k)}, \beta_2, \beta_3^{(k-1)}, \ldots, \beta_p^{(k-1)})$$

$$\beta_3^{(k)} \in \underset{\beta_3 \in \mathbb{R}}{\arg\min} f(\beta^{(k)}, \beta_2^{(k)}, \beta_3, \ldots, \beta_p^{(k-1)})$$

$$\vdots$$

$$\beta_p^{(k)} \in \underset{\beta_p \in \mathbb{R}}{\arg\min} f(\beta_1^{(k)}, \beta_2^{(k)}, \beta_3^{(k)}, \ldots, \beta_p)$$

$$k := k + 1$$

Motivation

- Coordinate descent can be very fast, especially if the design X is unstructured and sparse (otherwise see Forward-Backward)
- Convergence toward a minimum is guaranteed (for smooth or separable non-smooth functions *cf.* Tseng (2001))
- can visit the coordinate cyclically, randomly, etc.
- sometimes referred to as block methods : same idea but update a block of coordinates

Lasso : coordinate descent

$$\arg\min_{\beta \in \mathbb{R}^p} f(\beta) \text{ for } f(\beta) = \frac{1}{2}\|y - X\beta\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Minimize w.r.t β_j keeping β_k's $(k \neq j)$ fixed :

$$\begin{aligned}
\hat{\beta}_j &= \arg\min_{\beta_j \in \mathbb{R}} f(\beta_1, \cdots, \beta_p) \\
&= \arg\min_{\beta_j \in \mathbb{R}} \frac{1}{2}\|y - \sum_{k \neq j} \beta_k \mathbf{x}_k - \mathbf{x}_j \beta_j\|^2 + \lambda \sum_{k \neq j} |\beta_j| + \lambda |\beta_j| \\
&= \arg\min_{\beta_j \in \mathbb{R}} \frac{1}{2}\|\mathbf{x}_j\|^2 \beta_j^2 - \langle y - \sum_{k \neq j} \beta_k \mathbf{x}_k, \mathbf{x}_j \rangle \beta_j + \lambda |\beta_j| \\
&= \arg\min_{\beta_j \in \mathbb{R}} \|\mathbf{x}_j\|^2 \left[\frac{1}{2} \left(\beta_j - \|\mathbf{x}_j\|^{-2} \langle y - \sum_{k \neq j} \beta_k \mathbf{x}_k, \mathbf{x}_j \rangle \right)^2 + \frac{\lambda}{\|\mathbf{x}_j\|^2} |\beta_j| \right]
\end{aligned}$$

<u>Reminder</u> : $\eta_{\mathrm{ST},\lambda}(z) = \arg\min_{x \in \mathbb{R}} x \mapsto \frac{1}{2}(z - x)^2 + \lambda |x|$

Lasso : coordinate descent (II)

<u>Solution</u> : $$\hat{\beta}_j = \eta_{\mathrm{ST},\lambda/\|\mathbf{x}_\mathrm{j}\|^2}\left(\|\mathbf{x}_j\|^{-2}\langle y - \sum_{k\neq j}\beta_k \mathbf{x}_k, \mathbf{x}_j\rangle\right)$$

Initialize : parameter $\beta = 0 \in \mathbb{R}^p$, residual $\rho = y \in \mathbb{R}^n$
While not converged, pick $j \in [\![1, p]\!]$ and perform :

$$\begin{aligned}\rho^{\mathrm{int}} &\leftarrow \rho + \mathbf{x}_j\beta_j \\ \beta_j &\leftarrow \eta_{\mathrm{ST},\lambda/\|\mathbf{x}_\mathrm{j}\|^2}\left(\mathbf{x}_j^\top \rho^{\mathrm{int}}/\|\mathbf{x}_j\|^2\right) \\ \rho &\leftarrow \rho^{\mathrm{int}} - \mathbf{x}_j\beta_j\end{aligned}$$

<u>Rem</u>: again, pick coordinates cyclically or (uniformly) at random
<u>Rem</u>: low memory impact storing ρ and β
<u>Rem</u>: interesting to choose $\|\mathbf{x}_j\|_2^2 = 1$

Outline

Variable selection and sparsity
- Motivation and variable selection variants
- ℓ_0 and ℓ_1 penalties
- Sub-gradients / sub-differential

Lasso extensions and improvements
- LSLasso : Least-Square Lasso
- Lasso variants : Elastic Net
- Group structure
- Multivariate / Multi-task regression

Optimization for the Lasso
- Coordinate descent
- Proximal methods — Forward / Backward

Theoretical results for the Lasso
- Prediction error
- Estimation error

Composite minimization

One aims at minimizing : $F = f + g$

Rem: for the Lasso $f(\beta) = \|X\beta - y\|_2^2/2$ and $g = \lambda\|\beta\|_1$

