High-Dimensional Statistics

Christophe Giraud

Université Paris-Sud et Ecole Polytechnique

M2 DS & ISG

Lasso estimators

Reminder on the Lasso

The lasso estimator is defined by

$$\widehat{\beta}_\lambda \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \mathcal{L}_\lambda(\beta) \quad \text{where} \quad \mathcal{L}_\lambda(\beta) = \|Y - \mathbf{X}\beta\|^2 + \lambda|\beta|_1$$

Analytic solution : when the columns $\mathbf{X}_j$ are orthonormal

$$\left[\widehat{\beta}_\lambda\right]_j = \mathbf{X}_j^T Y \left(1 - \frac{\lambda}{2|\mathbf{X}_j^T Y|}\right)_+$$

Example

We have $n = 60$ noisy observations

$$Y_i = F^*(i/n) + \varepsilon_i, \quad i = 1, \ldots, 60$$

of an unknown undulatory signal $f : [0,1] \to \mathbb{R}$.

We expand the signal on the Fourier basis $\{\varphi_j : j \geq 0\}$

$$Y_i = \sum_j \beta_j^* \varphi_j(i/n) + \varepsilon_i, \quad i = 1, \ldots, 60.$$

To an estimator $\widehat{\beta}$ of β^* we associate an estimator of $F^*(x)$:

$$\widehat{F}(x) = \sum_j \widehat{\beta}_j \varphi_j(x).$$

Shrinkage bias of the Lasso estimator

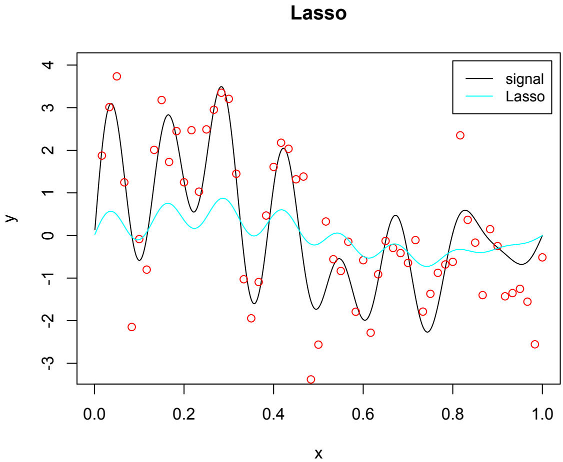


Figure: In black the unknown signal, in red the noisy observations and in cyan the Lasso estimator.

Gauss-lasso estimator

Gauss-Lasso estimator

$$\widehat{f}_\lambda^{\text{Gauss}} = \text{Proj}_{\widehat{S}_\lambda} Y, \quad \text{where} \quad \widehat{S}_\lambda = \text{span}\left\{\mathbf{X}_j : j \in \widehat{m}_\lambda\right\}.$$

In other words,

$$\widehat{f}_\lambda^{\text{Gauss}} = \widehat{f}_{\widehat{m}_\lambda} \quad \text{where} \quad \widehat{m}_\lambda = \text{supp}(\widehat{\beta}_\lambda).$$

Gauss-Lasso estimator

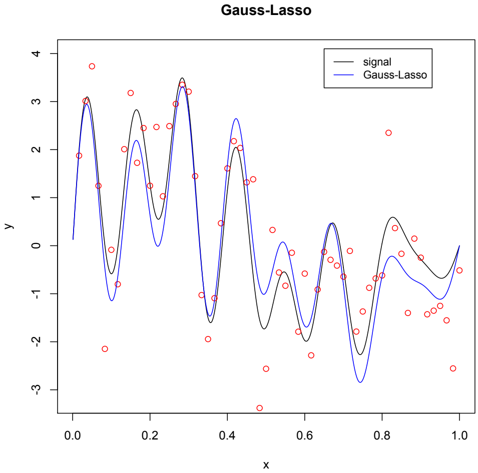


Figure: In black the unknown signal, in red the noisy observations and in blue the Gauss-Lasso estimator.

