Structured regression

Christophe Giraud

Université Paris-Sud

M2 DS

Why this course?

Goal of the lectures

1. to provide some theoretical guidelines for (high-dimensional) data analysis;
2. to highlight some delicate issues;
3. to learn to read a research paper: find the take-home message and understand the limits of the message;
4. to learn to question research papers.

Maths or No-Maths inside?

- we will speak all along about maths results,
- but we will not prove maths results during the lectures.

Goal: to learn to understand and question theoretical papers, not to produce them.

An interesting quote

"This is the first time that we two read an article in statistics on a state-of-the-art subject in detail. It was really not obvious at the beginning. We did not understand the notations and were not familiar with this domain, etc. But after reading it 4 or 5 times, the structure and the logic of the paper became clearer and clearer to us and we became more and more confident. So we would like to say that we are happy to have such experience of mini research in statistics. This will help us to be more confident when possible challenges in this domain occur to us in the future."

(Data Science 2016-17)

Organisation

Structures of the lectures

- Lecture to explain the topic of the session and some related issues
- first (supervised) reading of a research paper
- Discussion of the results of the paper

Between the lectures

Full reading of the research paper

Final "project"

Explain and discuss one of the paper exposed during the lectures (see below).

Please, ask questions!

Topics

1. Strength and weakness of the Lasso
2. False discoveries, multiple testing, online issue
3. Adaptive data analysis
4. Unsupervised dimension reduction: some limits
5. Robust learning

⚠ No deep learning inside!

Requirement

Download the papers before the lectures

http://www.math.u-psud.fr/~giraud/MSV/statsDS.html

Evaluation

Project

Due to mid-february

Mandatory

To attend to all lectures

Rapport à rendre: en binôme

The reports must be sent by email by February 15 in a zip file including:

1. the report in pdf format (10 to 20 pages)
2. if there is some numerics: the notebook (or source code)

Attendu

1) présenter le contexte et les principaux résultats du papier (moitié du rapport maximum).

Il ne s'agit pas de donner un panorama complet du papier, et encore moins un compte rendu littéral. Il s'agit de:

- sélectionner les résultats qui vous semblent les plus importants
- expliquer intuitivement les résultats et (si approprié) les idées sous-jacente à l'algorithme étudié
- commenter leurs implications

2) faire une analyse critique du papier.

- quelles portées des résultats? quelles limitations?
- quel message retenir?

Attendu (suite)

3) procéder à une exploration personnelle, de nature mathématique ou numérique

Côté maths: cela peut être

- expliquer les grandes lignes d'une preuve, les points cruciaux et proposer (de façon argumentée) des possibles extensions pour généraliser ou transposer les résultats.
- une étude théorique comparative des résultats à d'autres résultats récents de la littérature

Côté numérique: il s'agit d'explorer une ou plusieurs problématiques pratiques:

- définir la problématique, le plan d'expérience pour étudier cette problématique (justifier le plan);
- réaliser les expériences et rédiger un notebook explicatif (ou à défaut un code source bien annoté pour comprendre ce qui est fait)
- faire un choix pertinent des résultats à montrer et à commenter
- commenter les résultats et conclure

Critères d'évaluation

Evaluation

1. compréhension de l'article (contexte, motivation, apport, contresens, etc)
2. prise de recul (capacité à expliquer les idées et résultats, leurs implications et leur portée/limite)
3. analyse personnelle:
   - **maths:** compréhension et discernement des points importants, profondeur d'analyse et importance de la contribution personnelle
   - **numérique:** intérêt de la problématique étudiée, pertinence des expériences, qualités des résultats, de leur analyse et discussion

https://www.math.u-psud.fr/~giraud/MSV/statsDSevaluation.html

Projet

Projet

- en binôme
- prendre un des articles du cours et
  1. expliquer le contexte et son message
  2. en cerner/discuter les limites
  3. questionner/discuter numériquement ou théoriquement le papier
- A rendre pour le 15 février minuit.

The reports must be sent by email in a zip file including:

- the report in **pdf format**: 10 to 20 pages;
- the source code for the numerics.

Let's start!

