Introduction to high-dimensional statistics

Vincent Rivoirard

Université Paris Dauphine - PSL

12 janvier 2023

High-dimensional statistics

- In modern science, we collect more observations but we also record radically larger numbers of variables on individuals. Such data are said to be high-dimensional.

High-dimensional statistics

- In modern science, we collect more observations but we also record radically larger numbers of variables on individuals. Such data are said to be high-dimensional.
- Therefore, for such data, the dimension of the data vectors, denoted p, can be comparable to or even much larger than the sample size n. We have $p \geq n$ We speak about the "p larger than n setting“.

High-dimensional statistics

- In modern science, we collect more observations but we also record radically larger numbers of variables on individuals. Such data are said to be high-dimensional.
- Therefore, for such data, the dimension of the data vectors, denoted p, can be comparable to or even much larger than the sample size n. We have $p \geq n$ We speak about the "p larger than n setting“.
- Being able to collect a large amount of information on each individual seems to be good news.

High-dimensional statistics

- In modern science, we collect more observations but we also record radically larger numbers of variables on individuals. Such data are said to be high-dimensional.
- Therefore, for such data, the dimension of the data vectors, denoted p, can be comparable to or even much larger than the sample size n. We have $p \geq n$ We speak about the "p larger than n setting".
- Being able to collect a large amount of information on each individual seems to be good news. Unfortunately:

High-dimensional statistics

- In modern science, we collect more observations but we also record radically larger numbers of variables on individuals. Such data are said to be high-dimensional.
- Therefore, for such data, the dimension of the data vectors, denoted p, can be comparable to or even much larger than the sample size n. We have $p \geq n$ We speak about the "p larger than n setting“.
- Being able to collect a large amount of information on each individual seems to be good news. Unfortunately:

  1. Separating the signal from the noise is a very hard task for high-dimensional data, in full generality impossible.

High-dimensional statistics

- In modern science, we collect more observations but we also record radically larger numbers of variables on individuals. Such data are said to be high-dimensional.
- Therefore, for such data, the dimension of the data vectors, denoted p, can be comparable to or even much larger than the sample size n. We have $p \geq n$ We speak about the "p larger than n setting“.
- Being able to collect a large amount of information on each individual seems to be good news. Unfortunately:
  1. Separating the signal from the noise is a very hard task for high-dimensional data, in full generality impossible.
  2. Extracting the "good information" is more than challenging, consisting in finding a needle in a haystack.

High-dimensional statistics

- In modern science, we collect more observations but we also record radically larger numbers of variables on individuals. Such data are said to be high-dimensional.
- Therefore, for such data, the dimension of the data vectors, denoted p, can be comparable to or even much larger than the sample size n. We have $p \geq n$ We speak about the "p larger than n setting“.
- Being able to collect a large amount of information on each individual seems to be good news. Unfortunately:
    1. Separating the signal from the noise is a very hard task for high-dimensional data, in full generality impossible.
    2. Extracting the "good information" is more than challenging, consisting in finding a needle in a haystack.
    3. Using traditional techniques, often based on asymptotic arguments with the dimension p held fixed as the sample size n increases, is not possible.

High-dimensional statistics

- In modern science, we collect more observations but we also record radically larger numbers of variables on individuals. Such data are said to be high-dimensional.
- Therefore, for such data, the dimension of the data vectors, denoted p, can be comparable to or even much larger than the sample size n. We have $p \geq n$ We speak about the ”p larger than n setting“.
- Being able to collect a large amount of information on each individual seems to be good news. Unfortunately:
    1. Separating the signal from the noise is a very hard task for high-dimensional data, in full generality impossible.
    2. Extracting the ”good information” is more than challenging, consisting in finding a needle in a haystack.
    3. Using traditional techniques, often based on asymptotic arguments with the dimension p held fixed as the sample size n increases, is not possible.
- This phenomenon is often called the curse of dimensionality, terminology introduced by Richard Bellman, in 1961.

Plan

1. Curse of dimensionality
2. Linear regression setting
3. Classical estimation
4. Ridge estimation: ℓ_2-penalization
5. Model selection (à la Birgé-Massart): ℓ_0-penalization
6. Lasso estimation: ℓ_1-penalization
7. 9 variations of the Lasso

Plan

1. Curse of dimensionality
2. Linear regression setting
3. Classical estimation
4. Ridge estimation: ℓ_2-penalization
5. Model selection (à la Birgé-Massart): ℓ_0-penalization
6. Lasso estimation: ℓ_1-penalization
7. 9 variations of the Lasso

Curse of dimensionality. The Gaussian distribution

The p-dimensional standard Gaussian density is:

$$f(x) = \frac{1}{\left(\sqrt{2\pi}\right)^p} \exp\left(-\frac{\|x\|^2}{2}\right)$$

Curse of dimensionality. The Gaussian distribution

The *p*-dimensional standard Gaussian density is:

$$f(x) = \frac{1}{\left(\sqrt{2\pi}\right)^p} \exp\left(-\frac{\|x\|^2}{2}\right)$$

- The *p*-dimensional standard Gaussian density is very flat:

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

Curse of dimensionality. The Gaussian distribution

The p-dimensional standard Gaussian density is:

$$f(x) = \frac{1}{\left(\sqrt{2\pi}\right)^p} \exp\left(-\frac{\|x\|^2}{2}\right)$$

- The p-dimensional standard Gaussian density is very flat:

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

- For p large, the mass of the Gaussian distribution concentrates in its tails

Proposition

Let $X \sim \mathcal{N}(0, I_p)$. For any $K > 0$,

$$\mathbb{P}\left(\|X\| \leq K\right) \leq \mathbb{E}\left[e^{-\|X\|^2/2}\right] e^{K^2/2} = 2^{-p/2} e^{K^2/2}.$$

Consequence: $\mathbb{P}\left(\|X\| \leq K\right)$ non-negligible requires $K \gtrsim \sqrt{p}$

Curse of dimensionality - Neighborhood

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

$$\mathbb{E}[Y|X] = m(X)$$

Curse of dimensionality - Neighborhood

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.

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

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 - Neighborhood

- 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 - Neighborhood

- 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:

Curse of dimensionality - Neighborhood

- 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

Curse of dimensionality - Neighborhood

- 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.

Curse of dimensionality - Neighborhood

- 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.

  Theoretical arguments: Consider $X = (X_1, \dots, X_p)$ a random vector whose coordinates are i.i.d. and bounded by b a.s.
  - Hoeffding's inequality implies that for any $K > 0$, with $m^2 = \mathbb{E}[X_1^2]$,

$$\mathbb{P}\Big(\|X\| \notin \Big[m\sqrt{p} - b^2 m^{-1} K ; m\sqrt{p} + b^2 m^{-1} K\Big]\Big) \leq 2\exp(-K^2/2).$$

Curse of dimensionality - Neighborhood

- 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.

Theoretical arguments: Consider $X = (X_1, \ldots, X_p)$ a random vector whose coordinates are i.i.d. and bounded by b a.s.

