High-dimensional Statistical Methods

Skoltech

Q. Paris*

National Research University HSE
Faculty of Computer Science
Moscow, Russia

Introduction

*email:qparis@hse.ru, teaching material: http://www.qparis-math.com/teaching.

Foreword

These notes aim to introduce, in a simple framework, the foundations of some modern statistical methods in the context of high-dimensional data. The topics presented below are in no way an exhaustive overview of the subject and are biased by the personal choices of the author. The material gathered here relies on important reference books or lecture notes to which we refer the reader for additional material and details. As the main references for this course, we recommend the lecture notes by Rigollet (2015) or the book by Giraud (2015). The reader is also invited to use the additional and very valuable references van Handel (2016) or Vershynin (2016). Related material may also be found in Buhlmann and van de Geer (2011); Koltchinskii (2011); Hastie et al. (2015) and the references therein. Below is a tentative program for the course, that may be modified along the way.

1 Fixed design regression

Let $n \geq 1$ and $z_1, \ldots, z_n$ be known and deterministic points (called design or input points) in some input space $\mathcal{Z}$ fixed by the statistician or practitioner. Suppose that, to each of the z_i's, corresponds an observation (or measurement) $Y_i \in \mathbf{R}$ of the form

$$Y_i = f^\star(z_i) + \xi_i, \tag{1.1}$$

where $\xi_1, \ldots, \xi_n$ denote real-valued, centered and independent random variables and $f^\star$ denotes an unknown function $f^\star : \mathcal{Z} \to \mathbf{R}$. In this setting, Y_i stands for some physical measurement $f^\star(z_i)$, relative to z_i, corrupted by some noise ξ_i. From a statistical point of view, the goal is to estimate (or recover) the true vector

$$\mu^\star = \begin{bmatrix} \mu_1^\star \\ \vdots \\ \mu_n^\star \end{bmatrix} \in \mathbf{R}^n \quad \text{where} \quad \mu_i^\star = f^\star(z_i),$$

based on the only knowledge of the observations $Y_1, \ldots, Y_n$.

2 Mean squared error

Given an estimator $\hat{\mu} \in \mathbf{R}^n$ of $\mu^\star$ based on the observations $Y_1, \ldots, Y_n$, a natural measure of its performance is the mean squared error

$$\mathcal{E}(\hat{\mu}) = \frac{1}{n} \sum_{i=1}^{n} (\hat{\mu}_i - \mu_i^\star)^2 = \frac{1}{n} \|\hat{\mu} - \mu^\star\|_2^2,$$

where $\|.\|_2$ denotes the Euclidean norm. We stress that, of course, since $\hat{\mu}$ is constructed from the random Y_i's, the mean squared error $\mathcal{E}(\hat{\mu})$ is itself a (positive) random quantity.

3 Expansion in a dictionary

As will be made clear below, the nature of the input space $\mathcal{Z}$ plays no role in the analysis insofar the input points $z_1, \ldots, z_n$ are considered as known or fixed by the statistician. In particular, $\mathcal{Z}$ could be infinite-dimensional. For instance, the input points could be curves, images, signals, etc. The high-dimensionality of the problem, that will play a fundamental role in the sequel, will be related to a reformulation of the initial problem described below.

Suppose given a collection, also referred to as dictionary, $\varphi_1, \ldots, \varphi_p$ of p known functions $\varphi_j : \mathcal{Z} \to \mathbf{R}$ and suppose for simplicity that the unknown function $f^\star$

can be expanded in the dictionary in the following sense: for some unknown coefficients $\beta_1^\star, \dots, \beta_p^\star \in \mathbf{R}$, we have

$$\forall i \in \{1, \dots, n\} : \quad f^\star(z_i) = \sum_{j=1}^{p} \beta_j^\star \varphi_j(z_i). \tag{3.1}$$

Then, introducing the notation

$$\boldsymbol{x}_i = \begin{bmatrix} \varphi_1(z_i) \\ \vdots \\ \varphi_p(z_i) \end{bmatrix} \in \mathbf{R}^p \quad \text{and} \quad \boldsymbol{\beta}^\star = \begin{bmatrix} \beta_1^\star \\ \vdots \\ \beta_p^\star \end{bmatrix} \in \mathbf{R}^p, \tag{3.2}$$

the reader may easily check that equation (1.1) becomes

$$Y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}^\star + \xi_i. \tag{3.3}$$

The points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbf{R}^p$ will also be called the design points (by extension) and the unknown vector $\boldsymbol{\beta}^\star \in \mathbf{R}^p$ the regression vector. We insist on the fact that the input points $z_1, \dots, z_n$ and the dictionary $\varphi_1, \dots, \varphi_p$ being known and deterministic, the design points are quantities within the knowledge of the statistician during the analysis. The functions in the dictionary can be considered, typically, as some basis functions in a given functional space such as Splines, Fourier or Wavelet basis functions. Other examples of dictionary functions can be tailored for specific applications.