Adaptive-Lasso estimator

Another trick: compute first the Gauss-Lasso estimator $\widehat{\beta}_{\lambda}^{\mathrm{Gauss}}$ and then estimate β with

Adaptive-Lasso estimator

$$\widehat{\beta}_{\lambda,\mu}^{\mathrm{adapt}} \in \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \left\{ \|Y-\mathbf{X}\beta\|^2 + \mu \sum_{j=1}^{p} \frac{|\beta_j|}{|(\widehat{\beta}_{\lambda}^{\mathrm{Gauss}})_j|} \right\}.$$

for $\beta \approx \widehat{\beta}_{\lambda}^{\mathrm{Gauss}}$ we have $\sum_j |\beta_j|/|(\widehat{\beta}_{\lambda}^{\mathrm{Gauss}})_j| \approx |\beta|_0$

This analogy suggests to take $\mu = (1+\sqrt{2\log(p)})^2$

Adaptive-Lasso estimator

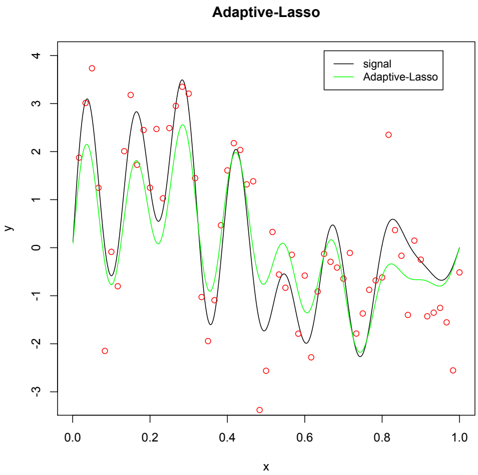


Figure: In black the unknown signal, in red the noisy observations and in green the Adaptive-Lasso estimator.

Estimator Selection

What shall I do with these data ?

Important steps

1. Write down a statistical model suited to analyze the data and answer the scientific question. This requires some
   - deep discussions with specialists (biologists, physicians, etc),
   - low level analyses (PCA, LDA, etc) to detect key features, outliers, etc
   - and ... experience !
2. Choose an estimation procedure
3. Check your results (residues, possible bias, stability, etc)

Setting

Gaussian regression with unknown variance:

- $Y_i = f_i^* + \varepsilon_i$ with $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$
- $f^* = (f_1^*, \ldots, f_n^*)^T$ and σ^2 are unknown
- we want to estimate f^*

Ex 1 : sparse linear regression

- $f^* = \mathbf{X}\beta^*$ with β^* "sparse" in some sense and $\mathbf{X} \in \mathbb{R}^{n\times p}$ with possibly $p > n$

A plethora of estimators

Sparse linear regression

- **Coordinate sparsity:** Lasso, Dantzig, Elastic-Net, Exponential-Weighting, Projection on subspaces $\{V_\lambda : \lambda \in \Lambda\}$ given by PCA, Random Forest, PLS, etc.

- **Structured sparsity:** Group-lasso, Fused-Lasso, Bayesian estimators, etc

Important practical issues

Which estimator shall I use?

Lasso? Group-Lasso? Random-Forest? Exponential-Weighting? Forward–Backward?

With which tuning parameter?

- which penalty level λ for the lasso?
- which beta for expo-weighting?
- etc

Difficulties

- No procedure is universally better than the others
- A sensible choice of the tuning parameters depends on
  - some unknown characteristics of f (sparsity, smoothness, etc)
  - the unknown variance σ^2.

Even if you are a pure Lasso-enthusiast, you miss some key informations in order to apply properly the lasso procedure !

The objective

Formalization

We have a collection of estimation schemes (lasso, group-lasso, etc) and for each scheme we have a grid of different values for the tuning parameters.

At the end, putting all the estimators together we have a collection $\{\hat{f}_\lambda, \lambda \in \Lambda\}$ of estimators.

Ideal objective

Select the "best" estimator among the collection $\{\hat{f}_\lambda, \lambda \in \Lambda\}$.

Cross-Validation

The most popular technique for choosing tuning parameters

Principle

split the data into a training set and a validation set: the estimators are built on the *training* set and the *validation* set is used for estimating their prediction risk.

Most popular cross-validation scheme

- Hold-out : a single split of the data for *training* and *validation*.
- V-fold CV : the data is split into V subsamples. Each subsample is successively removed for *validation*, the remaining data being used for *training*.
- Leave-one-out : corresponds to n-fold CV.
- Leave-q-out : every possible subset of cardinality q of the data is removed for *validation*, the remaining data being used for *training*.

Classical choice of V : between 5 and 10 (remains tractable).

V-fold CV

train	train	train	train	test
train	train	train	test	train
train	train	test	train	train
train	test	train	train	train
test	train	train	train	train

Recursive data splitting for 5-fold Cross-Validation

Pros and Cons

- **Universality:** Cross-Validation can be implemented in most statistical frameworks and for most estimation procedures.
- Usually (but not always!) give good results in practice.
- But **no theoretical garanties** in large dimensional settings.