- Hoeffding's inequality implies that for any $K > 0$, with $m^2 = \mathbb{E}[X_1^2]$,

$$\mathbb{P}\Big(\|X\| \notin \Big[m\sqrt{p} - b^2 m^{-1} K; m\sqrt{p} + b^2 m^{-1} K\Big]\Big) \leq 2\exp(-K^2/2).$$

- This can be generalized for any ℓ_q-norm on $\mathbb{R}^p$, $1 \leq q < \infty$.

Curse of dimensionality - Neighborhood

- 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.

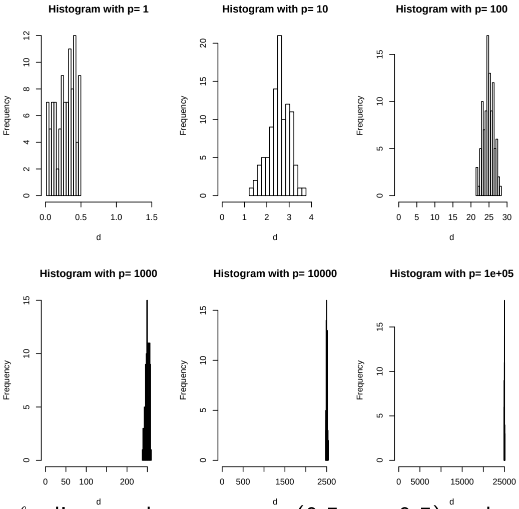
Theoretical arguments: Consider $X = (X_1, \dots, X_p)$ a random vector whose coordinates are i.i.d. and bounded by b a.s.
- Hoeffding's inequality implies that for any $K > 0$, with $m^2 = \mathbb{E}[X_1^2]$,

$$\mathbb{P}\Big(\|X\| \notin \Big[m\sqrt{p} - b^2 m^{-1} K; m\sqrt{p} + b^2 m^{-1} K\Big]\Big) \leq 2\exp(-K^2/2).$$

- This can be generalized for any ℓ_q-norm on $\mathbb{R}^p$, $1 \leq q < \infty$.
- It can also be generalized for the sup-norm: Let $K > 0$. With $p_k = \mathbb{P}(|X_1| > K)$,

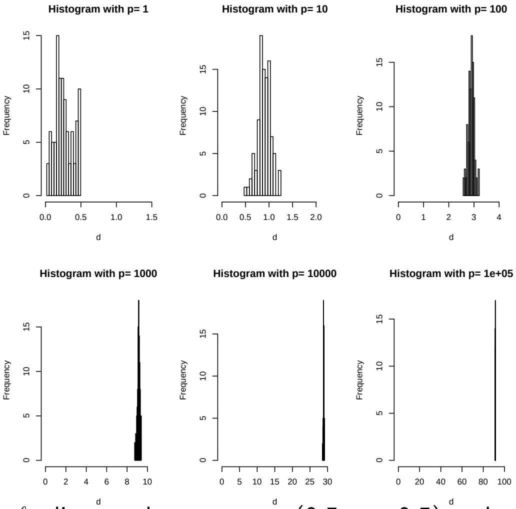
$$\mathbb{P}(\|X\|_\infty \leq K) = \exp(p\log(1 - p_k)).$$

Curse of dimensionality - Neighborhood



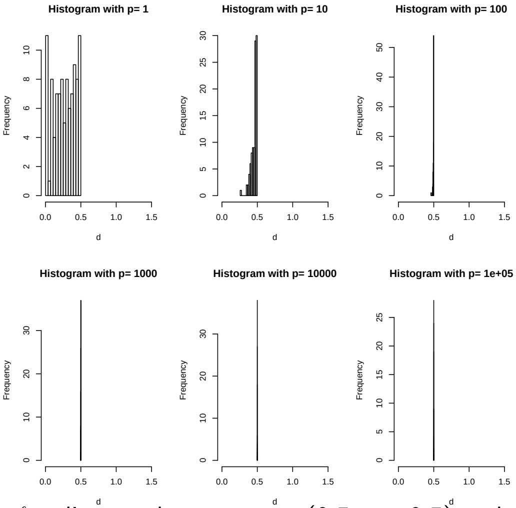
Histograms of the ℓ_1-distance between $x = (0.5, \dots, 0.5)$ and $n = 100$ random uniform variables on $[0, 1]^p$.

Curse of dimensionality - Neighborhood



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

Curse of dimensionality - Neighborhood



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

Curse of dimensionality

Main problem:

- The volume $V_p(r)$ of a p-dimensional ball of radius r for the ℓ_2-norm satisfies

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

Curse of dimensionality

Main problem:

- The volume $V_p(r)$ of a p-dimensional ball of radius r for the ℓ_2-norm satisfies

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

Other problems:

- The diagonal of the hypercube $[0,1]^p$ is almost orthogonal to its edges
- Accumulation of small fluctuations in many directions can produce a large global fluctuation.
- Computational complexity.

Circumventing the curse of dimensionality

In light of the few previous arguments, the situation seems desperate.

- 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.

References

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

- Donoho, David *High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality*, AMS Lectures, 2000.

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

- Hastie, Trevor; Tibshirani, Robert & Friedman, Jérome *The elements of statistical learning. Data mining, inference, and prediction. Second edition.* Springer Series in Statistics. Springer, New York, 2009.

- Wainwright, Martin J. *High-dimensional statistics: A non-asymptotic viewpoint.* Cambridge Series in Statistical and Probabilistic Mathematics, vol. 48, Cambridge University Press, Cambridge, 2019.

Plan

1. Curse of dimensionality
2. Linear regression setting
3. Classical estimation
4. Ridge estimation: ℓ_2-penalization
5. Model selection (à la Birgé-Massart): ℓ_0-penalization
6. Lasso estimation: ℓ_1-penalization
7. 9 variations of the Lasso

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

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

Sometimes, we further assume that

- σ^2 is known
- ϵ is Gaussian

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

Plan

1. Curse of dimensionality
2. Linear regression setting
3. Classical estimation
4. Ridge estimation: ℓ_2-penalization
5. Model selection (à la Birgé-Massart): ℓ_0-penalization
6. Lasso estimation: ℓ_1-penalization
7. 9 variations of the Lasso

Classical estimation

We consider the linear regression model

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

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

- We naturally estimate β^* by considering the ordinary least squares estimate $\widehat{\beta}^{ols}$ defined by
$$\widehat{\beta}^{ols} \in \arg\min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|.$$
- The estimator $\widehat{\beta}^{ols}$ is uniquely defined if and only if X^TX is invertible In this case, we have
$$\widehat{\beta}^{ols} = (X^TX)^{-1}X^TY.$$
- If X^TX is not invertible, introducing $(X^TX)^+$ the Moore-Penrose inverse of X^TX, we have:

$$\arg\min_{\beta \in \mathbb{R}^p} \|Y - X\beta\| = \left\{(X^TX)^+X^TY + u : \ u \in \ker(X^TX)\right\}$$

Classical estimation

We consider the linear regression model

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

with $\mathbb{E}[\epsilon] = 0$, $\mathrm{Var}(\epsilon) = \sigma^2 I_n$ and $X^T X$ invertible. The ordinary least squares estimate is

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

- We have

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

  and

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

- If the predictors are orthonormal

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

- Remark: $X^T X$ invertible ($\iff \ker(X) = \{0\}) \iff \mathrm{rank}(X) = p \Rightarrow p \leq n$ ☹

Plan

1. Curse of dimensionality
2. Linear regression setting
3. Classical estimation
4. Ridge estimation: ℓ_2-penalization
5. Model selection (à la Birgé-Massart): ℓ_0-penalization
6. Lasso estimation: ℓ_1-penalization
7. 9 variations of the Lasso

Ridge estimates

We still consider the linear regression model

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

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

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

which satisfies, with $\|\cdot\|$ the ℓ_2-norm,

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

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

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

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

Following Hoerl and Kennard (1970), this estimator is called the Ridge estimate.