- f smooth : often meaning ∇f is L-Lipschitz
- g proximable (prox-capable) : prox_g can be "efficiently" computed, where

$$\boxed{\mathrm{prox}_g(w) = \arg\min_{z \in \mathbb{R}^p} \left(\frac{1}{2}\|z - w\|_2^2 + g(z) \right)}$$

More details on prox properties in Parikh and Boyd (2013)

Examples of proximity operators

$$\boxed{\operatorname{prox}_g(w) = \operatorname*{arg\,min}_{z \in \mathbb{R}^p} \left(\frac{1}{2}\|z - w\|_2^2 + g(z) \right)}$$

- Null function : if $g = 0$, then $\operatorname{prox}_g = \mathrm{Id}$
- Indicator function : $g = \iota_C$ for a closed convex set $C \subset \mathbb{R}^p$,

$$\operatorname{prox}_g = \pi_C, \quad \text{projection over the set } C$$

- Soft-Thresholding : $g = \lambda|\cdot|$ (*i.e.*, $p = 1$ here), then

$$\operatorname{prox}_g(w) = \eta_{\mathrm{ST},\lambda}(w) = \operatorname{sign}(w)(|w| - \lambda)_+$$

- Vector Soft-Thresholding : $g = \lambda\|\cdot\|_1$, then

$$\operatorname{prox}_g(w) = (\eta_{\mathrm{ST},\lambda}(w_1), \ldots, \eta_{\mathrm{ST},\lambda}(w_1))^\top$$

Forward-Backward / Iterative Soft Thresholding

Extension of gradient descent for a sum of functions :

General Forward-Backward

Choose step size value : α
Initialization : $\beta = 0 \in \mathbb{R}^p$
While not converged
$\beta \leftarrow \mathrm{prox}_{\alpha g}\left(\beta - \alpha \nabla f(\beta)\right)$

Forward-Backward / Iterative Soft Thresholding

Extension of gradient descent for a sum of functions :

General Forward-Backward	Iterative Soft-thresholding
Choose step size value : α	Choose step size value : α
Initialization : $\beta = 0 \in \mathbb{R}^p$	Initialization : $\beta = 0 \in \mathbb{R}^p$
While not converged	While not converged
$\beta \leftarrow \mathrm{prox}_{\alpha g}\left(\beta - \alpha \nabla f(\beta)\right)$	$\beta \leftarrow \eta_{\mathrm{ST},\alpha\lambda}\left(\beta + \alpha X^\top (y - X\beta)\right)$

Forward-Backward / Iterative Soft Thresholding

Extension of gradient descent for a sum of functions :

General Forward-Backward	Iterative Soft-thresholding
Choose step size value : α	Choose step size value : α
Initialization : $\beta = 0 \in \mathbb{R}^p$	Initialization : $\beta = 0 \in \mathbb{R}^p$
While not converged	While not converged
$\beta \leftarrow \mathrm{prox}_{\alpha g}\left(\beta - \alpha \nabla f(\beta)\right)$	$\beta \leftarrow \eta_{\mathrm{ST},\alpha\lambda}\left(\beta + \alpha X^\top (y - X\beta)\right)$

Rem: Majorization-minimization : if $\alpha \leqslant 1/L$ one has a quadratic majorant, and the prox step consists in solving

$$\operatorname*{arg\,min}_{\beta' \in \mathbb{R}^p} \left(f(\beta) + \langle \nabla f(\beta), \beta' - \beta \rangle + \frac{1}{2\alpha} \|\beta' - \beta\|^2 + g(\beta') \right)$$

Forward-Backward / Iterative Soft Thresholding (II)

- Interesting when the operator $z \mapsto X^\top z$ can be performed efficiently : often the case in imaging, *e.g.*, for FFT, Wavelet transforms, etc.
- Requires α to be tuned/chosen : default is often $\alpha = 1/L = 1/\mu_{\max}(X^\top X)$ (spectral radius of $X^\top X$)
- Common acceleration : Fast Iterative Soft Thresholding Algorithm (FISTA) Nesterov (1983), Beck and Teboulle (2009)

Homotopy methods for the Lasso

Family of algorithms introduced by Osborne *et al.* (2000) ; the most famous variant is called LARS Efron *et al.* (2004)

It leverages the piecewise affine property of the Lasso w.r.t λ and least squares computation

- pros :
  - Provide all solutions up to interpolation
  - Only finite number of kinks computed
- cons :
  - Not stable for small λ's
  - can produce many solutions, up to $O((3^p + 1)/2)$
  - Do not generalize to group, logistic, etc.

cf. Mairal and Yu (2012) for more details on Lasso homotopy

Outline

Theoretical analysis of the lasso

Results require (hard to check) assumptions on the design X :

