Statistics for high-dimensional data

Vincent Rivoirard

Université Paris-Dauphine

High-dimensional data

- It has now become part of folklore to claim that the 21st century will be the *century of data*. Our society invests more and more in the collection and processing of data of all kinds (big data phenomenon).
- Now, data have then a strong impact on almost every branch of human activities including science, medicine, business or humanities.
- In traditional statistics, we assumed we had many observations and a few well-chosen variables.
- In modern science, we collect more observations but we also collect radically larger numbers of variables, which consist in thousands up to millions of features voraciously recorded on objects or individuals.
- Such data are said high-dimensional.

Examples of high-dimensional data

- **Consumers preferences data:** Websites gather informations about browsing and shopping behaviors of consumers. For example, recommendation systems collect consumer's preferences on various products, together with some personal data (age, sex, location,...) and predict which products could be of interest for a given consumer.
- **Traffic jams:** Many cities (for instance Boston) have developed programs, to improve traffic, based on big data collection (crowdsourced data) and their analyses.
- **Biotech data:** Recent technologies enable to acquire high-dimensional data on single individuals. For example, DNA microarrrays measure the transcription level of thousands of genes simultaneously.
- **Images and videos:** Large images or videos are continuously collected all around the world. Each image is made of thousands up to millions of pixels.

Characterization and problems of high-dimensional data

- Previous examples show that we are in the era of massive automatic data collection.
- For previous examples, the number of variables or parameters p is much larger than the number of observations n.
- Being able to collect a large amount of information on each individual seems to be good news.
- Unfortunately the mathematical and statistical reality clashes with this optimistic statement: Separating the signal from the noise is a very hard task for high-dimensional data, in full generality impossible.
- Extracting the "good information" is more than challenging, consisting in finding a needle in a haystack.
- This phenomenon is often called the curse of dimensionality, terminology introduced by Richard Bellman, in 1961.

Curse of dimensionality

The volume $V_p(r)$ of a p-dimensional ball of radius r for the euclidian distance satisfies

$$V_p(r) \overset{p\to+\infty}{\sim} \left(\frac{2\pi e r^2}{p}\right)^{p/2} (p\pi)^{-1/2}.$$

So, if $(X^{(i)})_{i=1\ldots,n}$ are i.i.d with uniform distribution on the hypercube $[-0.5, 0.5]^p$, then

$$\begin{aligned}\mathbb{P}(\exists i \in \{1,\ldots,n\} :\ X^{(i)} \in B_p(0,r)) &\leq n \times \mathbb{P}(X^{(1)} \in B_p(0,r)) \\ &\leq n V_p(r).\end{aligned}$$

So if $n = o(V_p(r)^{-1})$, then the last probability goes to 0.

Curse of dimensionality

Example: Classical regression problem. Estimation of the conditional expectation of a random variable. Data consist of n i.i.d. observations $(Y_i, X^{(i)})_{i=1\dots,n}$ with the same distribution as $(Y, X) \in \mathbb{R} \times \mathbb{R}^p$. We wish to estimate the function m where $\mathbb{E}[Y|X] = m(X)$. We consider the Nadaraya-Watson estimate:

$$\hat{m}(x) = \frac{\sum_{i=1}^n K_h(x - X^{(i)}) Y_i}{\sum_{i=1}^n K_h(x - X^{(i)})}, \quad x \in \mathbb{R}^p,$$

$$K_h(x) = \frac{1}{\prod_{j=1}^p h_j} K\left(\frac{x_1}{h_1}, \dots, \frac{x_p}{h_p}\right), \quad h = (h_j)_{j=1,\dots,p}$$

and K is a kernel (with at least one vanishing moment), i.e.

$$K(x) = \prod_{j=1}^p \mathbf{1}_{[-0.5;0.5]}(x_j), \quad K(x) = \frac{1}{(2\pi)^{p/2}} e^{-\frac{\|x\|_2^2}{2}}, \quad K(x) = \frac{1_{\{B_p(0,r)\}}(x)}{\text{Vol}(B_p(0,r))}$$

We have to determine the tuning parameter h which allows to select the variables Y_i associated with the "neighbors" of x among the $X^{(i)}$'s.

Curse of dimensionality

We have to determine the tuning parameter h which allows to select the variables Y_i associated with the "neighbors" of x among the $X^{(i)}$'s. Two problems:

1. We have no neighbor in high dimensions ☹
2. All the points are at a similar distance one from the others. ☹

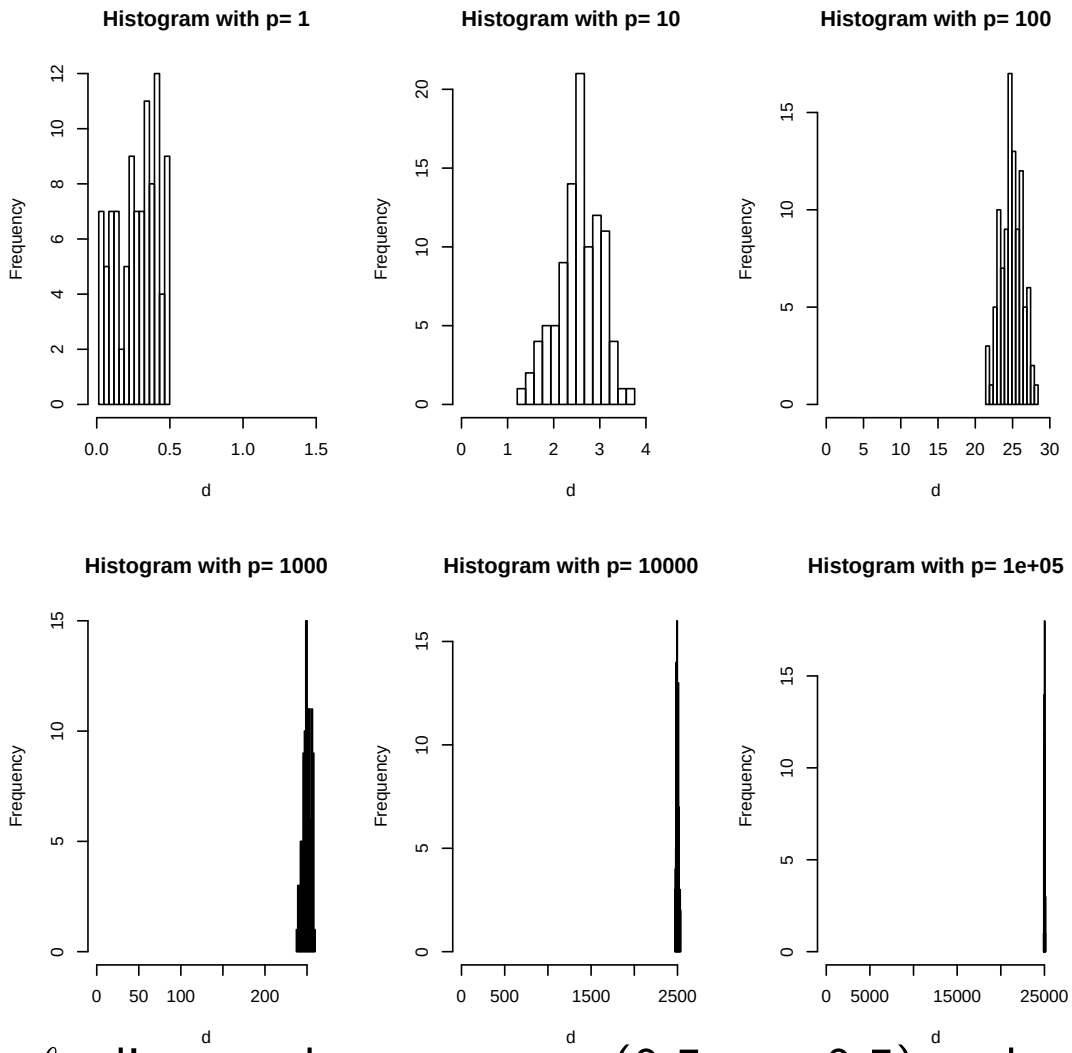
Illustration: Assume that coordinates of X are i.i.d. and $x = (x_1, x_1, \ldots, x_1)$,

$$\mathbb{E}\left[\|x - X\|_1\right] = \mathbb{E}\left[\sum_{j=1}^{p} |x_j - X_j|\right] = p \times \mathbb{E}[|x_1 - X_1|]$$

$$\text{sd}(\|x - X\|_1) = \sqrt{\text{Var}\left(\|x - X\|_1\right)} = \sqrt{p} \times \sqrt{\text{Var}(|x_1 - X_1|)}$$

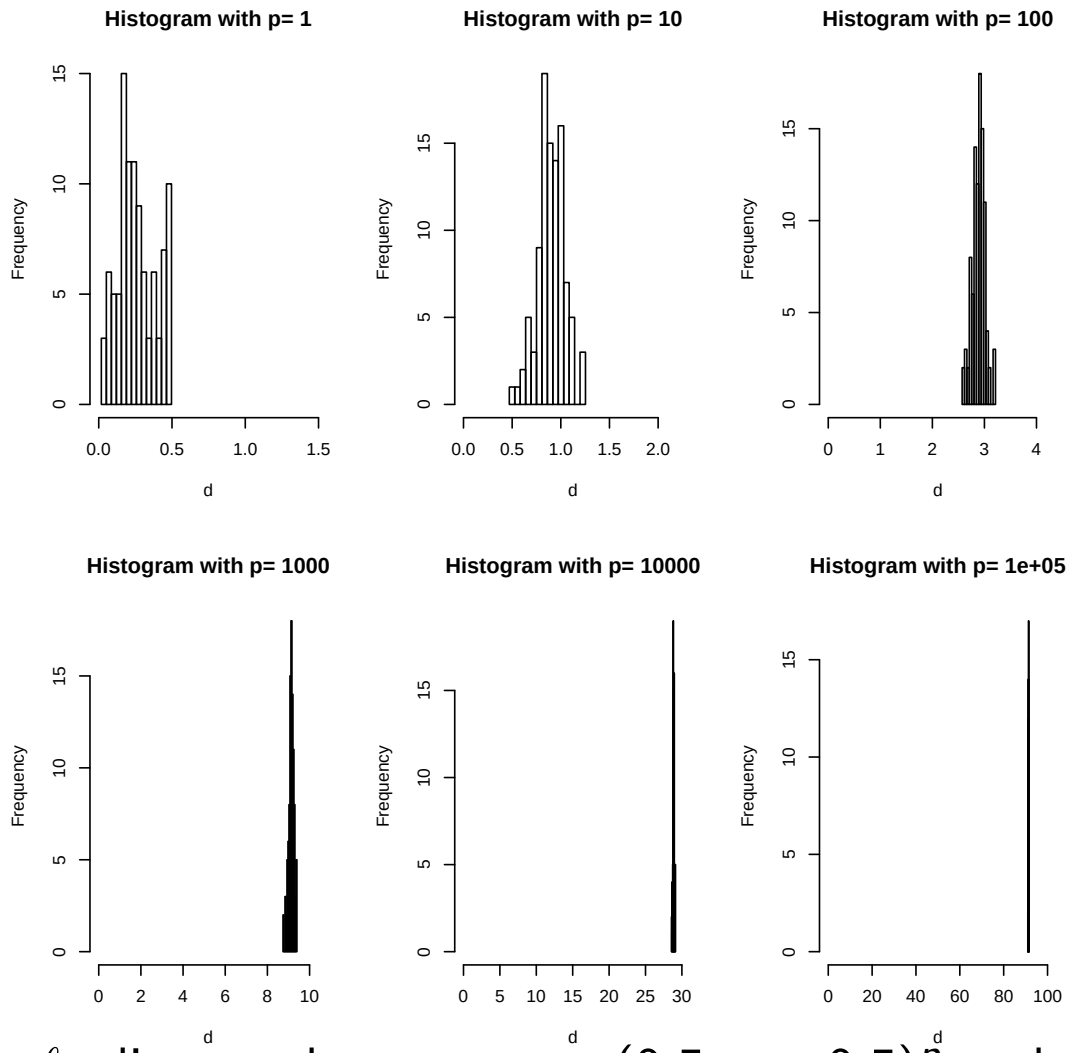
So, any estimator based on a local averaging will fail. ☹

Curse of dimensionality



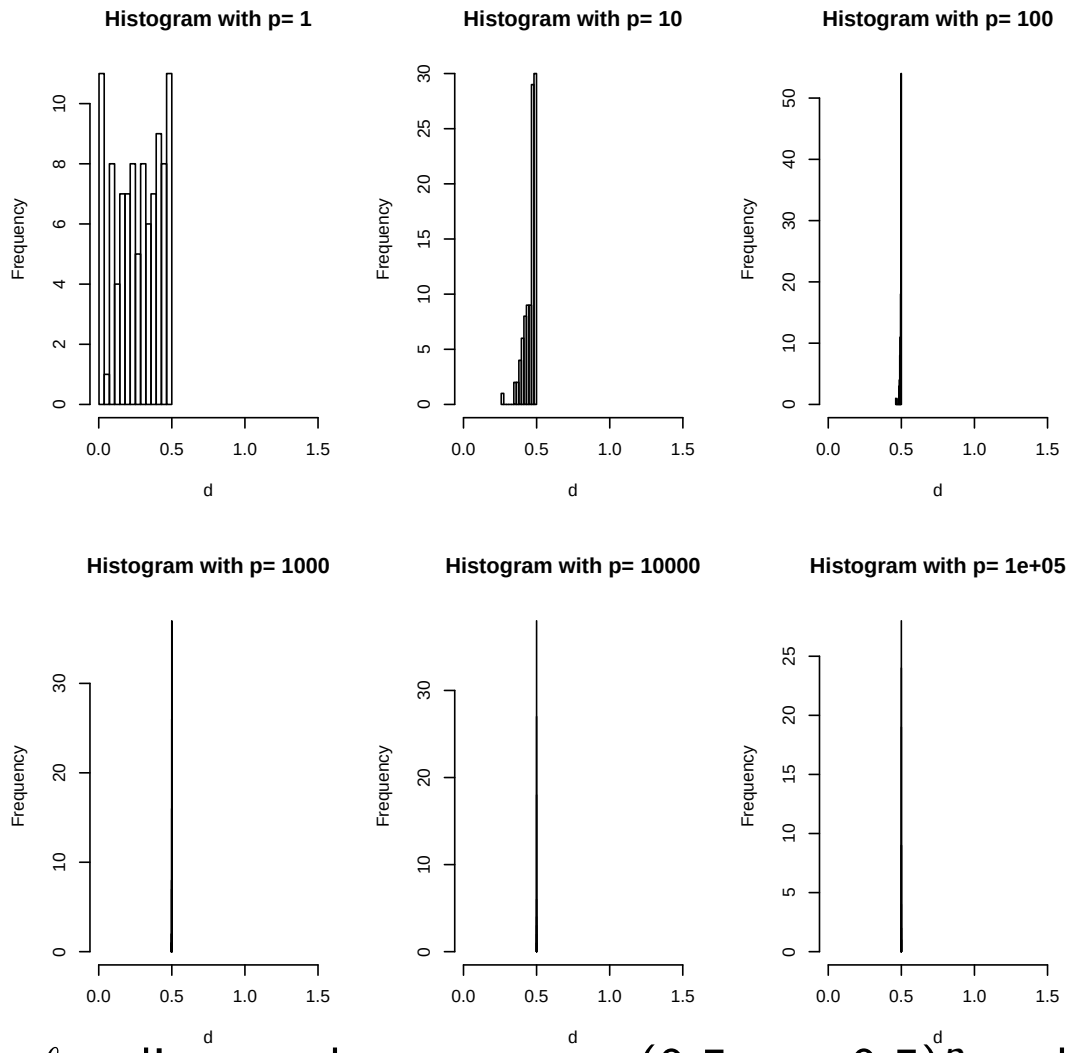
Histograms of the ℓ_1-distance between $x = (0.5, \ldots, 0.5)$ and $n = 100$ random uniform variables on $[0, 1]^p$.

Curse of dimensionality



Histograms of the ℓ_2-distance between $x = (0.5, \ldots, 0.5)^p$ and $n = 100$ random uniform variables on $[0, 1]^p$.

Curse of dimensionality



Histograms of the ℓ_∞-distance between $x = (0.5, \ldots, 0.5)^p$ and $n = 100$ random uniform variables on $[0, 1]^p$.

Curse of dimensionality - Other problems

Other strange phenomena in high dimensions :

- The multivariate Gaussian standard density is very flat:

$$\sup_{x \in \mathbb{R}^p} f(x) = (2\pi)^{-p/2}$$

- The diagonal of the hypercube $[0,1]^p$ is almost orthogonal to its edges

Other problems :

- Accumulation of small fluctuations in many directions can produce a large global fluctuation.
- An accumulation of rare events may not be rare.
- Computational complexity.

Circumventing the curse of dimensionality

At first sight, the high-dimensionality of the data seems to be good news but as explained previously, it is a major issue for extracting information. In light of the few examples described above, the situation may appear hopeless.

- Fortunately, high-dimensional data are not uniformly spread in $\mathbb{R}^p$ (for instance, pixel intensities of an image are not purely random and images have geometrical structures).
- Data are concentrated around low-dimensional structures (many variables have a negligible or even a null impact)....
- ... but this low-dimensional structure is much of the time unknown.