Ridge estimates : $\widehat{\beta}_\lambda^{ridge} := (X^TX + \lambda I_p)^{-1}X^TY$

- If ϵ is Gaussian, for any $\lambda \geq 0$

$$\mathbb{P}\left(\forall j \quad \widehat{\beta}_{\lambda,j}^{ridge} \neq 0\right) = 1$$

- The tuning parameter λ balances the bias and variance terms.

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

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

with $(\mu_j)_{j=1,\ldots,p} :=$ eigenvalues(X^TX). We deduce the risk:

$$\mathbb{E}\left[\|\widehat{\beta}_\lambda^{ridge} - \beta^*\|^2\right] = \lambda^2 \beta^{*T}(X^TX + \lambda I_p)^{-2}\beta^* + \sigma^2 \sum_{j=1}^{p} \frac{\mu_j}{(\mu_j + \lambda)^2}$$

Pros and cons:

1. We don't need the assumption $\ker(X) = \{0\}$ ☺
2. We can consider high dimensions: $p > n$ ☺
3. Linearity: Easy to compute for most problems ☺
4. Automatic selection is not possible ☹
5. The choice of the regularization parameter λ is intricate ☹

Plan

1. Curse of dimensionality
2. Linear regression setting
3. Classical estimation
4. Ridge estimation: ℓ_2-penalization
5. Model selection (à la Birgé-Massart): ℓ_0-penalization
6. Lasso estimation: ℓ_1-penalization
7. 9 variations of the Lasso

Sparsity

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

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,

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

with $\mathbb{E}[\epsilon] = 0$, $\mathrm{Var}(\epsilon) = \sigma^2 I_n$ and known σ^2, we assume that m^* the support of β^* is small, with

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

Note that m^* is unknown.

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,

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

with $\mathbb{E}[\epsilon] = 0$, $\mathrm{Var}(\epsilon) = \sigma^2 I_n$ and known σ^2, we assume that m^* the support of β^* is small, with

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

Note that m^* is unknown.

- In general, $\widehat{\beta}^{ols}$ and $\forall\ \lambda > 0$, $\widehat{\beta}_\lambda^{ridge}$ are not sparse.

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,

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

  with $\mathbb{E}[\epsilon] = 0$, $\mathrm{Var}(\epsilon) = \sigma^2 I_n$ and known σ^2, 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, $\widehat{\beta}^{ols}$ and $\forall\ \lambda > 0$, $\widehat{\beta}_\lambda^{ridge}$ are 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:
$$Y = f^* + \epsilon. \tag{5.1}$$

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:
$$Y = f^* + \epsilon. \tag{5.1}$$

- 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
$$S^* = \mathrm{span}\{X_j : \ j \in m^*\}.$$

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:
$$Y = f^* + \epsilon. \tag{5.1}$$

- 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 (5.1) under the constraint that the estimate of f^* belongs to S^*.

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:

$$Y = f^* + \epsilon. \tag{5.1}$$

- 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

$$S^* = \mathrm{span}\{X_j : \ j \in m^*\}.$$

Note that if $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ then $\hat{f}_{m^*}$ is the maximum likelihood estimate in the model (5.1) 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 $m \subset \{1, \dots, p\}$, 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.

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.

Oracle approach

- For any model $m \subset \{1, \dots, p\}$, 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 "oracle estimate'.

Oracle approach

- For any model $m \subset \{1, \dots, p\}$, 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 "oracle estimate'.
- More precisely, we wish to select $\hat{m} \in \mathcal{M}$ such that
$$\mathbb{E}\left[\|\hat{f}_{\hat{m}} - f^*\|^2\right] \lesssim \mathbb{E}\left[\|\hat{f}_{m_0} - f^*\|^2\right].$$

Oracle approach

Oracle model:

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

Some remarks:

- We allow $m^* \notin \mathcal{M}$.
- Even if $m^* \in \mathcal{M}$, m^* may be different from m_0.
- The oracle model m_0 is not random but depends on β^*. So, $\hat{f}_{m_0}$ 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 satisfies

$$m_0 := \arg\min_{m\in\mathcal{M}} R(\hat{f}_m)$$

and is the model which achieves the best trade-off between these two terms.

Mallows' C_p

Mallows' Recipe: Since we wish to minimize

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

it's natural to choose $\hat{m}$ as the minimizer of an estimate of $R(\hat{f}_m)$. The following lemma gives the main ingredient of the recipe (based on the replacement of f^* by Y).

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 (1973))

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

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

- Assumptions are mild. In particular the Mallows' criterion is distribution-free. ☺
- It achieves good performances in many situations, so is a very popular criterion. ☺

Mallows' C_p

Definition (Mallows (1973))

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

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

- Assumptions are mild. In particular the Mallows' criterion is distribution-free. ☺
- It achieves good performances in many situations, so is a very popular criterion. ☺
- Only based on unbiased estimation, this approach does not take into account fluctuations of $C_p(m)$ around its expectation $\mathbb{E}[C_p(m)] = R(\hat{f}_m) + \sigma^2 n$.

Mallows' C_p

Definition (Mallows (1973))

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

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

- Assumptions are mild. In particular the Mallows' criterion is distribution-free. ☺
- It achieves good performances in many situations, so is a very popular criterion. ☺
- Only based on unbiased estimation, this approach does not take into account fluctuations of $C_p(m)$ around its expectation $\mathbb{E}[C_p(m)] = R(\hat{f}_m) + \sigma^2 n$.
  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

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

Mallows' C_p

Definition (Mallows (1973))

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

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

- Assumptions are mild. In particular the Mallows' criterion is distribution-free. ☺
- It achieves good performances in many situations, so is a very popular criterion. ☺
- Only based on unbiased estimation, this approach does not take into account fluctuations of $C_p(m)$ around its expectation $\mathbb{E}[C_p(m)] = R(\hat{f}_m) + \sigma^2 n$.
  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

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

  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: AIC and BIC

- We assume that the distribution of observations is known. ☹

Other popular criteria: AIC and BIC

- We assume that the distribution of observations is known. ☹
- In this case, we can consider AIC and BIC criteria which are based on the likelihood.

Other popular criteria: AIC and BIC

- We assume that the distribution of observations is known. ☹
- In this case, 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) (Akaike (1973))

$$C(m) = -2L(m) + 2d_m$$

Other popular criteria: AIC and BIC

- We assume that the distribution of observations is known. ☹
- In this case, 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) (Akaike (1973))

$$C(m) = -2L(m) + 2d_m$$

  - for the Bayesian Information Criterion (BIC) (Schwarz (1978))

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

Other popular criteria: AIC and BIC

- We assume that the distribution of observations is known. ☹
- In this case, 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) (Akaike (1973))

    $$C(m) = -2L(m) + 2d_m$$

  - for the Bayesian Information Criterion (BIC) (Schwarz (1978))

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

- In the Gaussian setting with σ^2 known, AIC and Mallows' C_p are equivalent.