Example 3.1 (**Gene expression level**). *Suppose $\mathcal{Z} = \mathbf{R}$ and $f^\star$ is a piecewise constant function $f^\star : \mathcal{Z} \to \mathbf{R}$. Suppose in addition that the positions $t_1 < \cdots < t_p \in \mathbf{R}$ of potential jumps of $f^\star$ are know but that the value of $f^\star$ on each $[t_j, t_{j+1})$ is not. This situation arises, for instance, in biology (or genetics) when analysing the transcription level*[1] *of genes in a DNA sequence. In this context, the genes in the given sequence (represented by the t_j's here) are known but their transcription level is to be discovered. Here, observe that the function $f^\star$ may be represented as*

$$\forall z \in \mathbf{R} : \quad f^\star(z) = \sum_{j=1}^{p} \beta_j^\star \mathbf{1}\{z \geq t_j\},$$

for known dictionary functions $\varphi_j(z) = \mathbf{1}\{z \geq t_j\}$ and unknown coefficients $\beta_j^\star$. Note therefore that identity (3.1) holds here for any choice of the input points $z_1, \dots, z_n \in \mathbf{R}$.

The interest of the previous reduction is obviously that, provided assumption (3.1) holds, the introduction of the dictionary $\varphi_1, \dots, \varphi_p$ reduces the statistical problematic of finding an unknown function $f^\star$ (of potentially infinite-dimensional nature), to that of estimating a p-dimensional vector $\boldsymbol{\beta}^\star$. In particular, it should be expected that the estimation of $\mu^\star$, in this context, is

[1] The transcription level of a gene in a cell at a given time is the quantity of ARNm associated to this gene present at this time in the cell

highly simplified. However, for a decomposition such as (3.1) to be realistic, the number p (corresponding to the dimension of both $\boldsymbol{\beta}^\star$ and the $\boldsymbol{x}_i$'s) of basis functions is expected to be very large and, in particular, potentially much larger then the sample size n. This context conflicts severely with the scenario of classical statistics in which p is of reasonable magnitude and the sample size n is supposed very large. Here, statistical guarantees are required for fixed n and p with, possibly, $p \gg n$.

Another situation studied in the course corresponds to the case where the considered dictionary $\varphi_1, \ldots, \varphi_p$ is miss-specified, *i.e.*

$$\forall \boldsymbol{\beta} \in \mathbf{R}^p, \ \forall j \in \{1, \ldots, n\} : \quad f^\star(z_i) \neq \sum_{j=1}^{p} \beta_j \varphi_j(z_i).$$

This situation arises naturally when, given too little information on the unknown function $f^\star$, the statistician chooses a dictionary $\varphi_1, \ldots, \varphi_p$ (*i.e.* formulates a model for the unknown $f^\star$) in which one cannot expand the unknown function $f^\star$ as in (3.1). In this context, the goal of the statistician is to guarantee a form of robustness of the statistical methodology considered: given our fixed dictionary, our estimators should do almost as good as the best possible function $\bar{f}$ in the model, *i.e.*

$$\bar{f}(z) = \sum_{j=1}^{p} \bar{\beta}_j \varphi_j(z),$$

where vector $\bar{\boldsymbol{\beta}} = (\bar{\beta}_1, \ldots, \bar{\beta}_p)^\top \in \mathbf{R}^p$ is defined as any

$$\bar{\boldsymbol{\beta}} \in \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbf{R}^p} \frac{1}{n} \sum_{i=1}^{n} (f^\star(z_i) - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2,$$

and $\boldsymbol{x}_i$ is as in (3.2). From a technical point of view, the analysis will be carried out by deriving so called oracle inequalities, quantifying precisely the aforementioned robustness of the considered estimation techniques.

