High-dimensional statistics and probability

Christophe Giraud

Université Paris Saclay

M2 Maths Aléa & MathSV

Bias of Lasso estimators

Example

We have $n = 60$ noisy observations

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

of $F^* : [0,1] \to \mathbb{R}$.

We expand F^* on the Fourier basis $\{\varphi_j : j \geq 0\}$

$$Y_i = \sum_j \beta_j^* \underbrace{\varphi_j(i/n)}_{=X_{ij}} + \varepsilon_i, \quad i = 1, \dots, n.$$

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

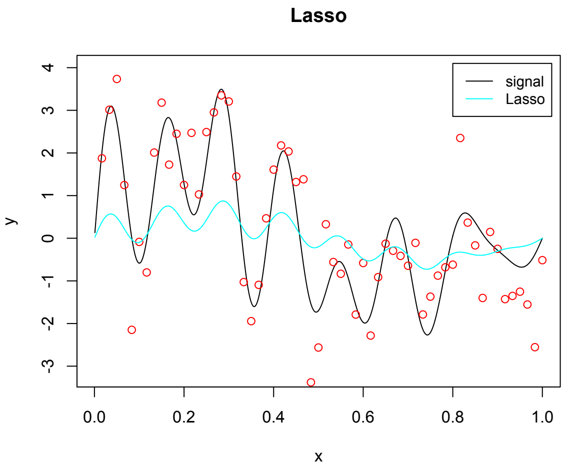

Figure: In black the unknown signal, in red the noisy observations and in cyan the Lasso estimator.

Why?

The lasso estimator is defined by

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

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

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

Gauss-lasso estimator

Gauss-Lasso estimator

We set

$$\widehat{m}_\lambda = \mathrm{supp}(\widehat{\beta}_\lambda)$$

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

In other words,

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

Gauss-Lasso estimator

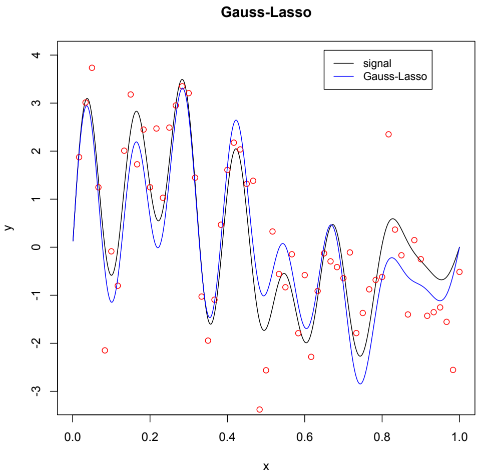

Figure: In black the unknown signal, in red the noisy observations and in blue the Gauss-Lasso estimator.

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

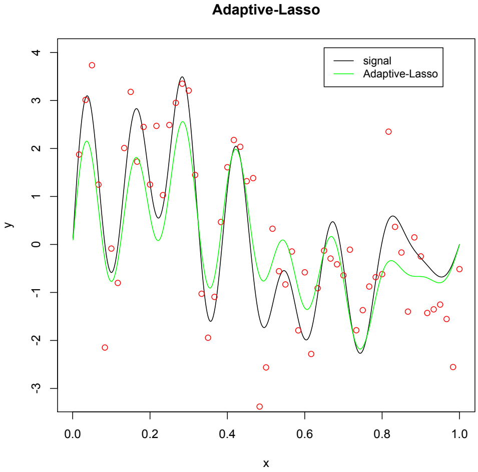

Figure: In black the unknown signal, in red the noisy observations and in green the Adaptive-Lasso estimator.

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

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

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

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

Example: scaled-Lasso

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