Other popular criteria: AIC and BIC

- We assume that the distribution of observations is known. ☹
- In this case, 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) (Akaike (1973))

$$C(m) = -2L(m) + 2d_m$$

  - for the Bayesian Information Criterion (BIC) (Schwarz (1978))

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

- In the Gaussian setting with σ^2 known, AIC and Mallows' C_p are equivalent.
- The use of BIC tends to prevent overfitting (larger penalty).

Theoretical analysis of AIC and BIC

- We illustrate drawbacks of AIC and BIC for large collections of models
- We take $n = p$

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

  with $X^T X = I_p$, $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ and σ^2 known. We take

  $$\beta^* = 0$$

  So that $m^* = m_0 = \emptyset$. Let $\mathcal{M} = \mathcal{P}(\{1, \dots, p\})$.

Theoretical analysis of AIC and BIC

- We illustrate drawbacks of AIC and BIC for large collections of models
- We take $n = p$

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

with $X^T X = I_p$, $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ and σ^2 known. We take

$$\beta^* = 0$$

So that $m^* = m_0 = \emptyset$. Let $\mathcal{M} = \mathcal{P}(\{1, \dots, p\})$.

- AIC (= Mallows):

$$\hat{m}_{\text{AIC}} = \arg\min_{m \in \mathcal{M}} \left\{ \| Y - \hat{f}_m \|^2 + 2\sigma^2 d_m \right\}$$

$$\mathbb{E}\left[\text{card}(\hat{m}_{\text{AIC}})\right] \overset{p \to +\infty}{\sim} 0.16p$$

Theoretical analysis of AIC and BIC

- We illustrate drawbacks of AIC and BIC for large collections of models
- We take $n = p$

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

with $X^T X = I_p$, $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ and σ^2 known. We take

$$\beta^* = 0$$

So that $m^* = m_0 = \emptyset$. Let $\mathcal{M} = \mathcal{P}(\{1, \dots, p\})$.

- AIC (= Mallows):

$$\hat{m}_{\text{AIC}} = \arg\min_{m \in \mathcal{M}} \left\{ \|Y - \hat{f}_m\|^2 + 2\sigma^2 d_m \right\}$$

$$\mathbb{E}\left[\text{card}(\hat{m}_{\text{AIC}})\right] \overset{p \to +\infty}{\sim} 0.16p$$

- BIC:

$$\hat{m}_{\text{BIC}} = \arg\min_{m \in \mathcal{M}} \left\{ \|Y - \hat{f}_m\|^2 + \log(n)\sigma^2 d_m \right\}$$

$$\mathbb{E}[\text{card}(\hat{m}_{\text{BIC}})] \overset{p \to +\infty}{\sim} \sqrt{\frac{2p}{\pi \log(p)}}$$

Penalization for Gaussian regression

- We still consider
$$Y = f^* + \epsilon,$$
with $f^* = X\beta^*$, $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ and σ^2 known.
- Given $\mathcal{M}$, previous approaches show that for ℓ_2 estimation, estimates
$$\hat{f} := \hat{f}_{\hat{m}} := \arg\min_{m\in\mathcal{M}} C(m),$$
$$C(m) = \|Y - \hat{f}_m\|^2 + \sigma^2 \mathrm{pen}(m), \quad \hat{f}_m = \Pi_{S_m} Y$$
may be suitable when pen, called the penalty, is proportional to d_m, the dimension of 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 $R(\hat{f}) \lesssim R(\hat{f}_{m_0})$ (equivalently $R(\hat{f}) \lesssim R(\hat{f}_m)$ for any $m \in \mathcal{M}$)
- We have
$$R(\hat{f}_m) = \|(I_n - \Pi_{S_m})f^*\|^2 + \sigma^2 d_m.$$

Penalty

Let $m \in \mathcal{M}$ be fixed.

Penalty

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

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

Penalty

Let $m \in \mathcal{M}$ be fixed. 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 $K > 1$, with $\bar{S}_{\hat{m}} = \mathrm{span}(S_{\hat{m}}, f^*)$,

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

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

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

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

$$\mathrm{pen}(m) := K\left(\sqrt{d_m} + \sqrt{-2\log(\pi_m)}\right)^2. \tag{5.2}$$

If $K > 1$, concentration inequalities lead to

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

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

Theoretical result

Theorem (Birgé and Massart (2001))

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* (5.2) *and estimate* f^* *with* $\hat{f} = \hat{f}_{\hat{m}}$ *such that*

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

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

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

- The assumption $K > 1$ can't be relaxed.

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. ☺

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. ☺

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 $\mathcal{M} = \mathcal{P}(\{1, \dots, p\})$, $\mathrm{card}(\mathcal{M}) = 2^p$. 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: $X^TX = I_p$. We have $d_m := \dim(S_m) = \mathrm{card}(m)$. Consider a penalty proportional to d_m:

$$\mathrm{pen}(m) = 2cd_m \log(p).$$

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:

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

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

- In this case, we have:

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

and

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

Model selection corresponds to hard thresholding and implementation is easy.

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

- In this case, we have:

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

and

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

Model selection corresponds to hard thresholding and implementation is easy.

- Assume that $f^* = 0$ and $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. Mallows' C_p and BIC are overfitting procedures. When $p \to +\infty$, if $c > 1$,

$$\mathbb{P}(\hat{f}_{HT,c} \neq 0) = o(1)$$

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

- Akaike, H. Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory*, 267–281, 1973.
- Birgé, Lucien and Massart, Pascal Gaussian model selection. *J. Eur. Math. Soc. (JEMS)* 3, no. 3, 203–268, 2001.
- Hoerl, A. E.; R. W. Kennard Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12 (1): 55–67, 1970.
- Kerkyacharian, Gérard and Picard, Dominique Thresholding algorithms, maxisets and well-concentrated bases. *Test* 9, no. 2, 283–344, 2000.
- Lebarbier, Emilie & Mary-Huard, Tristan Une introduction au critère BIC: fondements théoriques et interprétation. *J. Soc. Fr. Stat.* 147, no. 1, 39–57, 2006.
- Mallows, Colin Some Comments on C_p. *Technometrics*, 15, 661–675, 1973.
- Massart, Pascal *Concentration inequalities and model selection*, volume 6, Springer, 2007.
- Schwarz, Gideon E. Estimating the dimension of a model, *Annals of Statistics*, 6, no. 2, 461–464, 1978.

Plan

1. Curse of dimensionality
2. Linear regression setting
3. Classical estimation
4. Ridge estimation: ℓ_2-penalization
5. Model selection (à la Birgé-Massart): ℓ_0-penalization
6. Lasso estimation: ℓ_1-penalization
7. 9 variations of the Lasso

Convexification

- We still consider the linear regression model and the estimation problem (i.e. estimation of β^*)
$$Y = X\beta^* + \epsilon,$$
with $\mathbb{E}[\epsilon] = 0$, $\mathrm{Var}(\epsilon) = \sigma^2 I_n$ and σ^2 known.