The goal of high-dimensional statistics is to identify these structures and to provide statistical procedures with a low computational complexity.

Take-home message

Whereas classical statistics provide a very rich theory for analyzing data with a small number p of parameters and a large number n of observations, in many fields, current data have different characteristics:

- a huge number p of parameters
- a sample size n, which is of the same size as p or sometimes much smaller than p.

The asymptotic classical analysis with p fixed and n going to $+\infty$ does not make sense anymore. We must change our point of view. We face with the curse of dimensionality.
Fortunately, the useful information usually concentrates around low-dimensional structures (that has to be identified), which allows us to circumvent the course of dimensionality.

References

- Bühlmann, P. and van de Geer, S. *Statistics for high-dimensional data. Methods, theory and applications.* Springer Series in Statistics. Springer, Heidelberg, 2011.

- Donoho, D. *High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality.* American Math. Society "Math Challenges of the 21st Century", 2000.

- Giraud, C. *Introduction to high-dimensional statistics.* Monographs on Statistics and Applied Probability, 139. CRC Press, Boca Raton, FL, 2015.

- Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A. *Wavelets, approximation, and statistical applications.* Lecture Notes in Statistics, 129. Springer-Verlag, New York, 1998.

- Hastie, T., Tibshirani, R. and Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, New York, 2009.

- Hastie, T., Tibshirani, R. and Wainwright, M. *Statistical learning with sparsity.* The lasso and generalizations. Monographs on Statistics and Applied Probability, 143. CRC Press, Boca Raton, FL, 2015

- Tibshirani, R. *Regression shrinkage and selection via the lasso.* J. Roy. Statist. Soc. Ser. B 58, no. 1, 267–288, 1996

Overview of the course

The goal of this course is to present modern statistical tools and some of their theoretical properties for estimation in the high-dimensional setting, including

1. Wavelets and thresholding rules.
2. Penalized estimators: model selection procedures, Ridge and Lasso estimates.
3. Generalizations and variations of the Lasso estimate: Group-Lasso, Fused-Lasso, elastic-net and Dantzig selectors. Links with Bayesian rules.
4. Statistical properties of Lasso estimators: study in the classical regression model. Extensions for the generalized linear model.

I shall concentrate on simple settings in order to avoid unessential technical details.

What is (unfortunately) not mentioned and notations

Due to lack of time or skill, I won't speak about some important themes (fortunately, some of them will be dealt with by Franck or Tristan):

1. Optimizations aspects
2. Matrix completion
3. Testing approaches
4. Graphical models
5. Multivariate methods (sparse PCA, etc.)
6. Classification methods

Notations:

- n: size of observations.
- p: dimension of the involved unknown quantity.
- $\|\cdot\|_q$: ℓ_q-norm in $\mathbb{R}^p$.
- For short if there is no ambiguity, $\|\cdot\| = \|\cdot\|_2$.
- For any vector β, $\|\beta\|_0 = \text{card}\{j : \ \beta_j \neq 0\}$.

1 Introduction

Chapter 1: Model selection

Contents of the chapter:

Linear regression setting

Consider the linear regression model

$$Y = X\beta^* + \epsilon,$$

with

- $Y = (Y_i)_{i=1,\dots,n}$ a vector of observations (response variable)
- $X = (X_{ij})_{i=1,\dots,n,\, j=1,\dots,p}$ a known $n \times p$-matrix.
- $\beta^* = (\beta^*_j)_{j=1,\dots,p}$ an unknown vector
- $\epsilon = (\epsilon_i)_{i=1,\dots,n}$ the vector of errors. It is assumed that

$$\mathbb{E}[\epsilon] = 0, \quad \mathrm{Var}(\epsilon) = \sigma^2 I_n$$

and σ^2 is known.

Columns of X, denoted X_j, are explanatory variables or predictors.

Linear regression setting

The regression model can be rewritten as

$$Y = \sum_{j=1}^{p} \beta_j^* X_j + \epsilon.$$

Several problems can be investigated:

- The estimation problem: Estimate β^*
- The prediction problem: Estimate $X\beta^*$
- The selection problem: Determine non-zero coordinates of β^*

Why linear regression?

- It models various concrete situations
- It is simple to use from the mathematical point of view
- It allows to introduce and to present new methodologies

Classical estimation

We naturally estimate β^* by considering the ordinary least squares estimate $\hat{\beta}^{ols}$ defined by

$$\hat{\beta}^{ols} := \arg\min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|.$$

Proposition

If $rank(X) = p$, *then*

$$\hat{\beta}^{ols} = (X^T X)^{-1} X^T Y$$

and

$$\mathbb{E}[\hat{\beta}^{ols}] = \beta^*, \quad \mathrm{Var}(\hat{\beta}^{ols}) = \sigma^2 (X^T X)^{-1}.$$

Furthermore

$$\mathbb{E}\left[\|\hat{\beta}^{ols} - \beta^*\|^2\right] = \sigma^2 \times Tr((X^T X)^{-1}).$$

Classical estimation

Lemma

We have for any matrix A with n columns:

$$\mathbb{E}\left[\|A\epsilon\|^2\right] = \sigma^2 Tr(AA^T)$$

Some remarks:

- $\text{rank}(X) = p$ implies $p \leq n$
- If the predictors are orthonormal

$$\mathbb{E}\left[\|\hat{\beta}^{ols} - \beta^*\|^2\right] = p\sigma^2,$$

which may be large in high dimensions.

Up to now, structural assumptions are very mild. In the sequel, we shall first consider sparsity assumptions.

Sparsity

- Loosely speaking, a sparse statistical model is a model in which only a relatively small number of parameters play an important role.
- In the regression model,

$$Y = \sum_{j=1}^{p} \beta_j^* X_j + \epsilon$$

we assume that m^* the support of β^* is small, with

$$m^* = \left\{ j \in \{1, \dots, p\} : \ \beta_j^* \neq 0 \right\}.$$

Note that m^* is unknown.

- In general, $\hat{\beta}^{ols}$ is not sparse.
- Model selection is a natural approach to select a good estimator in this setting. We describe and study this methodology in the oracle approach.

Oracle approach

- We now consider the prediction risk and set $f^* = X\beta^* \in \mathbb{R}^n$ the unknown vector of interest. So, we have:

$$Y = f^* + \epsilon. \tag{2.1}$$

- If m^* were known, a natural estimate of f^* would be

$$\hat{f}_{m^*} = \Pi_{S^*} Y,$$

with $\Pi_{S^*} : \mathbb{R}^n \mapsto \mathbb{R}^n$ the projection matrix on S^* and

$$S^* = \text{span}(X_j : \ j \in m^*).$$

- Note that if $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ then $\hat{f}_{m^*}$ is the maximum likelihood estimate in the model (2.1) under the constraint that the estimate of f^* belongs to S^*.
- Of course m^* is unknown and $\hat{f}_{m^*}$ cannot be used.

Oracle approach

- For any model $m \subset \{1, \ldots, p\}$, we set $\hat{f}_m = \Pi_{S_m} Y$, with $\Pi_{S_m} : \mathbb{R}^n \mapsto \mathbb{R}^n$ the projection matrix on S_m and

$$S_m = \text{span}(X_j : \ j \in m).$$

  With a slight abuse, we also call S_m model.
- Given $\mathcal{M}$, a collection of models, we wish to select $\hat{m} \in \mathcal{M}$ such that the risk of $\hat{f}_{\hat{m}}$ is as small as possible.
- We introduce the oracle model m_0 as

$$m_0 := \arg\min_{m \in \mathcal{M}} \mathbb{E}\left[\|\hat{f}_m - f^*\|^2\right].$$

  $\hat{f}_{m_0}$ is called the (pseudo) oracle estimate.
- More precisely, we wish to select $\hat{m} \in \mathcal{M}$ such that

$$\mathbb{E}\left[\|\hat{f}_{\hat{m}} - f^*\|^2\right] \approx \mathbb{E}\left[\|\hat{f}_{m_0} - f^*\|^2\right].$$

Oracle approach

Oracle model:

$$m_0 := \arg\min_{m \in \mathcal{M}} \mathbb{E}\left[\|\hat{f}_m - f^*\|^2\right].$$

Some remarks:

- m^* may be different from m_0. We may have $f^* \notin S_{m_0}$ and even $f^* \notin \cup_{m \in \mathcal{M}} S_m$.
- The oracle model m_0 is not random but depends on β^*. So, it cannot be used in practice.

Model selection procedure

Our approach is based on the minimization of $R(\hat{f}_m)$ on $\mathcal{M}$, with

$$R(\hat{f}_m) := \mathbb{E}\left[\|\hat{f}_m - f^*\|^2\right].$$

The following lemma based on the simple bias-variance decomposition gives an explicit expression of $R(\hat{f}_m)$. We denote

$$d_m := \dim(S_m).$$

Lemma

We have:

$$R(\hat{f}_m) = \|(I_n - \Pi_{S_m})f^*\|^2 + \sigma^2 d_m.$$

- The first term is a bias term which decreases when m increases, whereas the second term (a variance term) increases when m increases
- The oracle model m_0 is the model which achieves the best trade-off between these two terms.

Mallows' C_p

Mallows' Recipe: Since we wish to minimize $m \mapsto R(\hat{f}_m)$, it's natural to choose $\hat{m}$ as the minimizer of an estimate of $R(\hat{f}_m)$. We denote the latter $\hat{R}_m$ that will be based on $\|\hat{f}_m - Y\|^2$ (replacing f^* with Y and removing the expectation). The following lemma gives the last ingredient of the recipe.

Lemma

We have:

$$\mathbb{E}\left[\|\hat{f}_m - Y\|^2\right] = R(\hat{f}_m) - \sigma^2(2d_m - n)$$

Using the lemma, an unbiased estimate of $R(\hat{f}_m)$ is given by

$$\|\hat{f}_m - Y\|^2 + \sigma^2(2d_m - n).$$

It leads to the model selection procedure based on minimization of Mallows' criterion defined by:

$$C_p(m) := \|\hat{f}_m - Y\|^2 + 2\sigma^2 d_m$$

Mallows' C_p

Definition

Mallows' estimate of f^* is $\hat{f} := \hat{f}_{\hat{m}}$ with

$$\hat{m} = \arg\min_{m \in \mathcal{M}} C_p(m), \quad C_p(m) := \|\hat{f}_m - Y\|^2 + 2\sigma^2 d_m$$

- Assumptions are very mild. In particular the Mallows' criterion is distribution-free. It's a very popular criterion.
- Only based on unbiased estimation, this approach does not take into account fluctuations of $C_p(m)$ around its expectation. The larger $\mathcal{M}$, the larger the probability to have $\min_{m \in \mathcal{M}} C_p(m)$ far from $\min_{m \in \mathcal{M}} R(\hat{f}_m) + \sigma^2 n$. In particular, we may have for some $m \in \mathcal{M}$, $C_p(m) \ll R(\hat{f}_m) + \sigma^2 n$ and

$$C_p(m) < C_p(m_0), \quad R(\hat{f}_m) > R(\hat{f}_{m_0}).$$

- The last situation occurs when we have a large number of models for each dimension. $\hat{m}$ is much larger than m_0 leading to overfitting. It's the main drawback of Mallows' C_p.

Other popular criteria

When the distribution of observations is known, we can consider AIC and BIC criteria which are based on the likelihood. For any model $m \in \mathcal{M}$, we set $L(m)$ as the maximum of the log-likelihood on S_m. We still consider

$$\hat{m} := \arg\min_{m \in \mathcal{M}} C(m),$$

with

- for the Akaike Information Criterion (AIC)

$$C(m) = -L(m) + d_m$$

- for the Bayesian Information Criterion (BIC)

$$C(m) = -L(m) + \frac{\log n}{2} \times d_m$$

In the Gaussian setting, AIC and Mallows' C_p are equivalent. The use of BIC tends to prevent overfitting (larger penalty).

Penalization for Gaussian regression

- We assume
$$\epsilon \sim \mathcal{N}(0, \sigma^2 I_n).$$
- Mallows' approach shows that for ℓ_2 estimation, a criterion of the form
$$C(m) = \|\hat{f}_m - Y\|^2 + \sigma^2 \mathrm{pen}(m),$$
is suitable with pen, called the penalty, satisfying $\mathrm{pen}(m) \geq 2d_m$.
- We now investigate good choices of penalties. It has to depend on $\mathcal{M}$.
- Recall our benchmark: The oracle risk $R(\hat{f}_{m_0})$ with
$$m_0 := \arg\min_{m \in \mathcal{M}} R(\hat{f}_m), \quad R(\hat{f}_m) := \mathbb{E}\left[\|\hat{f}_m - f^*\|^2\right].$$
We wish $R(\hat{f}) \approx R(\hat{f}_{m_0})$.
- We have
$$R(\hat{f}_m) = \|(I_n - \Pi_{S_m}) f^*\|^2 + \sigma^2 d_m.$$

Penalty

Since for any $m \in \mathcal{M}$, $C(\hat{m}) \leq C(m)$, we have:

$$\|f^* - \hat{f}_{\hat{m}}\|^2 + 2\langle \epsilon, f^* - \hat{f}_{\hat{m}} \rangle + \sigma^2 \mathrm{pen}(\hat{m}) \leq \|f^* - \hat{f}_m\|^2 + 2\langle \epsilon, f^* - \hat{f}_m \rangle + \sigma^2 \mathrm{pen}(m)$$

Taking expectation, since $\mathrm{pen}(m)$ is deterministic,

$$R(\hat{f}) \leq \underbrace{R(\hat{f}_m)}_{I} + 2\underbrace{\mathbb{E}\Big[\langle \epsilon, f^* - \hat{f}_m \rangle\Big]}_{II} + \underbrace{\sigma^2 \mathrm{pen}(m)}_{III} + \underbrace{\mathbb{E}\Big[2\langle \epsilon, \hat{f} - f^* \rangle - \sigma^2 \mathrm{pen}(\hat{m})\Big]}_{IV}$$

Each term can be analyzed: I is ok.

$$II := \mathbb{E}\Big[\langle \epsilon, f^* - \hat{f}_m \rangle\Big] = \mathbb{E}\Big[\langle \epsilon, f^* - \Pi_{S_m} Y \rangle\Big] = -\mathbb{E}\Big[\|\Pi_{S_m} \epsilon\|^2\Big] = -\sigma^2 d_m \leq 0.$$

The function $\mathrm{pen}(\cdot)$ has to be large enough so that IV is negligible but small enough to have

$$III := \sigma^2 \mathrm{pen}(m) \lesssim R(\hat{f}_m).$$

Then,

$$R(\hat{f})) \lesssim \inf_{m \in \mathcal{M}} R(\hat{f}_m) + \text{negl. term.}$$

Analysis of the forth term

For any $0 < \delta < 1$, with $\bar{S}_{\hat{m}} = \text{span}(S_{\hat{m}}, f^*)$,

$$\begin{aligned} 2\langle \epsilon, \hat{f} - f^* \rangle &= 2\langle \Pi_{\bar{S}_{\hat{m}}} \epsilon, \hat{f} - f^* \rangle \\ &\leq \delta^{-1} \|\Pi_{\bar{S}_{\hat{m}}} \epsilon\|^2 + \delta \|\hat{f} - f^*\|^2. \end{aligned}$$

And, with $\chi^2(m) := \|\Pi_{\bar{S}_m}(\sigma^{-1}\epsilon)\|^2$,

$$\begin{aligned} IV &:= \mathbb{E}\Big[2\langle \epsilon, \hat{f} - f^* \rangle - \sigma^2 \text{pen}(\hat{m})\Big] \\ &\leq \delta^{-1}\sigma^2 \mathbb{E}\Big[\chi^2(\hat{m}) - \delta \text{pen}(\hat{m})\Big] + \delta R(\hat{f}) \\ &\leq \delta^{-1}\sigma^2 \mathbb{E}\Big[\max_{m \in \mathcal{M}} \{\chi^2(m) - \delta \text{pen}(m)\}\Big] + \delta R(\hat{f}) \\ &\leq \delta^{-1}\sigma^2 \sum_{m \in \mathcal{M}} \Big(\mathbb{E}\Big[\chi^2(m)\Big] - \delta \text{pen}(m)\Big) + \delta R(\hat{f}) \end{aligned}$$

Penalty

Definition

To the collection of models $\mathcal{M}$, we associate $(\pi_m)_{m\in\mathcal{M}}$ such that $0 < \pi_m \leq 1$ and

$$\sum_{m\in\mathcal{M}} \pi_m = 1.$$

Then, for any constant $K > 1$, we set

$$\text{pen}(m) := K\left(\sqrt{d_m} + \sqrt{-2\log(\pi_m)}\right)^2. \tag{2.2}$$

If $K > 1$, taking e.g. $\delta = K^{-1}$, concentration inequalities lead to

$$IV \leq C(K)\sigma^2 + K^{-1}R(\hat{f})$$

$$\begin{aligned} III &:= \sigma^2\text{pen}(m) \leq 2K\sigma^2 d_m + 4K\sigma^2\log(\pi_m^{-1}) \\ &\leq 2KR(\hat{f}_m) + 4K\sigma^2\log(\pi_m^{-1}) \end{aligned}$$

Theoretical result

Theorem (Birgé and Massart)

We consider the linear regression model

$$Y = f^* + \epsilon$$

and assume that $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, *with* σ^2 *known. Given* $K > 1$, *we define the penalty function as in* (2.2) *and estimate* f^* *with* $\hat{f} = \hat{f}_{\hat{m}}$ *such that*

$$\hat{m} := \arg\min_{m \in \mathcal{M}} \left\{ \|\hat{f}_m - Y\|^2 + \sigma^2 \mathrm{pen}(m) \right\}.$$

Then, there exists $C_K > 0$ *only depending on* K *such that*

$$\mathbb{E}\left[\|\hat{f} - f^*\|^2\right] \leq C_K \min_{m \in \mathcal{M}} \left\{ \mathbb{E}\left[\|\hat{f}_m - f^*\|^2\right] + \sigma^2 \log(\pi_m^{-1}) + \sigma^2 \right\}.$$

- If $\log(\pi_m^{-1}) \lesssim \alpha d_m$ then $\hat{f}$ achieves the same risk as the oracle.
- Mallows' C_p will be suitable if $\exists K > 1$ s.t.

$$K\left(\sqrt{d_m} + \sqrt{-2\log(\pi_m)}\right)^2 \sim 2d_m.$$

First illustration with the full collection of models

We first consider the case where $\mathcal{M} = \mathcal{P}(\{1, \dots, p\})$. We can take

$$\pi_m = \frac{(e-1)}{(1-e^{-p})} \frac{d_m!(p-d_m)!}{p!} e^{-d_m}.$$

Then

$$\log(\pi_m^{-1}) \lesssim d_m \log\left(\frac{p}{d_m}\right) \leq d_m \log(p)$$

and

$$\mathbb{E}\left[\|\hat{f} - f^*\|^2\right] \lesssim \log(p) \min_{m \in \mathcal{M}} \mathbb{E}\left[\|\hat{f}_m - f^*\|^2\right]$$

- The $\log(p)$-term is unavoidable
- We can prove that by taking $K < 1$, we select a very big model, leading to overfitting. It's the reason why Mallows' C_p is not suitable for this case.