- Prediction bounds Bickel *et al.* (2009) : controlling $\|X\hat{\beta}^{(\lambda)} - X\beta^*\|_2^2$
- Estimation bounds Bickel *et al.* (2009), Wainwright (2009) : controlling $\|\hat{\beta}^{(\lambda)} - \beta^*\|_\infty$ or $\|\hat{\beta}^{(\lambda)} - \beta^*\|_2$
- Support/sign recovery Lounici (2008) : controls when $\mathrm{sign}(\hat{\beta}^{(\lambda)}) = \mathrm{sign}(\beta^*)$ or $\mathrm{supp}(\hat{\beta}^{(\lambda)}) = \mathrm{supp}(\beta^*)$

Rem: the control could be in expectation or with high probability

Rem: large volume of literature on this field, hard to be exhaustive
A good book for this is *cf.* Buhlmann et van de Geer (2011)

Prediction error for the Lasso

Take away message : optimal prediction error (minimax sense)

Theorem Bickel *et al.* (2009)

Assume the noise is Gaussian and the atoms are normalized s.t. $\|\mathbf{x}_j\|_2^2 = n$, then for $\lambda > c_1 \sigma \sqrt{n \log(p)}$ the following holds with high probability :

$$\|X\hat{\beta}^{(\lambda)} - X\beta^*\|_2^2/n \leqslant c_X \sigma^2 \frac{\|\beta^*\|_0 \log(p)}{n}$$

where c_X is a constant depending on the design matrix X

Rem: the $\log(p)$ term is the price to pay for not knowing $\operatorname{supp}(\beta^*)$
Rem: the assumption needed on the design so that $c_X > 0$ is not computationally checkable but are satisfied for random matrices

Outline

Estimation and support recovery for the Lasso

Take away message : the Lasso recovers the true support with high probability

For this result to hold, similar assumptions on the design matrix are needed, but **more** is needed Wainwright (2009) :

The true support $\mathrm{supp}(\beta^*)$ needs to be well separated from zero, otherwise some variables might be missing : they could be interpreted as noise fluctuations

$$\min_{j \in \mathrm{supp}(\beta^*)} |\beta_j^*| > c\sigma\sqrt{n \log(p)}$$

Rem: the sign vector might also be recovered w.h.p
Rem: results for a thresholded Lasso estimator Lounici (2008)

Conclusion

Lasso and variants properties :

- Lasso introduces sparsity (and possibly bias)
- Introduction to non-smooth optimization
- Extension to (partially) reduce bias
- Convex algorithms to solve ℓ_1 type regularization

Points not addressed :

- Parameter(s) tuning : Cross Validation and variants such as Bolasso Bach (2008) or Stability Selection Meinshausen et Buhlmann (2010)
- Noise estimation : $\sqrt{\text{Lasso}}$ Belloni *et al.* (2011), Scaled Lasso Zhang and Zhang (2012)
- Non-convex penalties : *e.g.*, SCAD Fan and Li (2002), Adaptive-Lasso Zou (2006), reweighted ℓ_1 Candès *et al.* (2008), etc.

References I

- F. Bach.
  Bolasso : model consistent Lasso estimation through the bootstrap.
  In *ICML*, 2008.
- H. H. Bauschke and P. L. Combettes.
  Convex analysis and monotone operator theory in Hilbert spaces.
  Springer, New York, 2011.
- A. Belloni, V. Chernozhukov, and L. Wang.
  Square-root Lasso : Pivotal recovery of sparse signals via conic programming.
  Biometrika, 98(4) :791–806, 2011.
- L. Breiman.
  Random Forests.
  Mach. Learn., 45(1) :5–32, 2001.
- P. J. Bickel, Y. Ritov, and A. B. Tsybakov.
  Simultaneous analysis of Lasso and Dantzig selector.
  Ann. Statist., 37(4) :1705–1732, 2009.
- A. Beck and M. Teboulle.
  A fast iterative shrinkage-thresholding algorithm for linear inverse problems.
  SIAM J. Imaging Sci., 2(1) :183–202, 2009.

References II

- P. Bühlmann and S. van de Geer.
  Statistics for high-dimensional data.
  Springer Series in Statistics. Springer, Heidelberg, 2011.
  Methods, theory and applications.
- S. S. Chen, D. L. Donoho, and M. A. Saunders.
  Atomic decomposition by basis pursuit.
  SIAM J. Sci. Comput., 20(1) :33–61 (electronic), 1998.
- E. J. Candès, M. B. Wakin, and S. P. Boyd.
  Enhancing sparsity by reweighted l_1 minimization.
  J. Fourier Anal. Applicat., 14(5-6) :877–905, 2008.
- D. L. Donoho, A., and A. Montanari.
  Message-passing algorithms for compressed sensing.
  Proceedings of the National Academy of Sciences, 106(45) :18914–18919, 2009.
- C-A. Deledalle, N. Papadakis, and J. Salmon.
  On debiasing restoration algorithms : applications to total-variation and nonlocal-means.
  In *SSVM*, 2015.
- M. A. Efroymson.
  Multiple regression analysis.
  In *Mathematical methods for digital computers*, pages 191–203. Wiley, New York, 1960.