Convexification

- We still consider the linear regression model and the estimation problem (i.e. estimation of β^*)
$$Y = X\beta^* + \epsilon,$$
with $\mathbb{E}[\epsilon] = 0$, $\mathrm{Var}(\epsilon) = \sigma^2 I_n$ and σ^2 known.
- Given $\mathcal{M} \subset \mathcal{P}(\{1, \ldots, p\})$, a collection of models and setting for $m \in \mathcal{M}$, and setting for any $m \in \mathcal{M}$,
$$S_m = \mathrm{span}\{X_j : \ j \in m\}, \quad \hat{f}_m = \Pi_{S_m} Y,$$
we have studied
$$\begin{aligned} \hat{m} := \arg\min_{m\in\mathcal{M}} & \left\{ \|Y - \hat{f}_m\|^2 + \lambda(\sigma^2) d_m \right\}. \\ = \arg\min_{m\in\mathcal{M}} & \left\{ \min_{\beta\in\mathbb{R}^p : X\beta \in S_m} \|Y - X\beta\|^2 + \lambda(\sigma^2) d_m \right\} \\ = \arg\min_{m\in\mathcal{M}} & \left\{ \min_{\beta\in\mathbb{R}^p : X\beta \in S_m} \left\{ \|Y - X\beta\|^2 + \lambda(\sigma^2) \|\beta\|_{\ell_0} \right\} \right\} \end{aligned}$$
under the condition $d_m = \mathrm{card}(m)$ and setting
$$\|\beta\|_{\ell_0} = \sum_{j=1}^{p} 1_{\{\beta_j \neq 0\}}, \quad \beta \in \mathbb{R}^p.$$

Convexification

- We still consider the linear regression model and the estimation problem (i.e. estimation of β^*)
$$Y = X\beta^* + \epsilon,$$
with $\mathbb{E}[\epsilon] = 0$, $\mathrm{Var}(\epsilon) = \sigma^2 I_n$ and σ^2 known.
- Given $\mathcal{M} \subset \mathcal{P}(\{1,\dots,p\})$, a collection of models and setting for $m \in \mathcal{M}$, and setting for any $m \in \mathcal{M}$,
$$S_m = \mathrm{span}\{X_j : \ j \in m\}, \quad \hat{f}_m = \Pi_{S_m} Y,$$
we have studied
$$\hat{m} = \arg\min_{m\in\mathcal{M}} \left\{ \min_{\beta\in\mathbb{R}^p : X\beta\in S_m} \left\{ \|Y - X\beta\|^2 + \lambda(\sigma^2)\|\beta\|_{\ell_0} \right\} \right\}$$
under the condition $d_m = \mathrm{card}(m)$ and setting
$$\|\beta\|_{\ell_0} = \sum_{j=1}^{p} 1_{\{\beta_j \neq 0\}}, \quad \beta \in \mathbb{R}^p.$$
- We obtain the model selection estimate:
$$\widehat{\beta} = \arg\min_{\beta\in\mathbb{R}^p} \left\{ \|Y - X\beta\|^2 + \lambda(\sigma^2)\|\beta\|_{\ell_0} \right\}$$

Convexification

- Model selection estimate:

$$\widehat{\beta} = \arg\min_{\beta\in\mathbb{R}^p} \left\{ \|Y - X\beta\|^2 + \lambda(\sigma^2)\|\beta\|_{\ell_0} \right\}$$

But, because of the penalty term,

$$C(\beta) := \|Y - X\beta\|^2 + \lambda(\sigma^2)\|\beta\|_{\ell_0}$$

is not convex.

Convexification

- Model selection estimate:

$$\widehat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|^2 + \lambda(\sigma^2)\|\beta\|_{\ell_0} \right\}$$

But, because of the penalty term,

$$C(\beta) := \|Y - X\beta\|^2 + \lambda(\sigma^2)\|\beta\|_{\ell_0}$$

is not convex.

- We replace the penalty using a convexification of the ℓ_0-norm. Typically, we take a penalty proportional to $\|\beta\|_{\ell_\gamma}^{\gamma}$ for $\gamma \geq 1$:

$$\widehat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|^2 + \lambda\|\beta\|_{\ell_\gamma}^{\gamma} \right\}$$

Bridge estimates

Franck and Friedman (1993) introduced 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

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

and

$$\widehat{\beta}_{\lambda,\gamma} := \arg\min_{\beta \in \mathbb{R}^p} C_{\lambda,\gamma}(\beta). \tag{6.1}$$

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

1. $\gamma = 0$: Model Selection
2. $\gamma = 2$: Ridge Estimation
3. $\gamma = 1$: Lasso Estimation

The case $\lambda = 0$ corresponds to the Ordinary Least Squares estimate.

Bridge estimates

We use for $\beta \in \mathbb{R}^p$

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

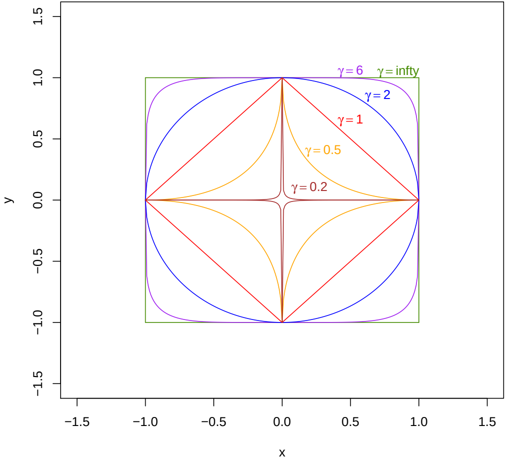
- If $0 \leq \gamma < 1$, then $C_{\lambda,\gamma}$ is not convex and it may be very hard to minimize it in high dimensions.
- If $\gamma = 1$, then $C_{\lambda,1}$ is convex and $C_{\lambda,1}$ has one minimizer if $\text{rank}(X) = p$.
- If $\gamma > 1$, then $C_{\lambda,\gamma}$ is strictly convex and $C_{\lambda,\gamma}$ has only one minimizer. If ϵ is Gaussian, almost surely, all coordinates of the minimizer of $C_{\lambda,\gamma}$ are non-zero.
- For $\gamma \geq 1$, one-to-one correspondence between the Lagragian problem

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

and the following constrained problem

$$\arg\min_{\{\beta\in\mathbb{R}^p:\ \|\beta\|_\gamma^\gamma \leq t\}} \|Y - X\beta\|^2.$$