Second illustration with a poor collection of models

We now consider the case where $\mathcal{M} = \{\{1, \dots, J\},\ 1 \leq J \leq p\}$. We can take for any constant $\alpha > 0$,

$$\pi_m = \frac{(e^\alpha - 1)}{(1 - e^{-\alpha p})} e^{-\alpha d_m}.$$

Then

$$\log(\pi_m^{-1}) \leq \alpha d_m + const.$$

Since

$$\mathbb{E}\left[\|\hat{f}_m - f^*\|^2\right] = \|(I_n - \Pi_{S_m}) f^*\|^2 + \sigma^2 d_m.$$

we have

$$\mathbb{E}\left[\|\hat{f} - f^*\|^2\right] \lesssim \min_{m \in \mathcal{M}} \mathbb{E}\left[\|\hat{f}_m - f^*\|^2\right]$$

- Under convenient choices of α and $K > 1$, we have $\mathrm{pen}(m) \sim 2 d_m$. Therefore, Mallows' C_p is suitable for this case.
- The choice $K < 1$ leads to overfitting

Pros and cons of model selection

- Under a convenient choice of penalty (based on concentration inequalities), the model selection methodology is able to select the "best" predictors to explain a response variable by only using data. ☺
- The model selection methodology (due to Birgé and Massart) has been presented in the Gaussian linear regression setting. But it can be extended to other settings: for density estimation, Markov models, counting processes, segmentation, classification, etc. ☺
- It is based on minimization of a penalized ℓ_2-criterion over a collection of models. Note that if $\mathcal{M} = \mathcal{P}(\{1, \ldots, p\})$, $\text{card}(\mathcal{M}) = 2^p$. When p is large, this approach is intractable due to a prohibitive computational complexity ($2^{20} > 10^6$). ☹

The orthogonal case

Assume that the matrix X is orthogonal: $X^T X = I_p$. We have $d_m := \dim(S_m) = \text{card}(m)$. Consider a penalty proportional to d_m:

$$\text{pen}(m) = 2K d_m \log(p).$$

Then, since

$$\hat{f}_m = \Pi_{S_m} Y = \sum_{j \in m} \hat{\beta}_j X_j, \quad \hat{\beta}_j := X_j^T Y$$

we obtain:

$$\begin{aligned}
\hat{m} &:= \arg\min_{m \in \mathcal{M}} \left\{ \|\hat{f}_m - Y\|^2 + \sigma^2 \text{pen}(m) \right\} \\
&= \arg\min_{m \in \mathcal{M}} \left\{ -\sum_{j \in m} \hat{\beta}_j^2 + 2K\sigma^2 \text{card}(m) \log(p) \right\} \\
&= \arg\min_{m \in \mathcal{M}} \left\{ -\sum_{j \in m} \left(\hat{\beta}_j^2 - 2K\sigma^2 \log(p) \right) \right\}
\end{aligned}$$

The orthogonal case and $\mathcal{M} = \mathcal{P}(\{1, \ldots, p\})$

- In this case, we have:

$$\hat{m} = \left\{ j \in \{1, \ldots, p\} : \ |\hat{\beta}_j| > \sigma \sqrt{2K \log(p)} \right\}$$

and

$$\hat{f} = \hat{f}_{HT,K} := \sum_{j=1}^{p} \hat{\beta}_j \mathbf{1}_{\left\{ |\hat{\beta}_j| > \sigma \sqrt{2K \log(p)} \right\}} X_j$$

Model selection corresponds to hard thresholding and implementation is easy.

- Assume that $f^* = 0$. Consider Mallows' C_p, BIC and hard thresholding alternatively. The first two are overfitting procedures: if $p \to +\infty$,
  1. with $\mathrm{pen}(m) = 2d_m$, $\mathbb{E}[\mathrm{card}(\hat{m}_{\mathrm{Mallows}})] \sim 0.16p$.
  2. if $n = p$ and $\mathrm{pen}(m) = \log(n) d_m$, $\mathbb{E}[\mathrm{card}(\hat{m}_{\mathrm{BIC}})] \sim \sqrt{\frac{2p}{\pi \log(p)}}$
  3. if $K > 1$, $\mathbb{P}(\hat{f}_{HT,K} \neq 0) = o(1)$

Take-home message

- This chapter presents in the Gaussian linear setting the model selection methodology, which consists in minimizing an ℓ_0-penalized criterion.
- Such procedures are very popular in the moderately large dimensions setting and can be extended to many statistical models.
- Using concentration inequalities, penalties can be designed to obtain adaptive and optimal procedures in the oracle setting and to overperform classical procedures, such as AIC, BIC and Mallows' C_p.
- When p is large and the model collection is wealthy, this approach may be intractable due to a prohibitive computational complexity. Alternatives have to be developed in very high dimensions.

References

- Birgé, Lucien and Massart, Pascal Gaussian model selection. *J. Eur. Math. Soc. (JEMS)* 3, no. 3, 203268, 2001.

- Giraud, Christophe *Introduction to high-dimensional statistics.* Monographs on Statistics and Applied Probability, 139. CRC Press, Boca Raton, FL, 2015.

- Mallows, Colin Some Comments on C_p. *Technometrics*, 15, 661-675, 1973.

- Massart, Pascal *Concentration inequalities and model selection*, volume 6, Springer, 2007.

- Verzelen, Nicolas Minimax risks for sparse regressions: ultra-high dimensional phenomenons. *Electron. J. Stat.*, 6, 38–90, 2012.

Chapter 2: From Ridge estimate to Lasso estimate

Ridge estimates

We still consider the linear regression model

$$Y = X\beta^* + \epsilon,$$

with $\mathbb{E}[\epsilon] = 0$, $\mathrm{Var}(\epsilon) = \sigma^2 I_n$ and σ^2 is known. If $\mathrm{rank}(X) = p$, then

$$\hat{\beta}^{ols} = (X^T X)^{-1} X^T Y$$

which satisfies

$$\mathbb{E}[\hat{\beta}^{ols}] = \beta^*, \quad \mathrm{Var}(\hat{\beta}^{ols}) = \sigma^2 (X^T X)^{-1}.$$

$$\mathbb{E}\left[\|\hat{\beta}^{ols} - \beta^*\|^2\right] = \sigma^2 \times Tr((X^T X)^{-1}).$$

In high dimensions, the matrix $X^T X$ can be ill-conditioned (i.e. may have small eigenvalues) leading to coordinates of $\hat{\beta}^{ols}$ with large variance. To overcome this problem while preserving linearity, we modify the OLS estimate and set

$$\hat{\beta}^{ridge}_{\lambda} = (X^T X + \lambda I_p)^{-1} X^T Y, \quad \lambda > 0$$

Ridge estimates

Since

$$\hat{\beta}_\lambda^{ridge} = (X^T X + \lambda I_p)^{-1} X^T Y, \quad \lambda > 0,$$

the tuning parameter λ balances the bias and variance terms.

$$\|\mathbb{E}[\hat{\beta}_\lambda^{ridge}] - \beta^*\|^2] = \lambda^2 \beta^{*T} (X^T X + \lambda I_p)^{-2} \beta^*$$

$$\mathbb{E}\left[\|\hat{\beta}_\lambda^{ridge} - \mathbb{E}[\hat{\beta}_\lambda^{ridge}]\|^2\right] = \sigma^2 \sum_{j=1}^{p} \frac{\mu_j}{(\mu_j + \lambda)^2},$$

with $(\mu_j)_{j=1,\dots,p} :=$ eigenvalues$(X^T X)$.

Pros and cons:

- We can consider very high dimensions: $p \gg n$ ☺
- Linearity: Easy to compute for many (?) problems ☺
- The choice of the regularization parameter λ is sensitive
- Automatic selection is not possible ☹

Bridge estimates

Definition

For $\lambda \geq 0$ and $\gamma \geq 0$, we set:

$$C_{\lambda,\gamma}(\beta) := \|Y - X\beta\|^2 + \lambda\|\beta\|_\gamma^\gamma$$

with

$$\|\beta\|_\gamma^\gamma = \begin{cases} \sum_{j=1}^p |\beta_j|^\gamma, & \text{if } \gamma > 0 \\ \sum_{j=1}^p \mathbf{1}_{\{\beta_j \neq 0\}}, & \text{if } \gamma = 0 \end{cases}$$

and

$$\hat{\beta}_{\lambda,\gamma} := \arg\min_{\beta \in \mathbb{R}^p} C_{\lambda,\gamma}(\beta). \tag{3.1}$$

Three interesting cases ($\lambda > 0$):

1. $\gamma = 0$: model selection
2. $\gamma = 2$: Ridge estimation
3. $\gamma = 1$: Lasso Estimation

Bridge estimates

- Assume that $\gamma = 0$. Then the bridge estimate exists if X is one-to-one.
- Assume that $\gamma > 0$. Then the bridge estimate exists.
- If $0 \leq \gamma < 1$, then $C_{\lambda,\gamma}$ is not convex and it may be very hard to minimize it in high dimensions.
- Assume that $\gamma = 1$. The penalized criterion $C_{\lambda,1}$ is then convex and $C_{\lambda,1}$ has one minimizer if X is one-to-one.
- Assume that $\gamma > 1$. The penalized criterion $C_{\lambda,\gamma}$ is then strictly convex and $C_{\lambda,\gamma}$ has only one minimizer. Almost surely, all coordinates of the bridge estimate are non-zero.

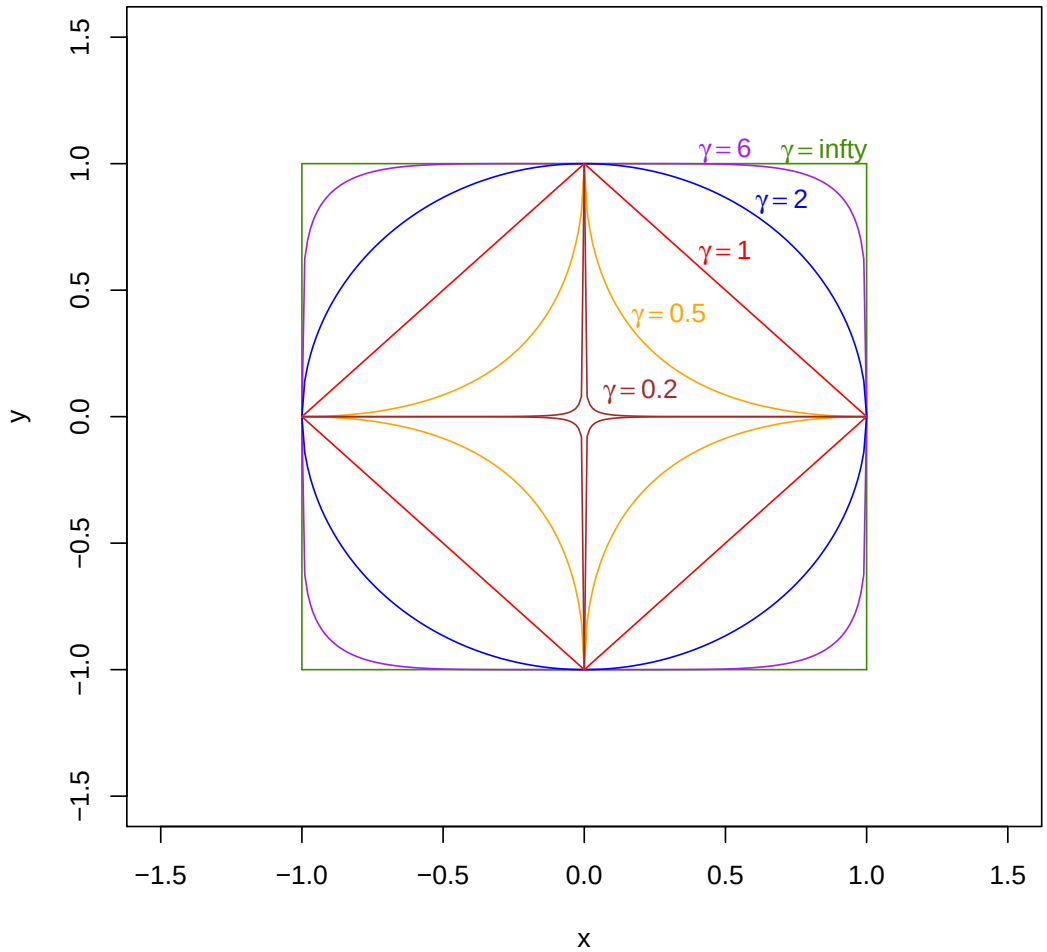
For $\gamma \geq 1$, one-to-one correspondence between the Lagragian problem

$$\hat{\beta}_{\lambda,\gamma} := \arg\min_{\beta \in \mathbb{R}^p} C_{\lambda,\gamma}(\beta), \quad C_{\lambda,\gamma}(\beta) := \|Y - X\beta\|^2 + \lambda \|\beta\|_\gamma^\gamma$$

and the following constrained problem

$$\arg \min_{\{\beta \in \mathbb{R}^p:\ \|\beta\|_\gamma^\gamma \leq t\}} \|Y - X\beta\|.$$

Bridge estimates



Constraints regions $\|\beta\|_\gamma^\gamma \leq 1$ for different values of γ. The region is convex if and only if $\gamma \geq 1$.