4 The LS criterion and its variants

4.1 Classical LS

In the context of representation (3.3), the reader should be familiar with the idea that, given the data points $(Y_1, \boldsymbol{x}_1), \ldots, (Y_n, \boldsymbol{x}_n) \in \mathbf{R} \times \mathbf{R}^p$, a natural method to estimate $\boldsymbol{\beta}^\star$, and therefore $\mu^\star$, is to minimize the least-squares criterion, *i.e.* consider

$$\hat{\boldsymbol{\beta}}^{\text{ls}} \in \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbf{R}^p} \mathcal{C}^{\text{ls}}(\boldsymbol{\beta}) \quad \text{where} \quad \mathcal{C}^{\text{ls}}(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2, \tag{4.1}$$

and form the estimator $\hat{\mu}^{\mathrm{ls}}$ of $\mu^\star$ with coordinates

$$\hat{\mu}_i^{\mathrm{ls}} = \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}^{\mathrm{ls}}.$$

As will be described in the first chapter, the least squares estimator performs quite badly in most of the cases, when $p \gg n$. In particular, we will prove that, up to universal constants, the expected mean squared error of the least squares estimator satisfies

$$\mathbf{E}\,\mathcal{E}(\hat{\mu}^{\mathrm{ls}}) \approx \frac{\sigma^2 r}{n},$$

(where statement '$\approx$' will be precised) if the noise variables $\xi_1, \ldots, \xi_n$ are independent, centered and sub-gaussian variables with variance proxy σ^2 (see the appendix devoted to this notion) and if r denoted the dimension of the subspace spanned by the design points. This rate of convergence is especially bad if the design is of full rank $r = \min\{n, p\}$. Hence, new estimation techniques are called for to deal efficiently with the high-dimensional nature of the problem.

The bad performance of the least-squares estimator, even though proved formally in the course, can be understood intuitively as follows. The (convex) minimization problem (4.1) encodes absolutely no restriction on the potential candidates $\boldsymbol{\beta} \in \mathbf{R}^p$ leading to an optimal value of the criterion $\mathcal{C}^{\mathrm{ls}}(\boldsymbol{\beta})$. To speak very roughly, the LS estimator results, in a sense, from an 'agnostic' search (no restriction on the candidates $\boldsymbol{\beta}$) of a pin (the vector $\boldsymbol{\beta}^\star$) in a haystack (the space $\mathbf{R}^p$) with only one eye (the sample size n is small compared to p).

4.2 Constrained LS

In some applications, the practitioner (or the expert) may have some intuition or knowledge on the location, or on some characteristics (usually of geometrical nature), of the unknown $\boldsymbol{\beta}^\star$. In this context, one is naturally brought to consider a constrained LS problem of the form

$$\hat{\boldsymbol{\beta}}_{\mathcal{K}}^{\mathrm{ls}} \in \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathcal{K}} \mathcal{C}^{\mathrm{ls}}(\boldsymbol{\beta}), \tag{4.2}$$

where $\mathcal{K}$ denotes a strict subset of $\mathbf{R}^p$. From a mathematical point of view, the analysis of such an estimator can be carried out (at least) for any totally bounded set $\mathcal{K}$. For practical and optimization purposes, one is however led to favor convex sets and a number of our lectures will be devoted to this problem.

4.3 Penalized LS

The framework of constained LS can be, in some instances, considered as too restrictive since one has to specify exactly a subset $\mathcal{K}$ on which to perform the minimization. A more general approach is to penalize the LS criterion.

Precisely, if the intuition on $\boldsymbol{\beta}^\star$ can be formalized by saying that, for some given function $\Omega : \mathbf{R}^p \to [0, +\infty)$, the quantity $\Omega(\boldsymbol{\beta}^\star)$ is likely to be small, a natural alternative to constrained LS is to minimize a penalized version of the least-squares criterion

$$\mathcal{C}^{\text{pls}}(\boldsymbol{\beta}) = \mathcal{C}^{\text{ls}}(\boldsymbol{\beta}) + \lambda\,\Omega(\boldsymbol{\beta}), \tag{4.3}$$

for some parameter $\lambda > 0$ (possibly depending on the sample size n) to be fixed by the statistician. A natural correspondence between the penalized and constrained approach, given by Lagrangian duality, will be established along the way. As for constrained LS, the statistician should (for practical purposes) favor a convex penalty function Ω as the resulting criterion $\boldsymbol{\beta} \mapsto \mathcal{C}^{\text{pls}}(\boldsymbol{\beta})$, also convex, can therefore be optimized using the powerful tools from convex (and stochastic) optimization. This observation will be crucial in the subsequent analysis. On the subject of convex optimization and its recent developments, we refer the reader to Bubeck (2015).