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

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

and the constrained problem becomes

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

- We compare the Ridge estimate

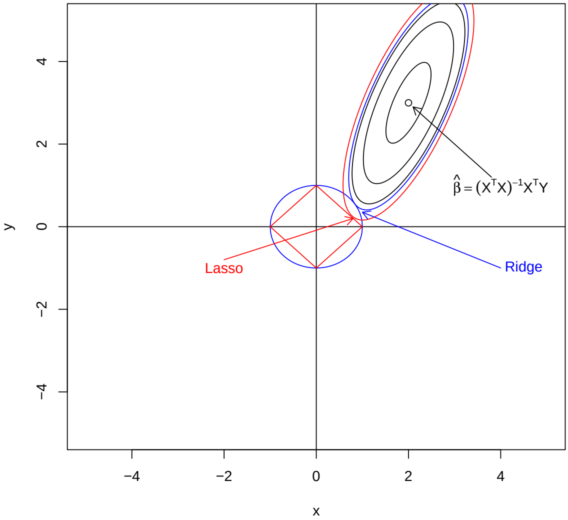
$$\widehat{\beta}_\lambda^{ridge} := \arg \min_{\{\beta \in \mathbb{R}^p: \ \|\beta\|^2 \leq t\}} \left\{(\beta - \widehat{\beta}^{ols})^T X^T X(\beta - \widehat{\beta}^{ols})\right\}$$

and the Lasso estimate

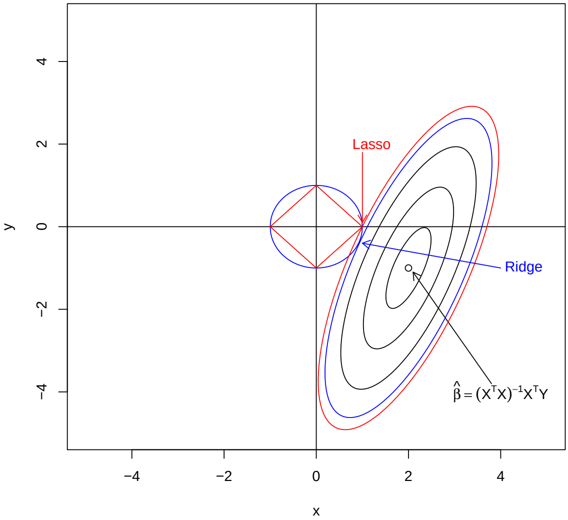
$$\widehat{\beta}_\lambda^{lasso} := \arg \min_{\{\beta \in \mathbb{R}^p: \ \|\beta\|_1 \leq t\}} \left\{(\beta - \widehat{\beta}^{ols})^T X^T X(\beta - \widehat{\beta}^{ols})\right\}$$

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

Graphical illustration for $p = 2$



Graphical illustration for $p = 2$



Characterization of the Lasso

The Lasso, proposed by Tibshirani (1996), is the minimizer of $C_{\lambda,\gamma}$ with $\gamma = 1$:

$$\widehat{\beta}_\lambda^{lasso} \in \arg\min_{\beta\in\mathbb{R}^p} \Big\{ \underbrace{\|Y - X\beta\|^2 + \lambda\|\beta\|_1}_{C_{\lambda,1}} \Big\}$$

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

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

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

$$\mathcal{E}_\lambda := \Big\{ j : \ |2X_j^T(Y - X\widehat{\beta}_\lambda)| = \lambda \Big\}$$

Uniqueness of the Lasso

Lasso estimate:

$$\widehat{\beta}_\lambda^{lasso} \in \arg\min_{\beta\in\mathbb{R}^p} \Big\{ \underbrace{\|Y - X\beta\|^2 + \lambda\|\beta\|_1}_{C_{\lambda,1}} \Big\}$$

Let $\widehat{\beta}_\lambda$ a global minimizer of $C_{\lambda,1}$. If $\widehat{\beta}'_\lambda$ is another global minimizer of $C_{\lambda,1}$, then

$$X\widehat{\beta}_\lambda = X\widehat{\beta}'_\lambda \quad \text{and} \quad \|\widehat{\beta}_\lambda\|_1 = \|\widehat{\beta}'_\lambda\|_1.$$

Conditions for uniqueness (Tibshirani (2013)):

1. $\widehat{\beta}_\lambda$ is the unique minimizer of $C_{\lambda,1}$ if $X_{\mathcal{E}_\lambda}$ is one to one with

$$\mathcal{E}_\lambda := \Big\{ j: \ |2X_j^T(Y - X\widehat{\beta}_\lambda)| = \lambda \Big\}$$

2. Note that $\widehat{S}_\lambda$, the support of $\widehat{\beta}_\lambda$, satisfies

$$\widehat{S}_\lambda := \Big\{ j: \ \widehat{\beta}_{\lambda,j} \neq 0 \Big\} \subset \mathcal{E}_\lambda.$$

So, $\widehat{\beta}_\lambda$ is the unique minimizer of $C_{\lambda,1}$ if $X_{\widehat{S}_\lambda}$ is one to one and $\forall j \notin \widehat{S}_\lambda$,

$$|2X_j^T(Y - X\widehat{\beta}_\lambda)| < \lambda.$$

3. If the entries of X are drawn from a continuous probability distribution on $\mathbb{R}^{np}$, then for any $\lambda > 0$, the lasso solution is unique with probability one.

The orthogonal case

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

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

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

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

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

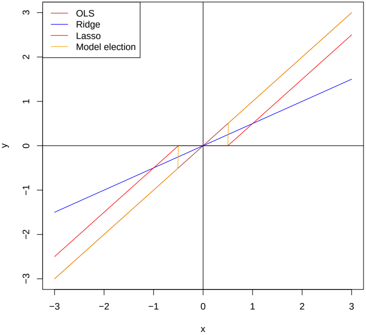
The orthogonal case - Comparison

We assume that the matrix X is orthogonal: $X^TX = I_p$. We set

$$a_j := X_j^T Y$$

and compare

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



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

Theoretical guarantees - Support recovery

We study the problem of estimating the support of β^* with the support of $\widehat{\beta}_\lambda^{lasso}$. Let

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

Under the Irrepresentable Condition (or the incoherence condition), we have $\hat{S}_\lambda = S^*$.

Theorem (Wainwright (2009))

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

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

$$\max_{j \notin S^*} \|(X_{S^*}^T X_{S^*})^{-1} X_{S^*}^T X_j\|_1 \leq 1 - \gamma. \qquad (6.2)$$

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

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

Condition (6.2) is (almost) necessary. See also Zou (2006) and Zhao and Yu (2006).

Theoretical guarantees - Bounds for prediction

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

Let us consider $\lambda \geq 4 \max_{j=1,\dots,p} |(X^T\epsilon)_j|$. *Then,*

$$\|X\widehat{\beta}_\lambda^{lasso} - X\beta^*\|^2 \leq 2\|\beta^*\|_1\lambda.$$

Theoretical guarantees - Bounds for prediction

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

Let us consider $\lambda \geq 4\max_{j=1,\ldots,p} |(X^T\epsilon)_j|$. *Then,*

$$\|X\widehat{\beta}_\lambda^{lasso} - X\beta^*\|^2 \leq 2\|\beta^*\|_1\lambda.$$

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

Let us consider $\lambda \geq 3\max_{j=1,\ldots,p} |(X^T\epsilon)_j|$. *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 $\kappa(\beta) > 0$,

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

Theoretical guarantees - Bounds for prediction

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

Let us consider $\lambda \geq 4 \max_{j=1,\ldots,p} |(X^T\epsilon)_j|$. *Then,*

$$\|X\widehat{\beta}_\lambda^{lasso} - X\beta^*\|^2 \leq 2\|\beta^*\|_1\lambda.$$

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

Let us consider $\lambda \geq 3 \max_{j=1,\ldots,p} |(X^T\epsilon)_j|$. *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 $\kappa(\beta) > 0$,

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

The Restricted Eigenvalues Condition, $\kappa(\beta) > 0$, expresses the lack of orthogonality of columns of X. Milder conditions can be used (see Hunt *et al.* (2019))).