Graphical illustration for $p = 2$

- We take $X^T X = \begin{pmatrix} 4 & 1.4 \\ 1.4 & 1 \end{pmatrix}$ and $t = 1$.
- Note that

$$\|Y - X\beta\|^2 = (\beta - \hat{\beta}^{ols})^T X^T X (\beta - \hat{\beta}^{ols}) + \|Y - X\hat{\beta}^{ols}\|^2$$

and the constrained problem becomes

$$\arg \min_{\{\beta \in \mathbb{R}^p:\ \|\beta\|_\gamma^\gamma \leq t\}} \left\{ (\beta - \hat{\beta}^{ols})^T X^T X (\beta - \hat{\beta}^{ols}) \right\}.$$

- We compare the Ridge estimate

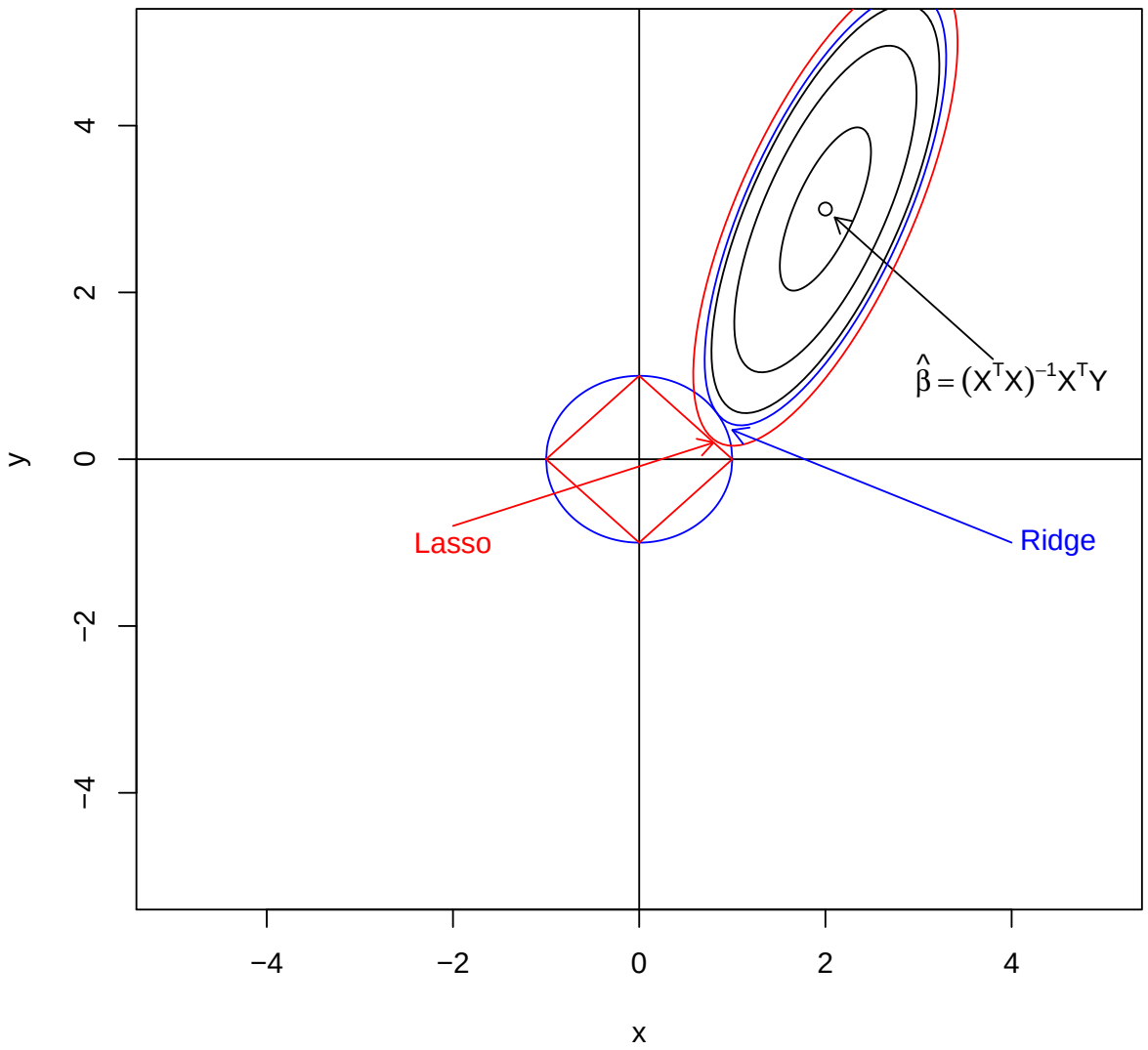
$$\hat{\beta}_\lambda^{ridge} := \arg \min_{\{\beta \in \mathbb{R}^p:\ \|\beta\|^2 \leq t\}} \left\{ (\beta - \hat{\beta}^{ols})^T X^T X (\beta - \hat{\beta}^{ols}) \right\}$$

and the Lasso estimate

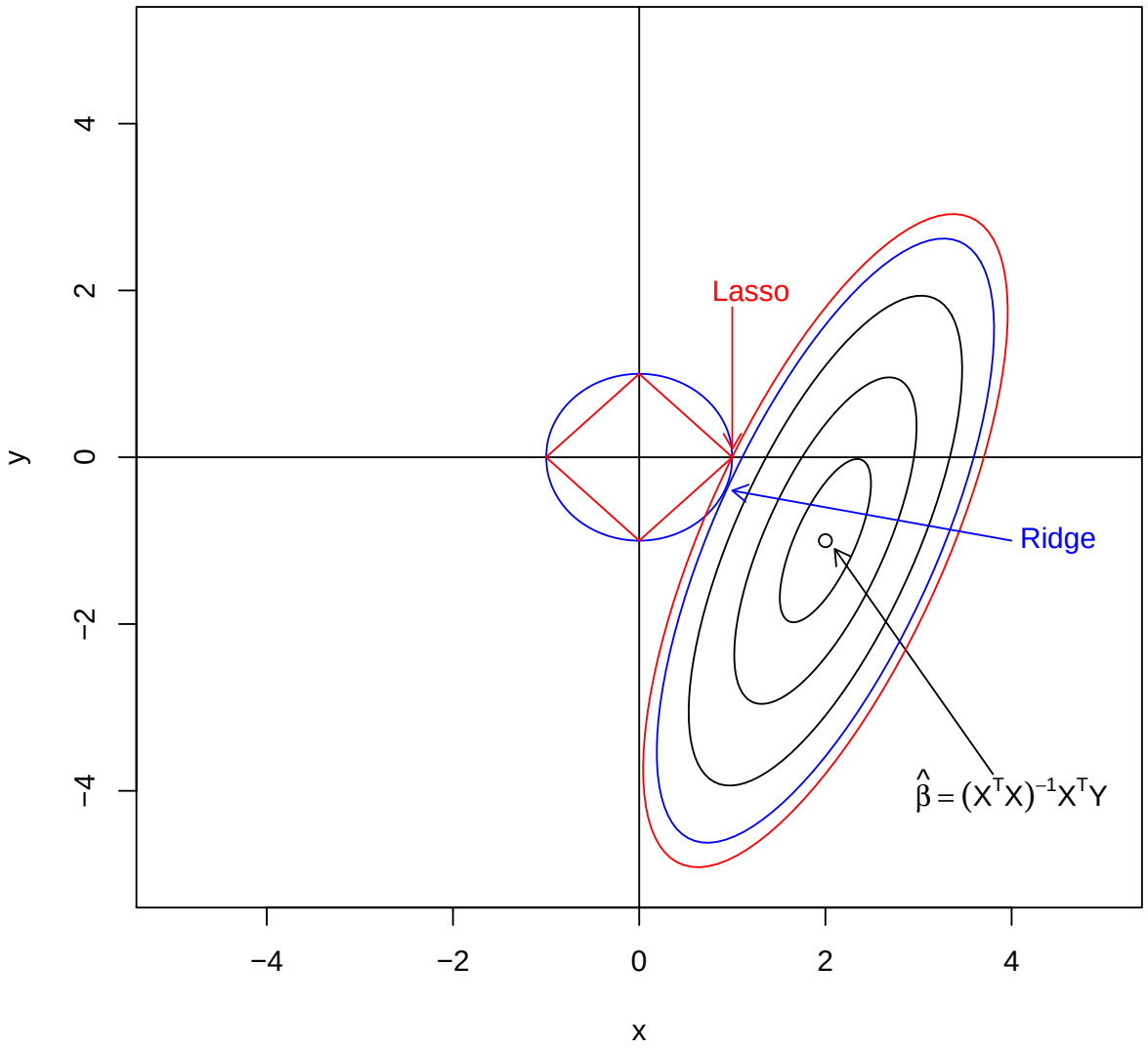
$$\hat{\beta}_\lambda^{lasso} := \arg \min_{\{\beta \in \mathbb{R}^p:\ \|\beta\|_1 \leq t\}} \left\{ (\beta - \hat{\beta}^{ols})^T X^T X (\beta - \hat{\beta}^{ols}) \right\}$$

- Of course both estimates are close (for same values of t) but, depending on $\hat{\beta}^{ols}$, Lasso estimate may have null coordinates.

Graphical illustration for $p = 2$



Graphical illustration for $p = 2$



Specific study of the case $\gamma = 1$ (the Lasso)

The Lasso, proposed by Tibshirani (1996), is the bridge estimate with $\gamma = 1$:

$$\hat{\beta}_\lambda^{lasso} := \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|^2 + \lambda \|\beta\|_1 \right\}$$

It has two specific properties:

1. It is obtained from the minimization of a convex criterion (so, with low computational cost) ☺
2. It may provide sparse solutions if the tuning parameter λ (resp. t) is large (resp. small) enough and allows for automatic selection. ☺

Theorem (Characterization of the Lasso)

A vector $\hat{\beta} \in \mathbb{R}^p$ is a global minimizer of $C_{\lambda,1}$ if and only if $\hat{\beta}$ satisfies following conditions: For any j,

- *if $\hat{\beta}_j \neq 0$, $2X_j^T(Y - X\hat{\beta}) = \lambda sign(\hat{\beta}_j)$*
- *if $\hat{\beta}_j = 0$, $|2X_j^T(Y - X\hat{\beta})| \leq \lambda$*

Furthermore, $\hat{\beta}$ is the unique minimizer if $X_{\mathcal{E}}$ is one to one with

$$\mathcal{E} := \left\{ j : \ |2X_j^T(Y - X\hat{\beta})| = \lambda \right\}$$

Specific study of the case $\gamma = 1$ (the Lasso)

- Sketch of the proof:
  1. For a convex function f, $\hat{\beta}$ is a minimum of f if and only if $0 \in \partial f(\hat{\beta})$, with
  $$\partial f(\hat{\beta}) := \left\{ s \in \mathbb{R}^p : \quad f(y) \geq f(\hat{\beta}) + \langle s, y - x \rangle, \ \forall y \right\}.$$
  2. If f is differentiable at $\hat{\beta}$, $\partial f(\hat{\beta}) = \{\nabla f(\hat{\beta})\}$
  3. If $f(\hat{\beta}) = \|\hat{\beta}\|_1$
  $$\partial f(\hat{\beta}) = \left\{ g \in \mathbb{R}^p : \ \|g\|_\infty \leq 1, \ \langle g, \hat{\beta} \rangle = \|\hat{\beta}\|_1 \right\}$$
- Note that $\hat{S} := \left\{ j : \ \hat{\beta}_j \neq 0 \right\} \subset \mathcal{E}$. So if $X_{\hat{S}}$ is one to one and $\forall j \notin \hat{S}$, $|2X_j^T(Y - X\hat{\beta})| < \lambda$, then we have uniqueness. Indeed, in this case, $\hat{S} = \mathcal{E}$.
- If $\hat{\beta}$ and $\hat{\beta}'$ are two global minimizers of $C_{\lambda,1}$, then
$$X\hat{\beta} = X\hat{\beta}' \quad \text{and} \quad \|\hat{\beta}\|_1 = \|\hat{\beta}'\|_1.$$

The orthogonal case

Assume that the matrix X is orthogonal: $X^T X = I_p$.

$$\begin{aligned}\hat{\beta}_\lambda^{lasso} &:= \arg\min_{\beta\in\mathbb{R}^p} \left\{ \|Y - X\beta\|^2 + \lambda\|\beta\|_1 \right\} \\ &= \arg\min_{\beta\in\mathbb{R}^p} \left\{ \sum_{j=1}^{p} \left(\beta_j^2 - 2(X_j^T Y)\beta_j + \lambda|\beta_j| \right) \right\}.\end{aligned}$$

Orthogonality allows for a coordinatewise study of the minimization problem. Straightforward computations lead to

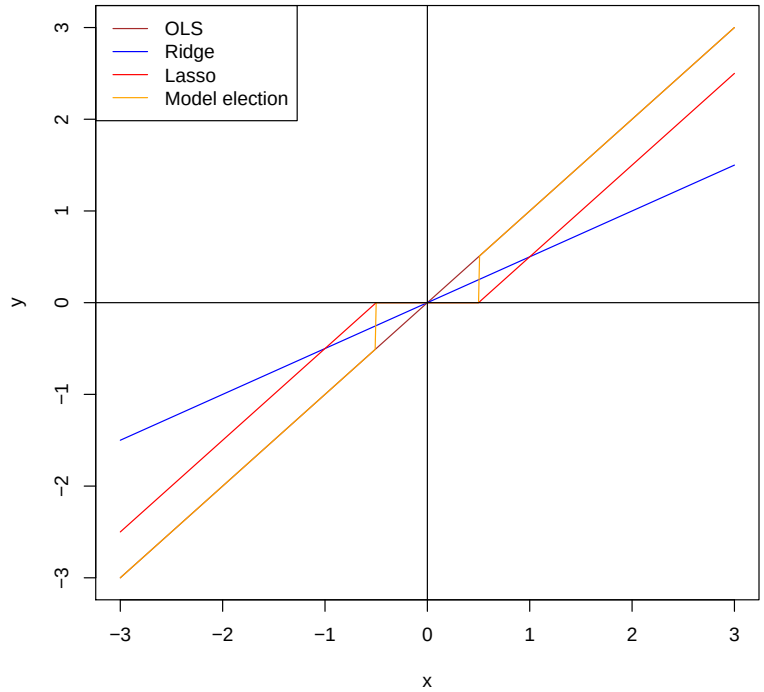
$$\begin{aligned}\hat{\beta}_{\lambda,j}^{lasso} &= \text{sign}(X_j^T Y) \times \left(|X_j^T Y| - \frac{\lambda}{2} \right)_+ \\ &= \begin{cases} X_j^T Y - \frac{\lambda}{2} & \text{if } X_j^T Y \geq \frac{\lambda}{2} \\ 0 & \text{if } -\frac{\lambda}{2} \leq X_j^T Y \leq \frac{\lambda}{2} \\ X_j^T Y + \frac{\lambda}{2} & \text{if } X_j^T Y \leq -\frac{\lambda}{2} \end{cases}\end{aligned}$$

The LASSO (Least Absolute Shrinkage and Selection Operator) procedure corresponds to a soft thresholding algorithm.

The orthogonal case - Comparison

We assume that the matrix X is orthogonal: $X^T X = I_p$. We compare (with $a_j =: X_j^T Y$):

- The OLS estimate: $\hat{\beta}_j^{ols} = a_j$
- The Ridge estimate ($\gamma = 2$):
$$\hat{\beta}_{\lambda,j}^{ridge} = (1+\lambda)^{-1} a_j$$
- The Lasso estimate or soft-thresholding rule ($\gamma = 1$):
$$\hat{\beta}_{\lambda,j}^{lasso} = \text{sign}(a_j) \times \left(|a_j| - \frac{\lambda}{2} \right)_+$$
- The Model Selection estimate or hard-thresholding rule ($\gamma = 0$):
$$\hat{\beta}_{\lambda,j}^{m.s.} = a_j \times \mathbf{1}_{\{|a_j| > \sqrt{\lambda}\}}$$



Comparison of 4 estimates for the orthogonal case with $\lambda = 1$.

Tuning the Lasso - V-fold Cross-validation

- We write the model

$$Y_i = x_i^T \beta + \epsilon_i, \quad i = 1, \ldots, n$$

with $x_i \in \mathbb{R}^p$ and ϵ_i i.i.d. $\mathbb{E}[\epsilon_i] = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2$.

- For a number V, we split the training pairs into V parts (or "folds"). Commonly, $V = 5$ or $V = 10$.
- V-fold cross-validation considers training on all but the kth part, and then validating on the kth part, iterating over $k = 1, \ldots, V$.
- When $V = n$, we call this leave-one-out cross-validation, because we leave out one data point at a time.

Tuning the Lasso - V-fold Cross-validation

1. Choose V and a discrete set Λ of possible values for λ.
2. Split the training set $\{1, \ldots, n\}$ into V subsets, $B_1, \ldots, B_V$, of roughly the same size.
3. For each value of $\lambda \in \Lambda$, for $k = 1, \ldots, V$, compute the estimate $\hat{\beta}_\lambda^{(-k)}$ on the training set $((x_i, Y_i)_{i \in B_\ell})_{\ell \neq k}$ and record the total error on the validation set B_k:
$$e_k(\lambda) := \frac{1}{\text{card}(B_k)} \sum_{i \in B_k} \left(Y_i - x_i^T \hat{\beta}_\lambda^{(-k)}\right)^2.$$
4. Compute the average error over all folds,
$$CV(\lambda) := \frac{1}{V} \sum_{k=1}^{V} e_k(\lambda) = \frac{1}{V} \sum_{k=1}^{V} \frac{1}{\text{card}(B_k)} \sum_{i \in B_k} \left(Y_i - x_i^T \hat{\beta}_\lambda^{(-k)}\right)^2.$$
5. We choose the value of tuning parameter that minimizes this function CV on Λ:
$$\hat{\lambda} := \underset{\lambda \in \Lambda}{\text{argmin}}\, CV(\lambda).$$

Tuning the Lasso - Degrees of freedom

We write the model

$$Y_i = x_i^T \beta + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \quad i = 1, \dots, n.$$

Definition (Efron (1986))

The degrees of freedom of a function $g : \mathbb{R}^n \mapsto \mathbb{R}^n$ with coordinates g_i is defined by

$$\mathrm{df}(g) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathrm{cov}(g_i(Y), Y_i).$$

The degrees of freedom may be viewed as the true number of independent pieces of informations on which an estimate is based. Example with $\mathrm{rank}(X) = p$: We estimate $X\beta^*$ with

$$g(Y) = X(X^T X)^{-1} X^T Y$$

$$\mathrm{df}(g) = \sigma^{-2} \sum_{i=1}^{n} \mathbb{E}[x_i^T (X^T X)^{-1} X^T \epsilon \times \epsilon_i] = p$$

Tuning the Lasso - Degrees of freedom

Efron's degrees of freedom is the main ingredient to generalize Mallows' C_p in high dimensions:

Proposition

Let $\hat{\beta}$ an estimate of β. If

$$C_p := \|Y - X\hat{\beta}\|^2 - n\sigma^2 + 2\sigma^2 df(X\hat{\beta}),$$

then we have:

$$\mathbb{E}[C_p] = \mathbb{E}[\|X\hat{\beta} - X\beta\|^2].$$

Assume that for any $\lambda > 0$, we have $\widehat{\mathrm{df}}$ an estimate of $\mathrm{df}(X\hat{\beta}_\lambda)$, where $\hat{\beta}_\lambda$ is the Lasso estimate associated with λ. Then, we can choose λ by minimizing

$$\lambda \mapsto \|Y - X\hat{\beta}_\lambda\|^2 + 2\sigma^2\widehat{\mathrm{df}}$$

Tuning the Lasso - Degrees of freedom

Theorem (Zou, Hastie and Tibshirani (2007))

Assume $rank(X) = p$. *Then, with*

$$\hat{S}_\lambda := \left\{ j : \quad \hat{\beta}_{\lambda,j} \neq 0 \right\},$$

we have

$$\mathbb{E}[card(\hat{S}_\lambda)] = df(X\hat{\beta}_\lambda).$$

Theorem (Tibshirani and Taylor (2012))

With

$$\mathcal{E}_\lambda := \left\{ j : \ |2X_j^T(Y - X\hat{\beta}_\lambda)| = \lambda \right\},$$

we have

$$\mathbb{E}[rank(X_{\mathcal{E}_\lambda})] = df(X\hat{\beta}_\lambda), \quad \mathbb{E}[rank(X_{\hat{S}_\lambda})] = df(X\hat{\beta}_\lambda)$$

This gives three possible estimates for $\mathrm{df}(X\hat{\beta}_\lambda)$.