References III

- B. Efron, T. Hastie, I. M. Johnstone, and R. Tibshirani.
  Least angle regression.
  Ann. Statist., 32(2) :407–499, 2004.
  With discussion, and a rejoinder by the authors.
- J. Fan and R. Li.
  Variable selection via nonconcave penalized likelihood and its oracle properties.
  J. Amer. Statist. Assoc., 96(456) :1348–1360, 2001.
- J. Fan and J. Lv.
  Sure independence screening for ultrahigh dimensional feature space.
  J. Roy. Statist. Soc. Ser. B, 70(5) :849–911, 2008.
- A. Gramfort, M. Kowalski, and M. Hämäläinen.
  Mixed-norm estimates for the M/EEG inverse problem using accelerated gradient methods.
  Physics in Medicine and Biology, 57(7) :1937–1961, 2012.
- A.-C. Haury, F. Mordelet, P. Vera-Licona, and J.-P. Vert.
  TIGRESS : Trustful Inference of Gene REgulation using Stability Selection.
  BMC systems biology, 6(1) :145, 2012.
- Bien J, J. Taylor, and R. Tibshirani.
  A lasso for hierarchical interactions.
  Ann. Statist., 41(3) :1111–1141, 2013.

References IV

- M. Lustig, D. L. Donoho, and J. M. Pauly.
  Sparse MRI : The application of compressed sensing for rapid MR imaging.
  Magnetic Resonance in Medicine, 58(6) :1182–1195, 2007.
- K. Lounici.
  Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators.
  Electron. J. Stat., 2 :90–102, 2008.
- N. Meinshausen and P. Bühlmann.
  Stability selection.
  J. Roy. Statist. Soc. Ser. B, 72(4) :417–473, 2010.
- J. Mairal, F. Bach, J. Ponce, and G. Sapiro.
  Online learning for matrix factorization and sparse coding.
  J. Mach. Learn. Res., pages 19–60, 2010.
- J. Mairal and B. Yu.
  Complexity analysis of the lasso regularization path.
  In *ICML*, 2012.
- S. Mallat and Z. Zhang.
  Matching pursuit with time-frequency dictionaries.
  IEEE Trans. Image Process., 41 :3397–3415, 1993.

References V

- Y. Nesterov.
  A method for solving a convex programming problem with rate of convergence $O(1/k^2)$.
  Soviet Math. Doklady, 269(3) :543–547, 1983.
- M. R. Osborne, B. Presnell, and B. A. Turlach.
  A new approach to variable selection in least squares problems.
  IMA J. Numer. Anal., 20(3) :389–403, 2000.
- N. Parikh, S. Boyd, E. Chu, B. Peleato, and J. Eckstein.
  Proximal algorithms.
  Foundations and Trends in Machine Learning, 1(3) :1–108, 2013.
- N. Simon, J. Friedman, T. Hastie, and R. Tibshirani.
  A sparse-group lasso.
  J. Comput. Graph. Statist., 22(2) :231–245, 2013.
- R. Tibshirani.
  Regression shrinkage and selection via the lasso.
  J. Roy. Statist. Soc. Ser. B, 58(1) :267–288, 1996.
- R. J. Tibshirani.
  The lasso problem and uniqueness.
  Electron. J. Stat., 7 :1456–1490, 2013.

References VI

- P. Tseng.
  Convergence of a block coordinate descent method for nondifferentiable minimization.
  J. Optim. Theory Appl., 109(3) :475–494, 2001.
- M. Wainwright.
  Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-Constrained quadratic programming (lasso).
  IEEE Trans. Inf. Theory, 55(5) :2183–2202, 2009.
- M. Yuan and Y. Lin.
  Model selection and estimation in regression with grouped variables.
  J. Roy. Statist. Soc. Ser. B, 68(1) :49–67, 2006.
- H. Zou and T. Hastie.
  Regularization and variable selection via the elastic net.
  J. Roy. Statist. Soc. Ser. B, 67(2) :301–320, 2005.
- T. Zhang.
  Adaptive forward-backward greedy algorithm for learning sparse representations.
  IEEE Trans. Inf. Theory, 57(7) :4689–4708, 2011.
- H. Zou.
  The adaptive lasso and its oracle properties.
  J. Am. Statist. Assoc., 101(476) :1418–1429, 2006.

References VII

- C.-H. Zhang and T. Zhang.
  A general theory of concave regularization for high-dimensional sparse estimation problems.
  Statistical Science, 27(4) :576–593, 2012.