Theoretical guarantees - Bounds for estimation

- From the previous theorem, we can deduce estimation bounds for ℓ_2 and ℓ_1 norms for estimating sparse vectors β^* (see Hunt *et al.* (2019)) :

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

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

Theoretical guarantees - Bounds for estimation

- From the previous theorem, we can deduce estimation bounds for ℓ_2 and ℓ_1 norms for estimating sparse vectors β^* (see Hunt *et al.* (2019)) :

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

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

- Deriving λ such that

$$\lambda \gtrsim \max_{j=1,\dots,p} |(X^T\epsilon)_j|$$

is satisfied with high probability is easy by using concentration inequalities. It provides a theoretical way to tune the Lasso (often too conservative in practice).

Tuning the Lasso - V-fold Cross-validation

1. We write the model
$$Y_i = x_i^T \beta^* + \epsilon_i, \quad i = 1, \dots, n$$
with $x_i \in \mathbb{R}^p$.
2. Choose V (commonly, $V = 5$ or $V = 10$) and a discrete set Λ of values for λ.
3. Split the training set $\{1, \dots, n\}$ into V subsets, $B_1, \dots, B_V$, of roughly the same size.
4. For each value of $\lambda \in \Lambda$, for $k = 1, \dots, V$, compute the estimate $\widehat{\beta}_\lambda^{(-k)}$ 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:
$$e_k(\lambda) := \frac{1}{\mathrm{card}(B_k)} \sum_{i \in B_k} \left(Y_i - x_i^T \widehat{\beta}_\lambda^{(-k)}\right)^2.$$
5. Compute the average error over all folds,
$$CV(\lambda) := \frac{1}{V} \sum_{k=1}^{V} e_k(\lambda) = \frac{1}{V} \sum_{k=1}^{V} \frac{1}{\mathrm{card}(B_k)} \sum_{i \in B_k} \left(Y_i - x_i^T \widehat{\beta}_\lambda^{(-k)}\right)^2.$$
6. We choose the value of tuning parameter that minimizes this function CV on Λ:
$$\widehat{\lambda} := \operatorname*{argmin}_{\lambda \in \Lambda} CV(\lambda).$$

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 rank$(X) = p$: We estimate $X\beta^*$ with

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

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

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 $\widehat{\beta}$ an estimate of β^*. If

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

then we have:

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

- Assume that for any $\lambda > 0$, we have $\widehat{\mathrm{df}}(\lambda)$ a good (unbiased) estimate of $\mathrm{df}(X\widehat{\beta}_\lambda)$, where $\widehat{\beta}_\lambda$ is the Lasso estimate associated with λ. Then, we can choose λ by minimizing
$$\lambda \longmapsto \|Y - X\widehat{\beta}_\lambda\|^2 + 2\sigma^2\widehat{\mathrm{df}}(\lambda)$$
- Denoting $\widehat{S}_\lambda$ the support of $\widehat{\beta}_\lambda^{lasso}$, we have:
$$\mathbb{E}[\mathrm{card}(\widehat{S}_\lambda)] = \mathbb{E}[\mathrm{rank}(X_{\widehat{S}_\lambda})] = \mathrm{df}\,(X\widehat{\beta}_\lambda).$$

See Zou, Hastie and Tibshirani (2007) and Tibshirani and Taylor (2012).

Tuning the Lasso - Alternatives

Tuning the parameter λ of

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

1. Theoretical tuning based on concentration inequalitites
2. Cross Validation procedures
3. The Efron's degree of freedom

Alternatives:

4. Interpret the Lasso procedure as a Bayesian procedure so that $\widehat{\beta}_\lambda^{lasso}$ is the posterior mode. The parameter λ can then be viewed as a hyperparameter of the prior distribution, which can be tuned by using hierarchical Bayes or Empirical Bayes approaches. See Park and Casella (2008).
5. We can combine Lasso and Model Selection: Each value λ of the grid Λ provides a specific model, namely $\widehat{S}_\lambda$. The choice of the best model can be performed by using a Model Selection criterion (Mallows' C_p, AIC, BIC, etc). See Lacroix (2022) for an extensive study and some extensions.

Computing the Lasso estimator

- Coordinate descent algorithm: We wish to compute

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

We assume that $\|X_j\| = 1$ for all j and we denote

$$\mathcal{C}(\beta) = \|Y - X\beta\|^2 + \lambda\|\beta\|_1.$$

- Since $\mathcal{C}$ is convex, the minimizer of

$$\beta_j \mapsto \mathcal{C}(\beta_1, \dots, \beta_{j-1}, \beta_j, \beta_{j+1}, \dots, \beta_p)$$

is

$$\beta_j = R_j \left(1 - \frac{\lambda}{2|R_j|}\right)_+, \quad R_j := X_j^T \left(Y - \sum_{k\neq j} \beta_k X_k\right).$$

Repeatedly computing $\beta_1 \dots, \beta_p$, $\beta_1 \dots, \beta_p$, etc. gives the coordinate descent algorithm summarized below. Thanks to the convexity of $\mathcal{C}$, this algorithm converges to the Lasso estimator.

Coordinate descent algorithm

The coordinate descent algorithm is implemented in the package glmnet.

- Initialization: $\beta = \beta_{init}$, with $\beta_{init} \in \mathbb{R}^p$ arbitrary
- Repeat, until convergence of β, the loop:
  for $j = 1, \ldots, p$

$$\beta_j = R_j \left(1 - \frac{\lambda}{2|R_j|}\right)_+, \quad R_j := X_j^T \left(Y - \sum_{k \neq j} \beta_k X_k\right).$$

- Output: β

Other algorithms:

- FISTA based on linearization of $\beta \longmapsto \|Y - X\beta\|^2$ and gradient descent
- LARS uses that $\lambda \longmapsto \widehat{\beta}_\lambda^{lasso}$ is piecewise affine

Illustration on real data

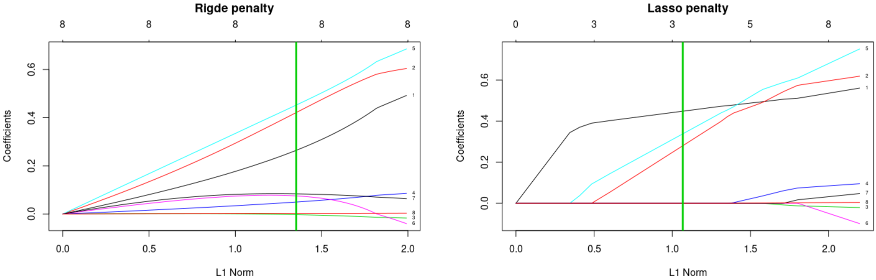
Analysis of the famous ”prostate data“ for $n = 97$ patients, which explains the log(prostate specific antigen) in function of

1. log(cancer volume)
2. log(prostate weight)
3. age
4. log(benign prostatic hyperplasia amount)
5. seminal vesicle invasion
6. log(capsular penetration)
7. Gleason score
8. percentage Gleason scores 4 or 5

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

```
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 = cv.glmnet(x=X,y=Y,,nfolds=10,alpha=0)
plot(ridge.out)
lasso.out = cv.glmnet(x=X,y=Y,,nfolds=10,alpha=1)
plot(lasso.out)
```

Illustration on real data



The x-axis corresponds to $\|\hat{\beta}_\lambda\|_1$. For each graph, the left-hand side corresponds to $\lambda = +\infty$, the right-hand side corresponds to $\lambda = 0$. The vertical green line corresponds to the value of λ determined by cross-validation.