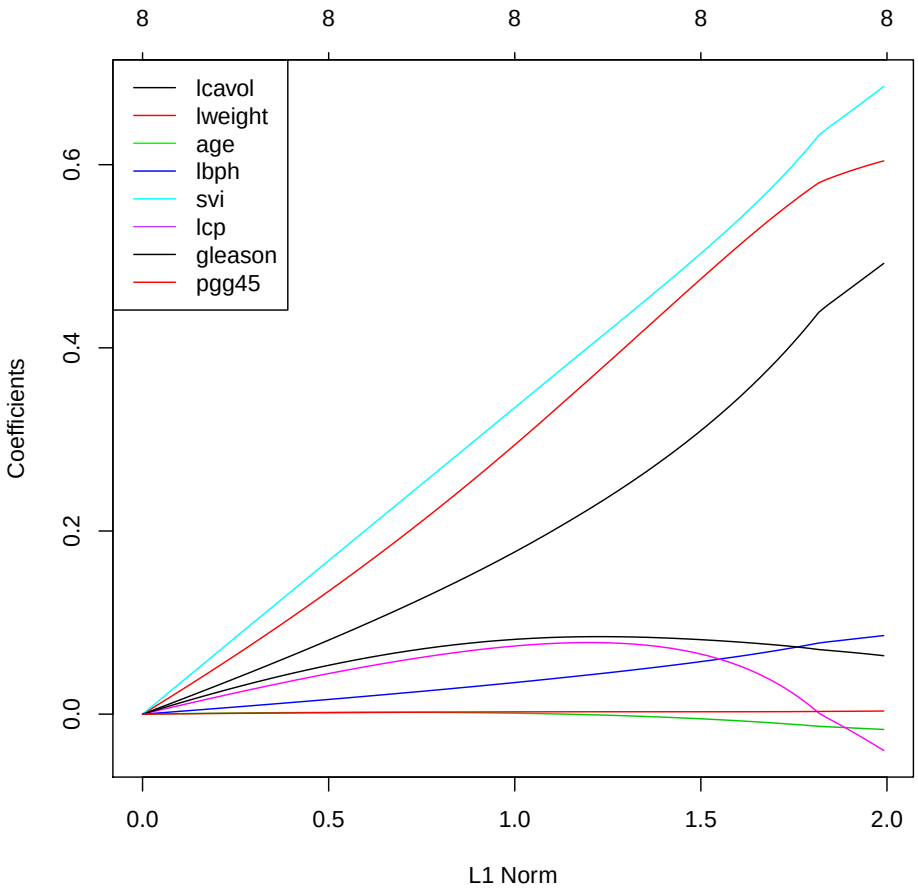
Illustration on real data

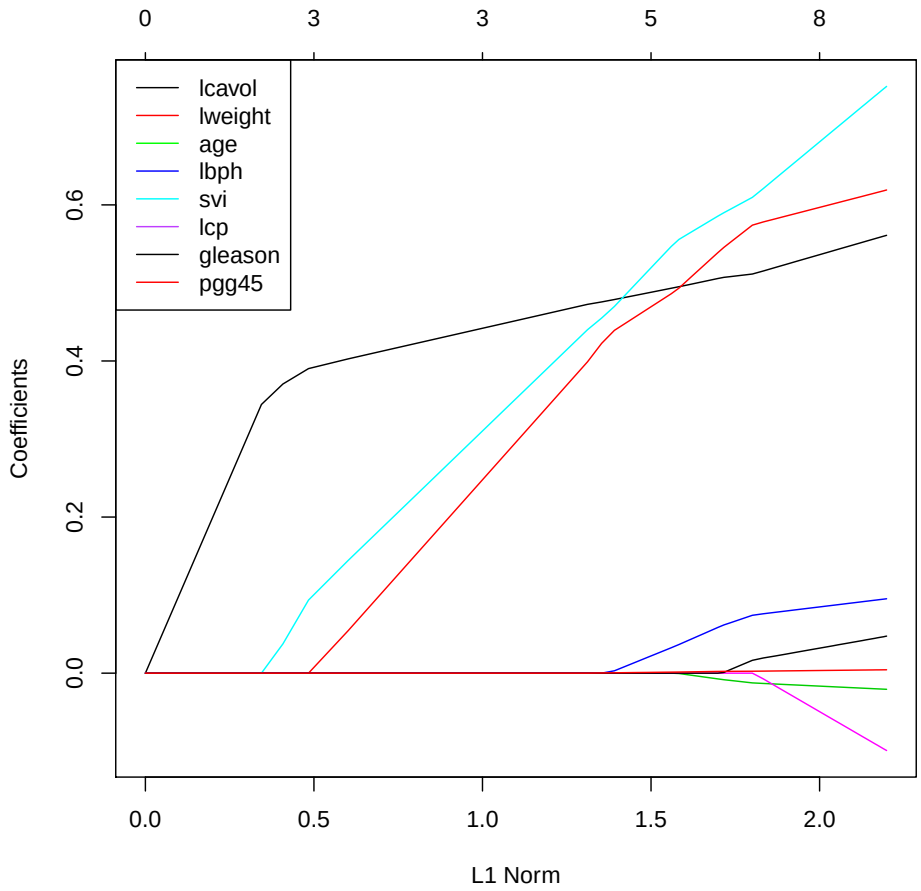
Analysis of the famous "prostate data", which records the prostate specific antigen, the cancer volume, the prostate weight, the age, the benign prostatic hyperplasia amount, the seminal vesicle invasion, the capsular penetration, the Gleason score, the percentage Gleason scores 4 or 5, for $n = 97$ patients.

```
install.packages("ElemStatLearn")
install.packages("glmnet")
library(glmnet)
data("prostate", package = "ElemStatLearn")
Y = prostate$lpsa
X = as.matrix(prostate[,names(prostate)!=c("lpsa","train")])
ridge.out = glmnet(x=X,y=Y,alpha=0)
plot(ridge.out)
lasso.out = glmnet(x=X,y=Y,alpha=1)
plot(lasso.out)
```

Theses R commands produce a plot of the values of the coordinates of the Ridge and Lasso estimates when λ decreases.

Illustration on real data





The x-axis corresponds to $\|\hat{\beta}_\lambda\|_1$. The left-hand side corresponds to $\lambda = +\infty$, the right-hand side corresponds to $\lambda = 0$.

Generalizations of the Lasso - the Dantzig selector

- Remember that the Lasso estimate satisfies the constraint

$$\max_{j=1,\dots,p} |2X_j^T(Y - X\hat{\beta}_\lambda^{lasso})| \leq \lambda.$$

We then introduce the convex set

$$\mathcal{D} := \left\{ \beta \in \mathbb{R}^p : \quad \max_{j=1,\dots,p} |2X_j^T(Y - X\beta)| \leq \lambda \right\},$$

which contains β^* with high probability if λ is well tuned.

- Remember also that we investigate sparse vectors where sparsity is measured by using the ℓ_1-norm.
- Therefore, Candès and Tao (2007) have suggested to use the Dantzig selector

$$\hat{\beta}_\lambda^{Dantzig} := \arg\min_{\beta \in \mathcal{D}} \|\beta\|_1.$$

- Note that $\|\hat{\beta}_\lambda^{Dantzig}\|_1 \leq \|\hat{\beta}_\lambda^{lasso}\|_1$. Numerical and theoretical performances of Dantzig and Lasso estimates are very close. In some cases, they may even coincide.

Generalization of the Lasso - "Adaptive" Lasso

- Due its "soft-thresholding nature", the Lasso estimation of large coefficients may suffer from a large bias. We can overcome this problem by introducing data-driven weights.
- Zou (2006) proposed an adaptive version of the classical Lasso:

$$\hat{\beta}_{\lambda}^{Zou} := \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j| \right\},$$

with

$$w_j = \frac{1}{|\hat{\beta}_j^{ols}|}.$$

- The larger $|\hat{\beta}_j^{ols}|$, the smaller w_j, which encourages large values for $\hat{\beta}_{\lambda,j}^{Zou}$.
- Instead of $\hat{\beta}^{ols}$, other preliminary estimates can be considered.

Generalizations of the Lasso - Relaxed Lasso

- Instead of introducing weights, Meinshausen (2007) suggests a two-step procedure:
  1. Compute
  $$\hat{\beta}_\lambda^{lasso} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|^2 + \lambda \|\beta\|_1 \right\}$$
  and set
  $$\hat{S}_\lambda := \left\{ j : \quad \hat{\beta}_{\lambda,j} \neq 0 \right\}.$$
  2. For $\delta \in [0,1]$,
  $$\hat{\beta}_{\lambda,\delta}^{relaxed} := \arg \min_{\beta \in \mathbb{R}^p,\, \text{supp}(\beta) \subset \hat{S}_\lambda} \left\{ \|Y - X\beta\|^2 + \delta\lambda \|\beta\|_1 \right\}$$
- If X is orthogonal,
$$\hat{\beta}_{\lambda,\delta,j}^{relaxed} = \begin{cases} X_j^T Y - \frac{\delta\lambda}{2} & \text{if } X_j^T Y \geq \frac{\lambda}{2} \\ 0 & \text{if } -\frac{\lambda}{2} \leq X_j^T Y \leq \frac{\lambda}{2} \\ X_j^T Y + \frac{\delta\lambda}{2} & \text{if } X_j^T Y \leq -\frac{\lambda}{2} \end{cases}$$
- The value $\delta = 0$ is commonly used.

Generalizations of the Lasso - The square-root Lasso

- A natural property of the Lasso estimate would be to satisfy for any $s > 0$

$$\begin{aligned}&\arg\min_{\beta\in\mathbb{R}^p}\left\{\|Y - X\beta\|^2 + \lambda\|\beta\|_1\right\}\\ &\qquad\overset{a.e.}{=} \arg\min_{\beta\in\mathbb{R}^p}\left\{\|sY - sX\beta\|^2 + \lambda\|s\beta\|_1\right\}.\end{aligned}$$

- If the tuning parameter is chosen independently of σ, the standard deviation of Y, then the Lasso estimate is not scaled invariant. The estimate

$$\arg\min_{\beta\in\mathbb{R}^p}\left\{\sigma^{-1}\|Y - X\beta\|^2 + \lambda\|\beta\|_1\right\}$$

  is scaled invariant but is based on the knowledge of σ.
- Alternatively, you can consider the square-root Lasso:

$$\arg\min_{\beta\in\mathbb{R}^p}\left\{\|Y - X\beta\| + \lambda\|\beta\|_1\right\},$$

  which also enjoys nice properties.

Generalizations of the Lasso - Elastic net

In the model $Y = X\beta^* + \epsilon$, consider

$$\hat{\beta}_\lambda^{lasso} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|^2 + \lambda \|\beta\|_1 \right\}$$

If we consider $\tilde{X} = [X, X_p]$ and if $\hat{\beta}_{\lambda,p}^{lasso} \neq 0$, then any vector $\tilde{\beta}_\lambda$ such that

$$\tilde{\beta}_{\lambda,j} = \begin{cases} \hat{\beta}_{\lambda,j}^{lasso} & \text{if } j \neq p \\ \alpha \hat{\beta}_{\lambda,p}^{lasso} & \text{if } j = p \\ (1-\alpha)\hat{\beta}_{\lambda,p}^{lasso} & \text{if } j = p+1 \end{cases},$$

with $\alpha \in [0, 1]$, is a solution of

$$\arg\min_{\beta \in \mathbb{R}^{p+1}} \left\{ \|Y - \tilde{X}\beta\|^2 + \lambda \|\beta\|_1 \right\}.$$

We have an infinite number of solutions.

Generalizations of the Lasso - Elastic net

- In practice, predictors are different but they may be strongly correlated. In this case, the Lasso estimate may hide the relevance of one of them, just because it is highly correlated to another one. Coefficients of two correlated predictors should be close.
- The elastic net procedure proposed by Zou and Hastie (2005) makes a compromise between Ridge and Lasso penalties: given $\lambda_1 > 0$ and $\lambda_2 > 0$,

$$\hat{\beta}^{e.n.}_{\lambda_1,\lambda_2} := \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|^2 \right\}.$$

  The criterion is strictly convex, so there is a unique minimizer.
- If columns of X are centered and renormalized and if Y is centered, then for $j \neq k$ such that $\hat{\beta}^{e.n.}_{\lambda_1,\lambda_2,j} \times \hat{\beta}^{e.n.}_{\lambda_1,\lambda_2,k} > 0$ then

$$\left| \hat{\beta}^{e.n.}_{\lambda_1,\lambda_2,j} - \hat{\beta}^{e.n.}_{\lambda_1,\lambda_2,k} \right| \leq \frac{\|Y\|_1}{\lambda_2} \sqrt{2(1 - X_j^* X_k)}.$$

- We can improve $\hat{\beta}^{e.n.}_{\lambda_1,\lambda_2}$ and consider $(1 + \lambda_2)\hat{\beta}^{e.n.}_{\lambda_1,\lambda_2}$

Generalizations of the Lasso - Fused Lasso

- For change point detection, for instance, for which coefficients remain constant over large portions of segments, Tibshirani, Saunders, Rosset, Zhu and Knight (2005) have introduced the fused Lasso: given $\lambda_1 > 0$ and $\lambda_2 > 0$,

$$\hat{\beta}^{fused}_{\lambda_1,\lambda_2} := \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=2}^{p} |\beta_j - \beta_{j-1}| \right\}.$$

- The first penalty is the familiar Lasso penalty which regularizes the signal. The second penalty encourages neighboring coefficients to be identical.
- We can generalize the notion of neighbors from a linear ordering to more general neighborhoods, for examples adjacent pixels in image. This leads to a penalty of the form

$$\lambda_2 \sum_{j \sim j'} |\beta_j - \beta_{j'}|.$$

Generalizations of the Lasso - Group Lasso

- To select simultaneously a group of variables, Yuan and Lin (2006) suggest to use the group-Lasso procedure. For this purpose, we assume we are given K known non-overlapping groups $G_1, G_2, \ldots, G_K$ and we set for $\lambda > 0$,

$$\hat{\beta}^{group} := \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|^2 + \lambda \sum_{k=1}^{K} \|\beta_{(k)}\| \right\},$$

  where $\beta_{(k)_j} = \beta_j$ if $j \in G_k$ and 0 otherwise.
- As for the Lasso, the group-Lasso can be characterized: $\forall\, k$,

$$\begin{cases} 2X_{(k)}^T(Y - X\hat{\beta}^{group}) = \lambda \times \frac{\hat{\beta}_{(k)}^{group}}{\|\hat{\beta}_{(k)}^{group}\|_2} & \text{if } \hat{\beta}_{(k)}^{group} \neq 0 \\ \left\| 2X_{(k)}^T(Y - X\hat{\beta}^{group}) \right\| \leq \lambda & \text{if } \hat{\beta}_{(k)}^{group} = 0 \end{cases}$$

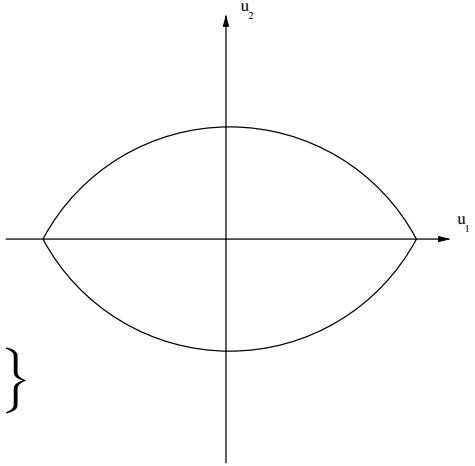
- The procedure keeps or discards all the coefficients within a block and can increase estimation accuracy by using information about coefficients of the same block.