Illustration

Expansion over a Fourier basis :

$$Y_i = \sum_{j=1}^{p} \beta_j^* \phi_j(z_i) + \varepsilon_i = f_{\beta^*}(z_i) + \varepsilon_i, \quad \text{for } i = 1, \ldots, n,$$

with

- $\phi_j(z) = \cos(\pi j z)$
- $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d with $\mathcal{N}(0,1)$ distribution
- β^* sparse and selected at random

Shrinkage bias of the Lasso estimator

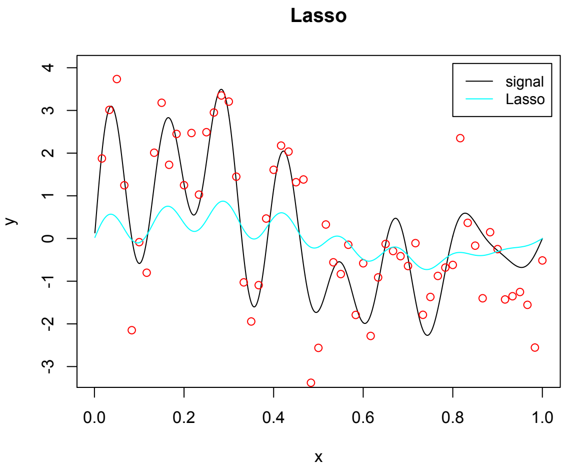


Figure: In black the unknown signal, in red the noisy observations and in cyan the Lasso estimator.

Reminder on the Lasso

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 orthogonal

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

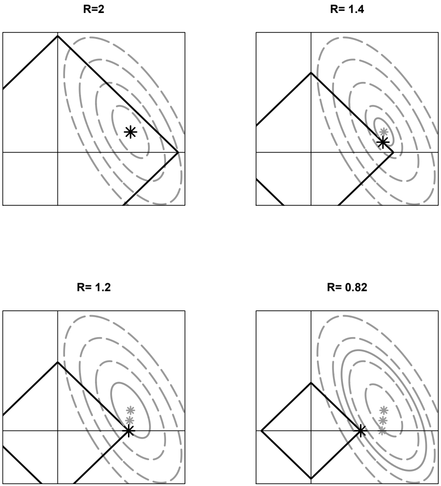
Lasso Path

Dual version of the Lasso

By Lagrangian duality, the Lasso estimator $\widehat{\beta}_\lambda$ is solution of

$$\widehat{\beta}_\lambda \in \operatorname*{argmin}_{\beta \in B_{\ell^1}(\widehat{R}_\lambda)} \|Y - \mathbf{X}\beta\|^2$$

where $\widehat{R}_\lambda = |\widehat{\beta}_\lambda|_1$.



Gauss-lasso estimator

Gauss-Lasso estimator

1. Compute the lasso estimator $\widehat{\beta}_\lambda$;
2. Extract the selected variables $\widehat{S}_\lambda = \mathrm{supp}(\widehat{\beta}_\lambda)$;
3. fit a Least-Square on the selected variables

$$\widehat{\beta}_\lambda^{\mathrm{Gauss}} = \operatorname*{argmin}_{\beta} \|Y - \sum_{j \in \widehat{S}_\lambda} \beta_j X_j\|^2.$$

Gauss-Lasso estimator

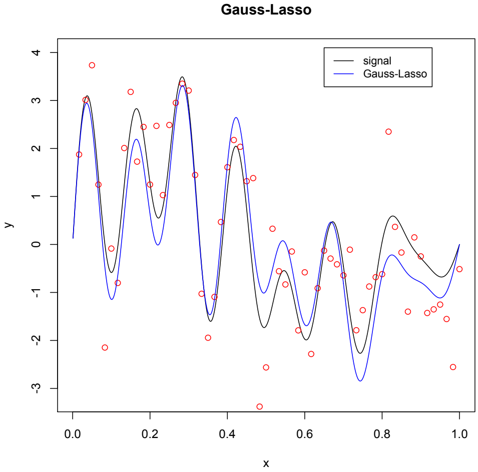


Figure: In black the unknown signal, in red the noisy observations and in blue the Gauss-Lasso estimator.

Adaptive-Lasso estimator

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

Adaptive-Lasso estimator

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

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

This analogy suggests to take $\mu \approx 2\sigma^2 \log(p)$

Adaptive-Lasso estimator

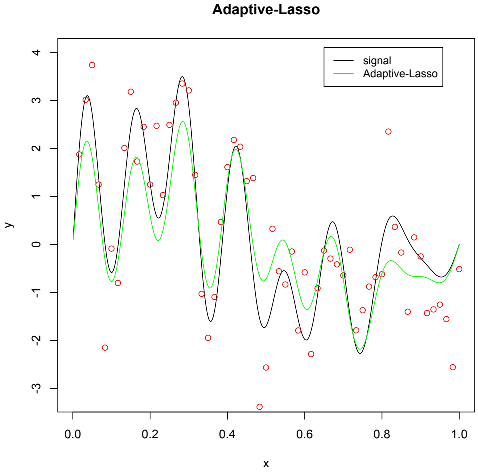


Figure: In black the unknown signal, in red the noisy observations and in green the Adaptive-Lasso estimator.

Scaled-Lasso

Automatic tuning of the Lasso

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)