4.4 Sparsity, BIC and the Lasso

In the high-dimensional framework, a possible geometric characteristic of the high-dimensional vector $\boldsymbol{\beta}^\star$, known to be of crucial impact on the performance of statistical methods, is sparsity. The vector $\boldsymbol{\beta}^\star$ is said to be sparse if only a few of its components are non-zero. Formally, $\boldsymbol{\beta}^\star$ is said to be k-sparse if its ℓ_0-norm [2]

$$\|\boldsymbol{\beta}^\star\|_0 = \sum_{j=1}^{p} \mathbf{1}\{\boldsymbol{\beta}^\star_j \neq 0\},$$

is smaller than k. When it is supposed that the unknown $\boldsymbol{\beta}^\star$ is sparse, then the choice of the penalty term $\Omega(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_0$ in (4.3) leads to the BIC [3] estimator,

$$\hat{\boldsymbol{\beta}}^{\text{bic}} \in \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbf{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \lambda^2 \|\boldsymbol{\beta}\|_0 \right\}. \tag{4.4}$$

As we will see, the theoretical performance of the BIC estimator is remarkable in that it adapts to the unknown sparsity of $\boldsymbol{\beta}^\star$ and, contrary to the classical least squares, is much less affected from the dimensionality p of the problem. Unfortunately, from a numerical point of view, computing the BIC estimator is usually unrealistic and no known computational method is known to be significantly better than an exhaustive search among the 2^p possible sparsity patterns of a p-dimensional vector.

[2] This quantity is abusively called a norm by convention.
[3] BIC stands for Bayes Information Criterion

A important alternative to the BIC is the Lasso [4]. The Lasso estimator,

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} \in \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbf{R}^p} \left\{ \frac{1}{n}\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})^2 + 2\lambda\|\boldsymbol{\beta}\|_1 \right\}, \tag{4.5}$$

may be seen as a convex relaxation of the BIC. Both computationally realistic and performant, the Lasso is arguably one of the most popular method in high-dimensional statistics nowadays and a significant part of the lectures will be devoted to its study.

5 Notation

We list below the most important notation used along the course. For any integers $p, q \geq 1$, we denote $\mathrm{M}_{p,q}(\mathbf{R})$ the set of $p \times q$ matrices with real coefficients and set $\mathrm{M}_p(\mathbf{R}) = \mathrm{M}_{p,p}(\mathbf{R})$. We define the vector and matrix notation

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \in \mathbf{R}^n, \quad \mathbf{X} = \begin{bmatrix} \boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{bmatrix} \in \mathrm{M}_{n,p}(\mathbf{R}), \quad \text{and} \quad \boldsymbol{\xi} = \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_n \end{bmatrix} \in \mathbf{R}^n. \tag{5.1}$$

In matrix notation, the least-squares criterion therefore reads

$$\mathcal{C}^{\text{ls}}(\boldsymbol{\beta}) = \frac{1}{n}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2.$$

The columns of the design matrix $\mathbf{X}$ will be denoted $\mathbf{x}^1, \ldots, \mathbf{x}^p \in \mathbf{R}^n$ with index in exponent. We introduce

$$\hat{\Sigma} = \frac{1}{n}\mathbf{X}^\top\mathbf{X} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top \in \mathsf{M}_p(\mathbf{R}).$$

Note in particular that, for any $k, \ell \in \{1, \ldots, p\}$, the entry (k, ℓ) of $\hat{\Sigma}$ is

$$\hat{\Sigma}_{k,\ell} = \frac{(\mathbf{x}^k)^\top\mathbf{x}^\ell}{n}.$$

In other words, the matrix $\hat{\Sigma}$ corresponds to the Gram matrix of the column vectors $\mathbf{x}^1, \ldots, \mathbf{x}^p \in \mathbf{R}^n$ up to a scaling factor $1/n$. Finally, for any $k \geq 1$ and any vector $v = (v_1, \ldots, v_k) \in \mathbf{R}^k$, we will use the classical notation $\|v\|_q$, $1 \leq q \leq +\infty$ to denote the norms

$$\|v\|_q = (\sum_{\ell=1}^{k}|v_\ell|^k)^{1/k}, \text{ if } 1 \leq q < +\infty, \quad \text{and} \quad \|v\|_\infty = \max_{1\leq\ell\leq k}|v_\ell|.$$

[4] Lasso stands for Least Absolute Shrinkage and Selection Operator.

References

S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

P. Buhlmann and S. van de Geer. *Statistics for High-Dimensional Data*. Springer, 2011.

C. Giraud. *Introduction to High-dimensional Statistics*. CRC Press, 2015.

T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning With Sparsity: The Lasso and Generalizations*. CRC Press, 2015.

V. Koltchinskii. Oracle inequalities in empirical risk minimization and sparse recovery problems. In *Lectures from the 38th Summer School on Probability Theory held in Saint-Flour in July, 2008*. Springer, 2011.

P. Rigollet. High-dimensional statistics. MIT lecture notes, 2015.

R. van Handel. Probability in high dimension. Princeton lecture notes, 2016.

R. Vershynin. High-dimensional probability: An introduction with applications in data science. Book in preparation, 2016.