Generalizations of the Lasso - Hierarchical group Lasso

We consider 2 predictors X_1 et X_2. Suppose we want X_1 to be included in the model before X_2. This hierarchy can be induced by defining the overlapping groups: We take $G_1 = \{1, 2\}$ et $G_2 = \{2\}$. This leads to

$$\hat{\beta}^{overlap} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|^2 + \lambda \left(\|\beta_1, \beta_2\| + |\beta_2| \right) \right\}$$

The contour plots of this penalty function is



Theorem (Zhao, Rocha et Yu (2009))

We assume we are given K known groups $G_1, G_2, \ldots, G_K$. Let $\mathcal{I}_1$ and $\mathcal{I}_2 \subset \{1, \ldots, p\}$ be two subsets of indices. We assume:

1. *For all $1 \leq k \leq K$, $\mathcal{I}_1 \subset G_k \Rightarrow \mathcal{I}_2 \subset G_k$.*
2. *There exists k_0 such that $\mathcal{I}_2 \subset G_{k_0}$ and $\mathcal{I}_1 \not\subset G_{k_0}$.*

Then, almost surely, $\hat{\beta}^G_{\mathcal{I}_2} \neq 0 \Rightarrow \hat{\beta}^G_{\mathcal{I}_1} \neq 0$.

Generalizations of the Lasso - The Bayesian Lasso

- In the Bayesian approach, the parameter is random and we write:

$$Y|\beta,\sigma^2 \sim \mathcal{N}(X\beta,\sigma^2 I_n)$$

- Park and Casella (2008) suggest to consider a Laplace distribution for β:

$$\beta|\lambda,\sigma \sim \prod_{j=1}^{p}\left[\frac{\lambda}{2\sigma}\exp\left(-\frac{\lambda}{\sigma}|\beta_j|\right)\right].$$

  Then, the posterior density is

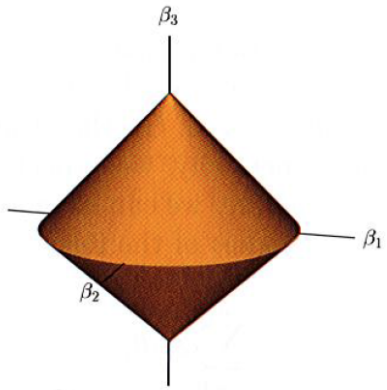
$$\propto \exp\left(-\frac{1}{2\sigma^2}\|Y-X\beta\|^2-\frac{\lambda}{\sigma}\|\beta\|_1\right)$$

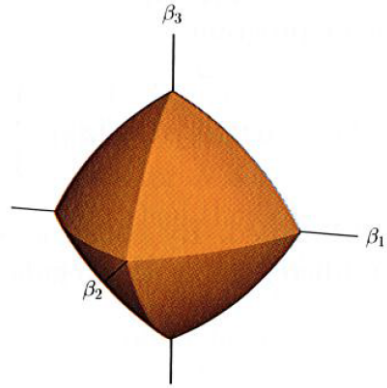
  and the posterior mode coincides with the Lasso estimate with smoothing parameter $\sigma\lambda$.
- The posterior distribution provides more than point estimates since it provides the entire posterior distribution.
- The procedure is tuned by including priors for σ^2 and λ.
- Most of Lasso-type procedures have a Bayesian interpretation.

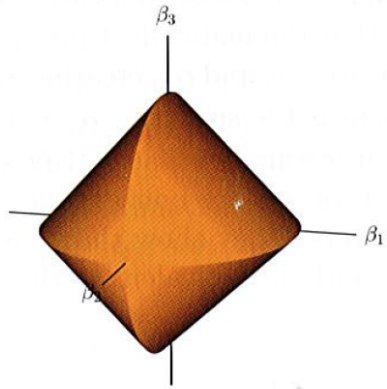
Geometric constraints area

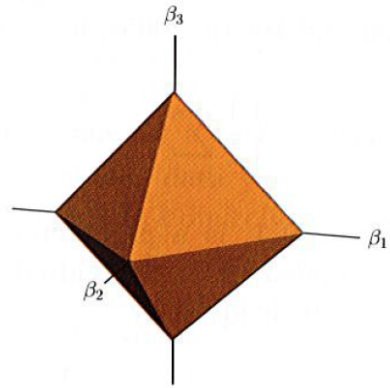
Exercice : Connect each methodology to its associated geometric constraints area.

1. Lasso
2. Elastic net
3. Group-Lasso
4. Overlap group-Lasso









Theoretical guarantees - Support recovery

Question: Are the non-zero entries of the Lasso estimate $\hat{\beta}_\lambda^{lasso}$ in the same positions as the true regression vector β^*? We set

$$S^* := \left\{j : \quad \beta_j^* \neq 0\right\}, \quad \hat{S}_\lambda := \left\{j : \quad \hat{\beta}_{\lambda,j}^{lasso} \neq 0\right\}$$

We can identify conditions under which $\hat{S}_\lambda = S^*$.

Theorem

We assume that for some $\gamma > 0$, $K > 0$ and $c_{min} > 0$,

$$\max_{j \notin S^*} \|(X_{S^*}^T X_{S^*})^{-1} X_{S^*}^T X_j\|_1 \leq 1 - \gamma,$$

$$\max_{j=1,\ldots,p} \|X_j\| \leq K, \quad eig(X_{S^*}^T X_{S^*}) \geq c_{min}.$$

Then, if $\lambda \geq \frac{8K\sigma\sqrt{\log p}}{\gamma}$, with probability larger than $1 - p^{-A}$,

$$\hat{S}_\lambda \subset S^*, \quad \|\hat{\beta}_\lambda^{lasso} - \beta^*\|_\infty \leq \lambda \left(\frac{4\sigma}{\sqrt{c_{min}}} + \|(X_{S^*}^T X_{S^*})^{-1}\|_\infty \right)$$

Theoretical guarantees - Prediction bounds

Theorem (Bunea *et al.* (2007))

Let us consider $\lambda \geq 3 \max_{j=1,\dots,p} |(X^T \epsilon)_j|$. *For any* $\beta \in \mathbb{R}^p$, *let*

$$\kappa(\beta) := \min_{\nu \in C(\beta)} \frac{\|X\nu\|^2}{\|\nu\|^2},$$

$$C(\beta) := \{\nu \in \mathbb{R}^p : \ 20\|\nu\|_{1,Supp(\beta)} > \|\nu\|_{1,Supp(\beta)^c}\}.$$

Then, if $\kappa(\beta) > 0$,

$$\|X\hat{\beta}_\lambda^{lasso} - X\beta^*\|^2 \leq \inf_{\beta \in \mathbb{R}^p} \left\{ 3\|X\beta - X\beta^*\|^2 + \frac{32\lambda^2 \|\beta\|_0}{\kappa(\beta)} \right\}.$$

- Deriving λ such that the first bound is satisfied with high probability is easy by using concentration inequalities.
- The Restricted Eigenvalues Condition expresses the lack of orthogonality of columns of X. Milder conditions can be used (see Negahban *et al.* (2012) or Jiang *et al.* (2017))

Theoretical guarantees - Prediction bounds

- From the previous theorem, we can deduce estimation bounds for ℓ_2 and ℓ_1 norms for estimating sparse vectors β^* (see Jiang *et al.* (2017)) :

$$\|\hat{\beta}_\lambda^{lasso} - \beta^*\|^2 \lesssim \lambda^2 \|\beta^*\|_0$$

$$\|\hat{\beta}_\lambda^{lasso} - \beta^*\|_1 \lesssim \lambda \|\beta^*\|_0$$

- The proof of the Theorem is based on the following lemma.

Lemma

Let $\lambda \geq 3 \max_{i=1,\dots,n} |(X^T \epsilon)_i|$ *and* $\beta \in \mathbb{R}^p$. *Then,*

$$\lambda \|\hat{\beta}_\lambda^{lasso} - \beta\|_{1,Supp(\beta)^c} \leq 3\|X\beta - X\beta^*\|^2 + 5\lambda \|\hat{\beta}_\lambda^{lasso} - \beta\|_{1,Supp(\beta)}$$

$$\|X\hat{\beta}_\lambda^{lasso} - X\beta^*\|^2 \leq \|X\beta - X\beta^*\|^2 + 2\lambda \|\hat{\beta}_\lambda^{lasso} - \beta\|_{1,Supp(\beta)}$$

- Better constants can be obtained via a more involved proof.

Take-home message

- To overcome prohibitive computational complexity of model selection, convex critera can be considered leading, in particular, to Lasso-type estimates.
- By doing so, we introduce some bias but reduce the variance of predicted values. Moreover, we can identify a small number of predictors that have the strongest effects and then makes interpretation easier for the practitioner.
- By varying the basic Lasso ℓ_1-penalty, we can reduce problems encountered by the standard Lasso or incorporate some prior knowledge about the model.
- In the linear regression setting, these estimates, which can be easily computed, are very popular for high dimensional statistics. They achieve nice theoretical and numerical properties.
- Even if some standard recipes can be used to tune the Lasso, its calibration remains an important open problem.

References

- Bunea, F., Tsybakov, A. and Wegkamp, M. Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.* 1, 169–194, 2007.

- Candes, E. and Tao, T. The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6), 2313–2351, 2007.

- Chen, S., Donoho, D. and Saunders, M. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.* 20 (1998), no. 1, 33–61, 1998.

- Efron, B. How biased is the apparent error rate of a prediction rule?, *Journal of the American Statistical Association: Theory and Methods* 81 (394), 461–470, 1986.

- Jiang X., Reynaud-Bouret P., Rivoirard V., Sansonnet L. and Willet R. A data-dependent weighted LASSO under Poisson noise. Submitted, 2017

- Meinshausen, N. Relaxed Lasso. *Comput. Statist. Data Anal.*, 52(1),374–393, 2007.

References

- Negahban, S., Ravikumar, P., Wainwright, M. and Yu, B. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.*, 27(4), 538–557, 2012.

- Park, T. and Casella, G. The Bayesian Lasso. *J. Amer. Statist. Assoc.*, 103(482), 681–686, 2008.

- Tibshirani, R. and Taylor, J. Degrees of freedom in lasso problems. *Ann. Statist.* 40(2), 1198–1232, 2012.

- Tibshirani, R. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* 58, no. 1, 267–288, 1996

- Tibshirani, R., Saunders, M. Rosset, S. Zhu, J. and Knight, K. Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(1), 91–108, 2005.

- Yuan, M. and Lin, Y. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(1), 49–67, 2006.

References

- Zhao, P., Rocha, G. and Yu, B. The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Statist.*, 37(6A), 3468–3497, 2009.

- Zou, H. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476), 1418–1429, 2006.

- Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(2), 301–320, 2005.

- Zou, H., Hastie, T. and Tibshirani, R. On the "degrees of freedom" of the lasso. *The Annals of Statistics*, 35 (5), 2173–2192, 2007.

Chapter 3: Generalized linear Models and related models

- In the last chapter, we considered estimation in the classical Gaussian linear model with a special emphasis on the Lasso estimate. We studied variations of the classical Lasso penalty.
- In this chapter, we generalize the previous setting by varying the loss function.

- Contents of the chapter: Among GLM, we shall lay special emphasis on
  1. The linear logistic model
  2. The Cox Proportional Hazards model
  3. The Poisson model

  The Poisson model is exemplified according to 3 illustrations:
  - a) Variable selection
  - b) Poisson inverse problem
  - c) Functional Poisson regression

Introduction

- Generalized Linear Models (McCullagh and Nelder (1989)) describe the response variable using a member of the exponential family, which includes the Bernoulli, Poisson and Gaussian as particular cases.
- For Generalized Linear Models (GLM), a transformation of the mean of the response variable is approximated by linear combinations of predictors: We assume that there exist a function g and $\beta^* \in \mathbb{R}^p$ such that for any $i = 1, \ldots, n$,

$$g\big(\mathbb{E}[Y_i]\big) = \sum_{j=1}^{p} \beta_j^* X_{ij},$$

  where the Y_i's are coordinates of the response vector assumed to be independent.
- We discuss two approaches based on minimizing penalized criteria. The criterion will be either the least squares criterion or the opposite of the log-likelihood (they coincide for the Gaussian case). The penalty will be a Lasso-type penalty.

First example: Logistic regression

- We consider the case where $Y_i \in \{0, 1\}$. Linear logistic regression models $\mathbb{E}[Y_i] = \mathbb{P}(Y_i = 1)$ by

$$\mathbb{E}[Y_i] = \frac{e^{\sum_{j=1}^p \beta_j^* X_{ij}}}{1 + e^{\sum_{j=1}^p \beta_j^* X_{ij}}} \iff \log\left(\frac{\mathbb{E}[Y_i]}{1 - \mathbb{E}[Y_i]}\right) = \sum_{j=1}^p \beta_j^* X_{ij}$$

  i.e. g is the logit function: $g(x) = \log\left(\frac{x}{1-x}\right)$. The log-likelihood is a concave function:

$$\mathcal{L}(\beta) = \sum_{i=1}^n \left[Y_i \sum_{j=1}^p \beta_j^* X_{ij} - \log\left(1 + e^{\sum_{j=1}^p \beta_j^* X_{ij}}\right) \right].$$

- Ideal for binary classification problems. Very popular in many fields.
- Classical Lasso estimate (Tibshirani (1996)):

$$\hat{\beta}_\lambda^{lasso} := \arg\min_{\beta \in \mathbb{R}^p} \{-\mathcal{L}(\beta) + \lambda \|\beta\|_1\}$$

First example: Logistic regression

- Group-Lasso estimate (Meier *et al.* (2008)): given K known non-overlapping groups $G_1, G_2, \ldots, G_K$, we set for $\lambda > 0$,

$$\hat{\beta}^{group} := \arg\min_{\beta \in \mathbb{R}^p} \left\{ -\mathcal{L}(\beta) + \lambda \sum_{k=1}^{K} \|\beta_{(k)}\| \right\},$$

  where $\beta_{(k)_j} = \beta_j$ if $j \in G_k$ and 0 otherwise.
- These estimates are consistent under mild assumptions.
- Convexity allows for deriving fast coordinate gradient descent algorithms.
- The previous approach can be extended for $K > 2$ outputs:

$$\mathbb{P}(Y_i = k) = \frac{e^{\sum_{j=1}^{p} \beta^*_{(k)j} x_{ij}}}{\sum_{\ell=1}^{K} e^{\sum_{j=1}^{p} \beta^*_{(\ell)j} x_{ij}}}, \quad k = 1, \ldots, K.$$

  Note that this model is over-parametrized and $(\beta^*_{(k)j} + c_j)_{k,j}$ and $(\beta^*_{(k)j})_{k,j}$ give the same model for any $c_j \in \mathbb{R}$.

Second example: Cox model

- We consider the usual setup of survival analysis and in particular right-censoring models that are very popular in biomedical problems.
- For each patient, we consider its lifetime T (with density f) that can be censored. We denote by C the independent censoring time. We face with censoring when, for instance, the patient drops out of a hospital study: the time of death is not observed, but we know that the patient was still alive when he left the study.
- So, we observe (Z, δ), with

$$Z = \min\{T, C\} \quad \text{and} \quad \delta = \mathbf{1}_{T \leq C}.$$

- A key quantity is the survival function defined by $S(t) = \mathbb{P}(T > t)$, the probability of surviving beyond the time t. We denote by h the common hazard rate of T:

$$h(t) := \lim_{\epsilon \to 0} \frac{\mathbb{P}(T \in (t, t+\epsilon) | T > t)}{\epsilon} = \frac{f(t)}{S(t)}.$$

It corresponds to the instantaneous probability of death at time t, given survival up till t.

Second example: Cox model

- We consider n patients and, for each patient i, we consider its lifetime T_i. We denote by C_i the independent censoring time.
- We observe (Z_i, δ_i), with

$$Z_i = \min\{T_i, C_i\} \quad \text{and} \quad \delta_i = \mathbf{1}_{T_i \leq C_i}.$$

  We assume that the vectors $(T_i, C_i)_{1 \leq i \leq n}$ are independent.
- We also observe for each patient, realizations of p predictors (for instance, the measure of p genes expression).
- For Cox regression:
  1. Given the vector of predictors $(X_{ij})_{j=1,\ldots,p}$ associated with patient i, the basic assumption is that any two patients have hazard functions whose ratio is constant (w.r.t t), so
  $$h^{(i)}(t) = h_0(t)\rho_i.$$
  2. It is assumed that there exists $\beta^* \in \mathbb{R}^p$ such that

$$\rho_i = \exp\left(\sum_{j=1}^{p} \beta_j^* X_{ij}\right) > 0 \Rightarrow h^{(i)}(t) = h_0(t) \exp\left(\sum_{j=1}^{p} \beta_j^* X_{ij}\right).$$

Second example: Cox model

- To estimate β^* in high dimensions, we penalize the (concave) log-likelihood function whose expression is given by

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} \delta_i \left(\sum_{j=1}^{p} \beta_j X_{ij} - \log \Big(\sum_{k:\ Z_k > Z_i} \exp \Big(\sum_{j=1}^{p} \beta_j X_{kj} \Big) \Big) \right).$$

  We have assumed that there are no ties (survival times are unique).

- For $\lambda > 0$, we set

$$\hat{\beta}_\lambda^{lasso} \quad := \quad \arg\min_{\beta \in \mathbb{R}^p} \left\{ -\mathcal{L}(\beta) + \lambda \|\beta\|_1 \right\}.$$

- Note that the baseline function h_0 does not play any role and we only investigate influence of predictors.
- See Tibshirani (1997) for a concrete study.