Plan

1. Curse of dimensionality
2. Linear regression setting
3. Classical estimation
4. Ridge estimation: ℓ_2-penalization
5. Model selection (à la Birgé-Massart): ℓ_0-penalization
6. Lasso estimation: ℓ_1-penalization
7. 9 variations of the Lasso

Variation of the Lasso - the Dantzig selector

- Remember that the Lasso estimate satisfies the constraint

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

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

Variation 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

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

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

Variation of the Lasso - Relaxed Lasso

- Instead of introducing weights, Meinshausen (2007) suggests a two-step procedure:
  1. Compute
  $$\widehat{\beta}_{\lambda}^{lasso} = \arg\min_{\beta\in\mathbb{R}^p} \left\{ \|Y - X\beta\|^2 + \lambda\|\beta\|_1 \right\}$$
  and set
  $$\hat{S}_{\lambda} := \left\{ j : \quad \hat{\beta}_{\lambda,j}^{lasso} \neq 0 \right\}.$$
  2. For $\delta \in [0,1]$,
  $$\hat{\beta}_{\lambda,\delta}^{relaxed} := \arg \min_{\beta\in\mathbb{R}^p,\, \mathrm{supp}(\beta)\subset\hat{S}_{\lambda}} \left\{ \|Y - X\beta\|^2 + \delta\lambda\|\beta\|_1 \right\}$$
- 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.

Variation of the Lasso - The square-root Lasso

- The Lasso estimate should be scaled invariant, meaning that for any $s > 0$

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

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

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

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.

Variation of the Lasso - Elastic net

The vanilla Lasso has two drawbacks:

1. In the $p > n$ case, the Lasso selects at most n variables

Variation of the Lasso - Elastic net

The vanilla Lasso has two drawbacks:

1. In the $p > n$ case, the Lasso selects at most n variables
2. In the model $Y = X\beta^* + \epsilon$, consider

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

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

$$\tilde{\beta}_{\lambda,j} = \left\{ \begin{array}{cl} \hat{\beta}_{\lambda,j}^{lasso} & \text{if } j \in \{1,\dots,p-1\} \\ \alpha\hat{\beta}_{\lambda,p}^{lasso} & \text{if } j = p \\ (1-\alpha)\hat{\beta}_{\lambda,p}^{lasso} & \text{if } j = p+1 \end{array} \right. ,$$

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.

Variation of the Lasso - Elastic net

The vanilla Lasso has two drawbacks:

1. In the $p > n$ case, the Lasso selects at most n variables
2. In the model $Y = X\beta^* + \epsilon$, consider

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

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

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

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.
More generally, if there is a group of variables among which the pairwise correlations are very high, then the Lasso tends to select only one variable from the group and does not care which one is selected.

Variation 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.

Variation 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.

Variation 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 $\hat{\beta}^{e.n.}_{\lambda_1,\lambda_2,j} \times \hat{\beta}^{e.n.}_{\lambda_1,\lambda_2,k} > 0$,

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

Variation 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 $\hat{\beta}^{e.n.}_{\lambda_1,\lambda_2,j} \times \hat{\beta}^{e.n.}_{\lambda_1,\lambda_2,k} > 0$,

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

- We can improve $\hat{\beta}^{e.n.}_{\lambda_1,\lambda_2}$ and consider $(1+\lambda_2)\hat{\beta}^{e.n.}_{\lambda_1,\lambda_2}$ (see Zou and Hastie (2005)).

Variation 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

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

- Parameters λ_1 and λ_2 are hard to tune.

Variation 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 and non-empty 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.

- If $K = p$, $\|\beta_{(k)}\| = |\beta_k|$ and the Group-Lasso is the vanilla Lasso.
- As for the Lasso, the group-Lasso can be characterized: For any $k \in \{1, \ldots, K\}$,

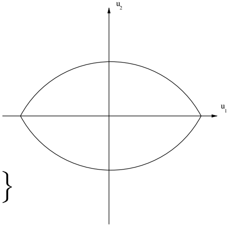
$$\begin{cases} 2X_{(k)}^T(Y - X\hat{\beta}^{group}) = \lambda \times \frac{\hat{\beta}_{(k)}^{group}}{\|\hat{\beta}_{(k)}^{group}\|_2} & \text{if } \hat{\beta}_{(k)}^{group} \neq 0 \\ \left\|2X_{(k)}^T(Y - X\hat{\beta}^{group})\right\| \leq \lambda & \text{if } \hat{\beta}_{(k)}^{group} = 0 \end{cases}$$

- 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.

Variation 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

The contour plots of this penalty function is



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

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 $1 \le k \le K$, $\mathcal{I}_1 \subset G_k \Rightarrow \mathcal{I}_2 \subset G_k$.*
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, $\hat{\beta}^G_{\mathcal{I}_2} \neq 0 \Rightarrow \hat{\beta}^G_{\mathcal{I}_1} \neq 0$.

Variation of the Lasso - The Bayesian Lasso

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

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

  if $\epsilon \sim \mathcal{N}(0,\sigma^2 I_n)$.
- 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.

Take-home message

- To overcome prohibitive computational complexity of model selection, convex critera can be considered leading, in particular, to Lasso-type estimates.

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.

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.

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.

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.

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.
- Lasso type estimates have been presented in the linear regression setting. But it can be extended to other settings: for Generalized Linear Models, density estimation, counting processes, etc.

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.

- Franck, I.E., and Friedman, J.H. A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, 35, 109–148, 1993.

- Hunt X. J., Reynaud-Bouret P., Rivoirard V., Sansonnet L. and Willet R. A data-dependent weighted LASSO under Poisson noise. *IEEE Transactions on Information Theory*, 65, no. 3, 1589–1613, 2019.

- Lacroix, P. *Contributions à la sélection de variables en grande dimension et ses utilisations en biologie*. PhD thesis, 2022.

References

- Meinshausen, N. Relaxed Lasso. *Comput. Statist. Data Anal.*, 52(1),374–393, 2007.
- 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. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* 58, no. 1, 267–288, 1996
- Tibshirani, R. and Taylor, J. Degrees of freedom in lasso problems. *Ann. Statist.* 40(2), 1198–1232, 2012.
- 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.
- Verzelen, Nicolas Minimax risks for sparse regressions: ultra-high dimensional phenomenons. *Electron. J. Stat.*, 6, 38–90, 2012.

References

- Wainwright, Martin J. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory*, 55, no. 5, 2183–2202, 2009.

- 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.

- 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.

- Zhao, P. and Yu, B. On model selection consistency of Lasso. *J. Mach. Learn. Res.* 7, 2541–2563, 2006.

- 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.

Thank you for your attention.

Questions and remarks are welcomed!