Complexity selection (LinSelect)

Principle

To adapt the ideas of model selection to estimator selection.

Pros and Cons

- Strong theoretical guaranties,
- Computationally feasible,
- Good performances in the Gaussian setting,

- But relies on the Gaussian assumption

Scaled-Lasso

Automatic tuning of the Lasso

Scaling issue

Change of units

Change of units of the observations: $Y \curvearrowright sY$

After change of units, we observe

$$sY = \mathbf{X}.(s\beta) + s\epsilon$$

A sensible estimator $\widehat{\beta} = \widehat{\beta}(Y, \mathbf{X})$ must fulfill

$$\widehat{\beta}(sY, \mathbf{X}) = s\widehat{\beta}(Y, \mathbf{X}).$$

Scale invariance

The estimator $\widehat{\beta}(Y, \mathbf{X})$ of β^* is scale-invariant if $\widehat{\beta}(sY, \mathbf{X}) = s\widehat{\beta}(Y, \mathbf{X})$ for any $s > 0$.

Example: the estimator

$$\widehat{\beta}(Y, \mathbf{X}) \in \underset{\beta}{\operatorname{argmin}} \|Y - \mathbf{X}\beta\|^2 + \lambda\Omega(\beta),$$

where Ω is homogeneous with degree 1 is not scale-invariant unless λ is proportional to σ.

In particular the Lasso estimator is not scale-invariant when λ is not proportional to σ.

Rescaling

Idea:

- estimate σ with $\widehat{\sigma} = \|Y - \mathbf{X}\beta\|/\sqrt{n}$.
- set $\lambda = \mu\widehat{\sigma}$
- divide the criterion by $\widehat{\sigma}$ to get a convex problem

Scale-invariant criterion

$$\widehat{\beta}(Y, \mathbf{X}) \in \operatorname*{argmin}_{\beta} \sqrt{n}\|Y - \mathbf{X}\beta\| + \mu\Omega(\beta).$$

Example: scaled-Lasso

$$\widehat{\beta} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \left\{ \sqrt{n}\|Y - \mathbf{X}\beta\| + \mu|\beta|_1 \right\}.$$

Pros and Cons

- Universal choice $\mu = 5\sqrt{\log(p)}$
- strong theoretical guaranties (Corollary 5.5)
- computationally feasible

- but poor performances in practice

Numerical experiments (1/2)

Tuning the Lasso

- 165 examples extracted from the literature
- each example e is evaluated on the basis of 400 runs

Comparison to the oracle $\widehat{\beta}_{\lambda^*}$

procedure	quantiles			
	0%	50%	75%	90%
Lasso 10-fold CV	1.03	1.11	1.15	1.19
Lasso LinSelect	0.97	1.03	1.06	1.19
Square-Root Lasso	1.32	2.61	3.37	11.2

For each procedure ℓ, quantiles of $\mathcal{R}\left[\widehat{\beta}_{\hat{\lambda}_\ell};\beta_0\right]/\mathcal{R}\left[\widehat{\beta}_{\lambda^*};\beta_0\right]$, for $e=1,\ldots,165$.

Numerical experiments (2/2)

Computation time

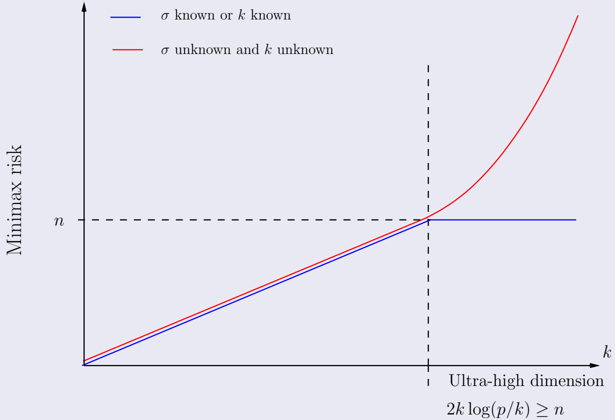
n	p	10-fold CV	LinSelect	Square-Root
100	100	4 s	0.21 s	0.18 s
100	500	4.8 s	0.43 s	0.4 s
500	500	300 s	11 s	6.3 s

Packages:

- `enet` for 10-fold CV and LinSelect
- `lars` for Square-Root Lasso (procedure of Sun & Zhang)

Impact of the unknown variance?

Case of coordinate-sparse linear regression



Minimax prediction risk over k-sparse signal as a function of k