Third example: Poisson regression

- When the response variable is nonnegative and represents a count, the Poisson distribution is extensively used.
- Poisson models are often used to model death rates.
- Illustration 1: Variable selection. Similarly, we write:

$$\log\left(\mathbb{E}[Y_i]\right) = \sum_{j=1}^{p} \beta_j^* X_{ij}$$

i.e. $g(x) = \log(x)$ in the GLM setting. With $\mu_i = \mathbb{E}[Y_i]$, since

$$Y_i \sim \mathcal{P}oisson(\mu_i) \iff \mathbb{P}(Y_i = k) = e^{-\mu_i} \frac{\mu_i^k}{k!}, \quad k = 0, 1, 2, \ldots,$$

assuming that the Y_i's are independent, the log-likelihood is

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} \left[Y_i \sum_{j=1}^{p} \beta_j X_{ij} - \exp\left(\sum_{j=1}^{p} \beta_j X_{ij} \right) \right]$$

and for $\lambda > 0$, we set again

$$\hat{\beta}_\lambda^{lasso} := \arg\min_{\beta \in \mathbb{R}^p} \left\{ -\mathcal{L}(\beta) + \lambda \|\beta\|_1 \right\}.$$

Poisson inverse problem

Illustration 2: Poisson inverse problem

- We observe a potentially random matrix $A = (a_{k,\ell})_{k,\ell} \in \mathbb{R}_+^{n \times p}$ and conditionally on A, we observe

$$Z|A \sim \mathcal{P}oisson(A\beta^*),$$

where $Z \in \mathbb{R}_+^n$, $\beta^* \in \mathbb{R}_+^p$ is sparse and the elements of Z are independent. The aim is to recover β^*.

- The matrix A corresponds to a sensing matrix which linearly projects β^* into another space before we collect Poisson observations.
- Example: Photon-limited compressive imaging
  A widely-studied compressed sensing measurement matrix is the Bernoulli ensemble, in which each element of A is drawn iid from a Bernoulli(q) distribution for $q \in (0, 1)$. (Typically, $q = 1/2$) The celebrated Rice single-pixel camera (Duarte *et al.* (2008)) uses exactly this model to position the micromirror array for each projective measurement.

Poisson inverse problem

- Remember the RE assumption for the matrix A: For any $\beta \in \mathbb{R}^p$,

$$\kappa(\beta) := \min_{\nu \in C(\beta)} \frac{\|A\nu\|^2}{\|\nu\|^2} > 0.$$

  But since elements of A are nonnegative, we cannot rely on the RE assumption that expresses that $A^T A$ is not far from I_p.
- In many settings, there is a proxy operator, denoted X, which is amenable to sparse inverse problems and is a simple linear transformation of the original operator A. A complementary linear transformation may then be applied to Z to generate proxy observations Y, and we use X and Y.
- Example: Photon-limited compressive imaging:

$$X = \frac{A}{\sqrt{nq(1-q)}} - \frac{q\mathbf{1}_{n\times 1}\mathbf{1}_{p\times 1}^T}{\sqrt{nq(1-q)}} \Rightarrow \mathbb{E}[X^T X] = I_p.$$

Poisson inverse problem

- Standard Lasso:

$$\hat{\beta}^{s.l.} := \arg\min_{\beta \in \mathbb{R}^p} \left\{ \| Y - X\beta \|^2 + \lambda \|\beta\|_1 \right\}$$

- Weighted Lasso:

$$\hat{\beta}^{w.l.} := \arg\min_{\beta \in \mathbb{R}^p} \left\{ \| Y - X\beta \|^2 + \sum_{k=1}^{p} \lambda_k |\beta_k| \right\}$$

- Weights have to satisfy

$$3|(X^T(Y - X\beta^*))_k| \leq \lambda \, (\text{resp. } \lambda_k), \quad \text{for } k = 1, \ldots, p. \tag{4.1}$$

The role of the weights is twofold:
- control of the random fluctuations of $X^T Y$ around its mean
- compensate for the ill-posedness due to X. Ill-posedness is strengthened by the heteroscedasticity of the Poisson noise.

- Example: Photon-limited compressive imaging : if

$$Y = \frac{1}{(n-1)\sqrt{nq(1-q)}} \left(nZ - \sum_{\ell=1}^{n} Z_\ell \mathbf{1}_{n \times 1} \right), \quad \mathbb{E}[X^T(Y - X\beta^*)] = 0$$

Poisson inverse problem

- Jiang *et al.* (2017) proved that, for photon-limited compressive imaging, the RE assumption is satisfied with $\min_\beta \kappa(\beta) > 0$. If β^* is s-sparse, under suitable conditions on s, if (4.1) is satisfied, then with large probability,

$$\|\hat{\beta}^{s.l.} - \beta^*\|^2 \lesssim s\lambda^2, \quad \|\hat{\beta}^{w.l.} - \beta^*\|^2 \lesssim \sum_{k=1}^{p} \lambda_k^2 \mathbf{1}_{\{\beta_k^* \neq 0\}}.$$

- We have to choose weights as small as possible such that (4.1) is satisfied.
- Still for photon-limited compressive imaging, assume

$$s \ll \frac{nq^2}{\log p} \quad \text{and} \quad q \gg \left(\frac{\log p}{n}\right)^{1/3}.$$

If q is small and s is large, $\hat{\beta}^{w.l.}$ achieves better rates since

$$\|\hat{\beta}^{s.l.} - \beta^*\|^2 \lesssim \frac{s\|\beta^*\|_1 \log p}{qn}, \quad \|\hat{\beta}^{w.l.} - \beta^*\|^2 \lesssim \frac{\log p \|\beta^*\|_1}{n}\left(s + \frac{1}{q}\right).$$

Functional Poisson regression

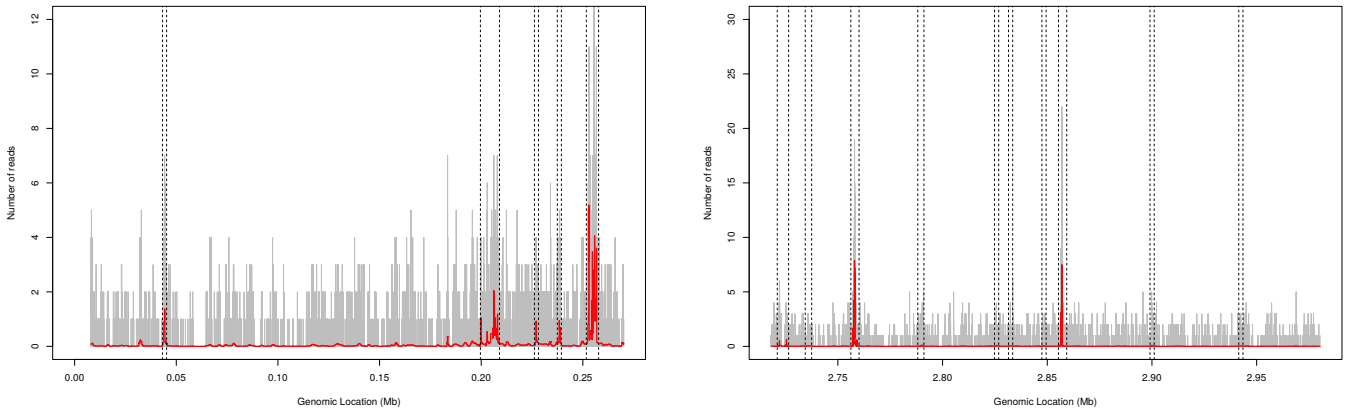
Illustration 3: Functional Poisson regression

- We consider the functional Poisson regression model, with n observed counts Y_i modeled by

$$Y_i \sim \mathcal{P}oisson(f_0(X_i)), \quad X_i \in [0,1].$$

The X_i's are naturally ordered (times or positions)

- Example (Ivanoff *et al.* (2016)) :



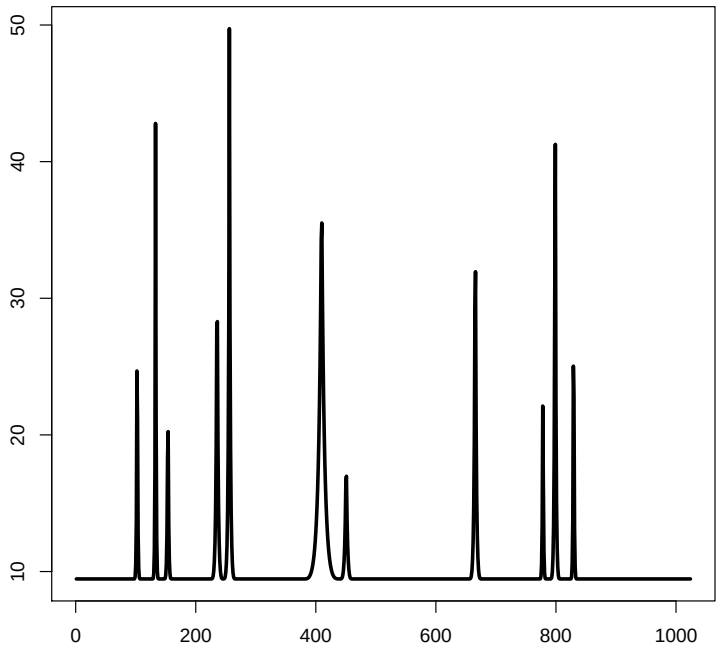
Estimation of the intensity function of Ori-Seq data (chromosomes 20 and X). Grey bars indicate the number of reads that match genomic positions (x-axis, in MegaBases). The red line corresponds to the estimated intensity function.

Functional Poisson regression

- In the model
$$Y_i \sim \mathcal{P}oisson(f_0(X_i)), \quad X_i \in [0,1]$$
the goal is to estimate the function f_0.
- Let f a candidate estimator of f_0 decomposed on a functional dictionary with p elements denoted $\Upsilon = \{\varphi_j\}_{j=1,\dots,p}$.
- f is assumed to be positive, so we set
$$\log(f) = \sum_{j=1}^{p} \beta_j \varphi_j, \quad \beta = (\beta_j)_{j=1,\dots,p}.$$
- We enrich the standard basis approach: We assume that $\log(f_0)$ is well approximated by a sparse linear combination of Υ.

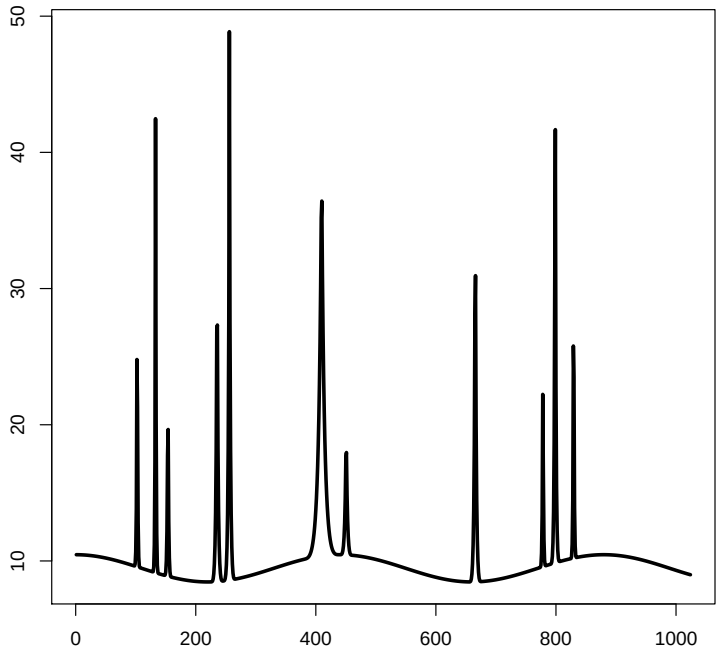
Functional Poisson regression

- The basis approach is designed to catch specific features of the signal ($p = n$).
- If many features are present simultaneously?
- Consider overcomplete dictionaries ($p > n$).
- Typical dictionaries: Histograms, Daubechies wavelets, Fourier,
- How to select the dictionary elements?



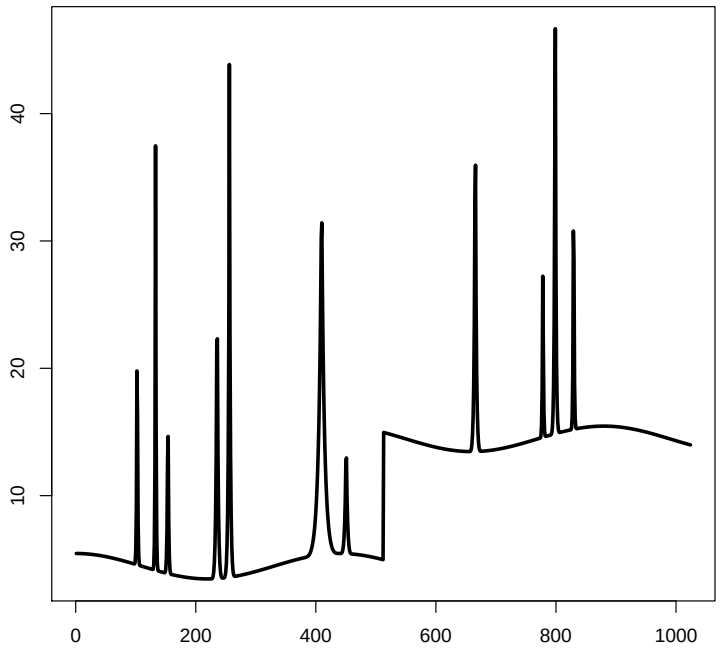
Functional Poisson regression

- The basis approach is designed to catch specific features of the signal ($p = n$).
- If many features are present simultaneously?
- Consider overcomplete dictionaries ($p > n$).
- Typical dictionaries: Histograms, Daubechies wavelets, Fourier,
- How to select the dictionary elements?



Functional Poisson regression

- The basis approach is designed to catch specific features of the signal ($p = n$).
- If many features are present simultaneously?
- Consider overcomplete dictionaries ($p > n$).
- Typical dictionaries: Histograms, Daubechies wavelets, Fourier,
- How to select the dictionary elements?



Functional Poisson regression

- We consider a likelihood-based penalized criterion to select β. Let $X_{ij} = \varphi_j(X_i)$. Then, the log-likelihood is, as previously,

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} \left[Y_i \sum_{j=1}^{p} \beta_j X_{ij} - \exp\left(\sum_{j=1}^{p} \beta_j X_{ij} \right) \right]$$

- Two different ways of penalizing $-\mathcal{L}(\beta)$ are proposed:
  1. Standard Lasso: Given positive weights $(\lambda_j)_j$,

$$\hat{\beta} := \arg\min_{\beta \in \mathbb{R}^p} \left\{ -\mathcal{L}(\beta) + \sum_{j=1}^{p} \lambda_j |\beta_j| \right\}.$$

  2. Group-Lasso: We partition the set of indices $\{1, \ldots, p\}$ into K non-empty groups: $\{1, \ldots, p\} = G_1 \cup G_2 \cup \cdots \cup G_K$.

$$\hat{\beta}^{gL} := \arg\min_{\beta \in \mathbb{R}^p} \left\{ -\mathcal{L}(\beta) + \sum_{k=1}^{K} \lambda_k^g \|\beta_{G_k}\|_2 \right\},$$

  where the λ_k^g's are positive weights and β_{G_k} stands for the sub-vector of β with elements indexed by the elements of G_k.

Functional Poisson regression

For any j, we choose a data-driven value for λ_j as small as possible so that with high probability, for any $j \in \{1, \dots, p\}$,

$$|(X^T(Y - \mathbb{E}[Y]))_j| \leq \lambda_j.$$

This leads to involved formula.

Theorem (Ivanoff *et al.* (2016))

Let $\gamma > 0$ be a constant. Define $\widehat{V}_j = \sum_{i=1}^n \varphi_j^2(X_i) Y_i$ and

$$\widetilde{V}_j = \widehat{V}_j + \sqrt{2\gamma \log p \widehat{V}_j \max_i \varphi_j^2(X_i)} + 3\gamma \log p \max_i \varphi_j^2(X_i).$$

Let

$$\lambda_j = \sqrt{2\gamma \log p \widetilde{V}_j} + \frac{\gamma \log p}{3} \max_i |\varphi_j(X_i)|,$$

then

$$\mathbb{P}\Big(|(X^T(Y - \mathbb{E}[Y]))_j| > \lambda_j\Big) \leq \frac{3}{p^\gamma}.$$

Functional Poisson regression

- For the group-Lasso, we choose the weights λ_k^g as small as possible so that with high probability, for any $k \in \{1, \ldots, K\}$,

$$\|X_{G_k}^T (Y - \mathbb{E}[Y])\| \leq \lambda_k^g$$

- This is the analog of standard Lasso weights but with absolute values replaced by ℓ_2-norms. Formula are much more involved but derived weights are nevertheless data-driven.
- We show that the associated Lasso/Group Lasso procedures are theoretically optimal (oracle inequalities for the Kullback divergence).
- With a theoretical form for the weights, much computing power is spared.
- See Ivanoff *et al.* (2016) for more details.

Functional Poisson regression

- We considered the classical Donoho & Johnstone functions (Blocks, bumps, doppler, heavisine).
- The intensity function f_0 is set such that (with $\alpha \in \{1, \dots, 7\}$)

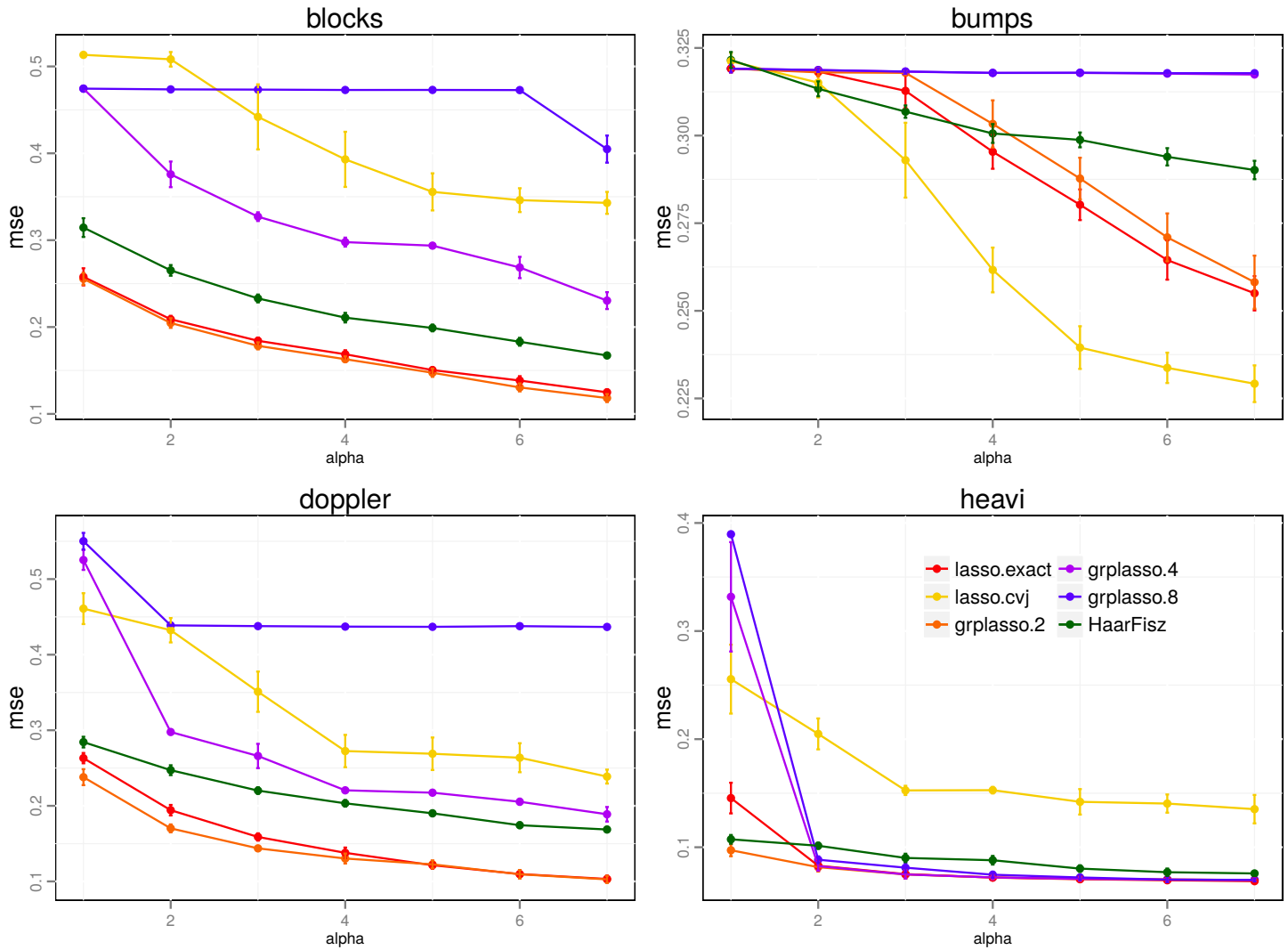
$$f_0 = \alpha \exp g_0$$

- Observations are sampled on a fixed regular grid ($n = 2^{10}$) with $Y_i \sim \mathcal{P}(f_0(X_i))$.
- Use Daubechies basis, Haar basis and Fourier as elements of the dictionary.
- Check the normalized reconstruction error:

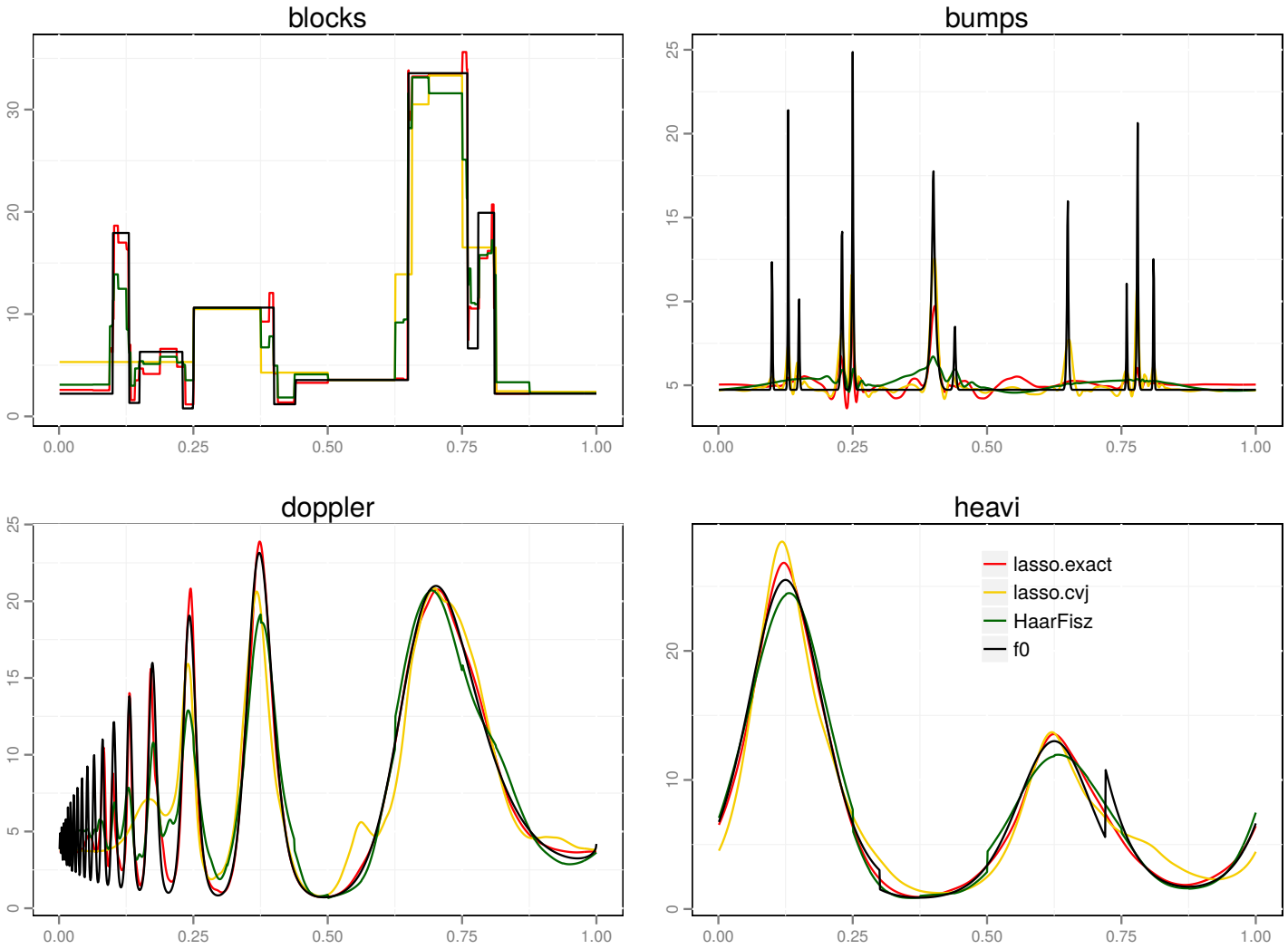
$$MSE = \frac{\|\widehat{f} - f_0\|_2^2}{\|f_0\|_2^2}$$

- Compete with the Haar-Fisz transform and cross-validation.

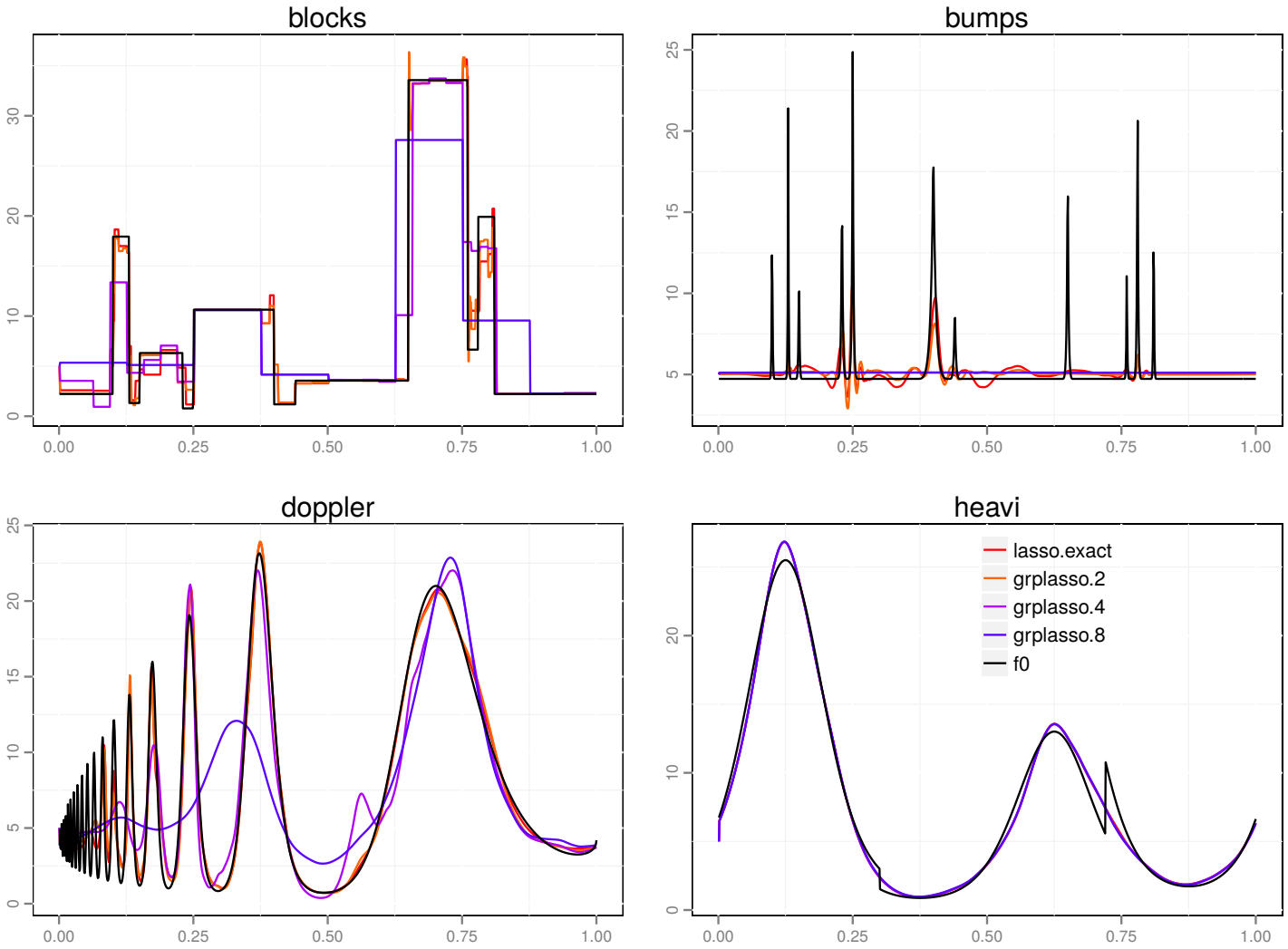
Reconstruction errors



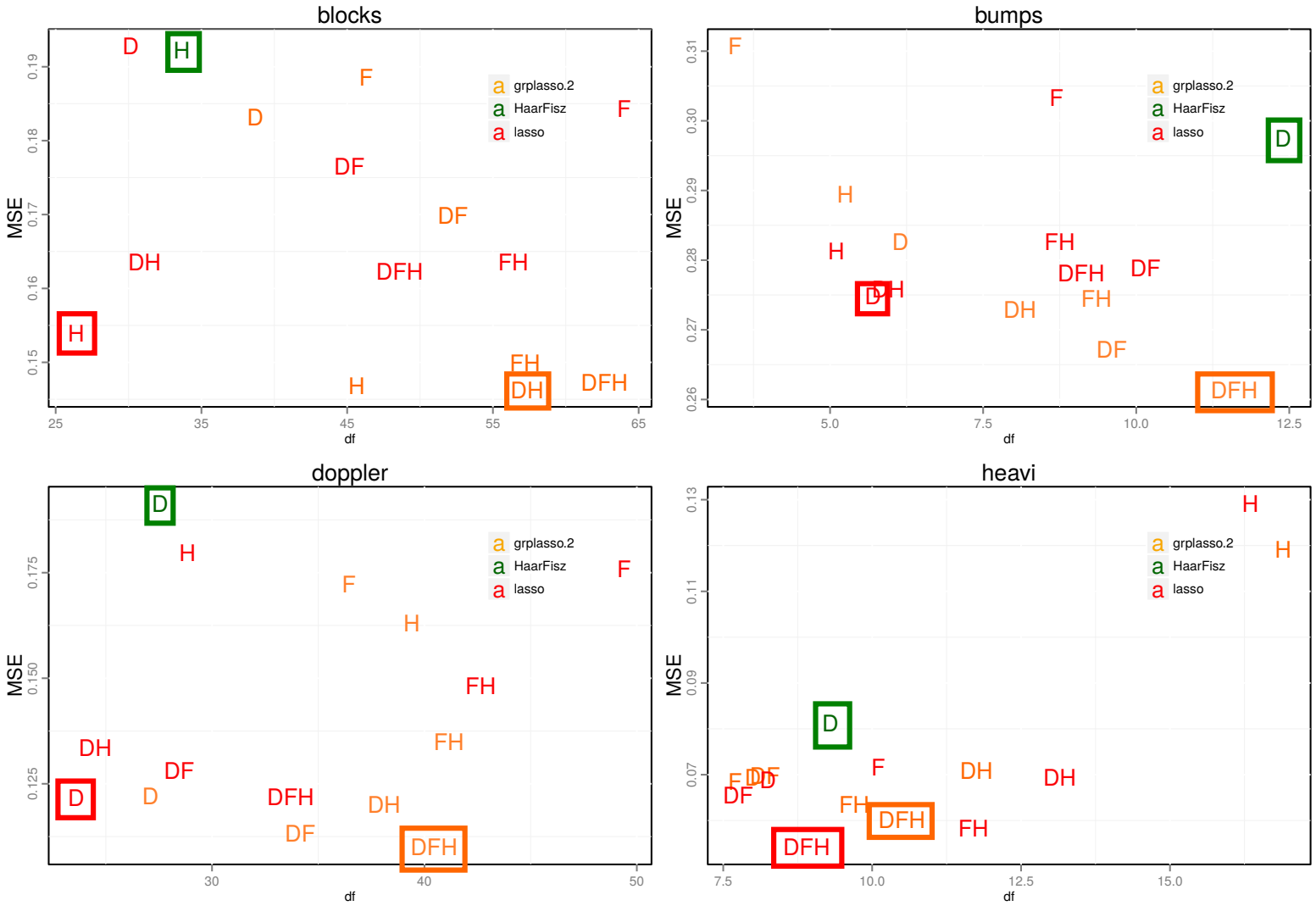
Estimated intensity functions (Lasso)



Estimated intensity functions (group-Lasso)



Choosing the best dictionary by cross-validation



Take-home message

- Methodologies proposed to linear regression models can be adapted to more intricate models, such as Generalized Linear Models or related models.
- Alternatively to the ℓ_2-criterion, we can use the likelihood-based criterion. In the last case, most of the times, the opposite of the log-likelihood, when convex, is penalized.
- For most of GLM, the noise level is not constant. Heteroscedasticity has to be taken into account to calibrate Lasso-type procedures.
- The last point is even more crucial for inverse problems.

References

- Duarte, M. F., Davenport, M. A., Takhar, D., Laska, J. N., Sun, T., Kelly, K. F. and Baraniuk, R. G. Single Pixel Imaging via Compressive Sampling. IEEE Sig. Proc. Mag., 25(2), 83–91, 2008

- Ivanoff S., Picard F. and Rivoirard V. Adaptive Lasso and group-Lasso for functional Poisson regression. *Journal of Machine Learning Research*, 17(55), 1–46, 2016

- Jiang X., Reynaud-Bouret P., Rivoirard V., Sansonnet L. and Willet R. A data-dependent weighted LASSO under Poisson noise. Submitted, 2017

- McCullagh, P.; Nelder, J. A. Generalized linear models. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1989

- Meier, L., van de Geer, S. and Bhlmann, P. The group Lasso for logistic regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 70, no. 1, 53–71, 2008

- Tibshirani, R. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* 58, no. 1, 267–288, 1996

- Tibshirani, R. The lasso method for variable selection in the Cox model. *Statistics in Medicine* 16, no. 4, 385–395, 1997