Topics in High-Dimensional Statistics

Lecture 1: Introduction to high-dimensional linear regression

Contents

1 Basic framework and goal

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

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

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

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

based on the only knowledge of the observations $Y_1, \ldots, Y_n$ and the input points $z_1, \ldots, z_n$. Given an estimator $\hat{\mu} \in \mathbb{R}^n$ of μ^* based on the observations $Y_1, \ldots, Y_n$, a natural measure of its performance is the ***mean squared error*** (abbreviated MSE) defined by

$$\mathrm{MSE}(\hat{\mu}) := \frac{1}{n} \sum_{i=1}^{n} (\hat{\mu}_i - \mu_i^*)^2 = \frac{1}{n} \|\hat{\mu} - \mu^*\|_2^2,$$

where $\|.\|_2$ denotes the Euclidean norm.

2 Linear modeling

At this point, a typical statistical strategy consists in linear modeling. This strategy consists in choosing a collection $\{f_1, \ldots, f_d\}$ (referred to as ***dictionary***) of d known functions $f_j : \mathcal{Z} \to \mathbb{R}$ and assuming that $f^* \in \mathcal{F}$ where

$$\mathcal{F} := \left\{ \sum_{j=1}^{d} \beta_j f_j : \beta_1, \ldots . \beta_d \in \mathbb{R} \right\}. \tag{2.1}$$

In this context, the assumption that $f^* \in \mathcal{F}$ (in which case we say that model $\mathcal{F}$ is ***well specified***) means that

$$\exists\, \beta_1^*, \ldots, \beta_d^* \in \mathbb{R},\ \forall z \in \mathcal{Z} : \quad f^*(z) = \sum_{j=1}^{d} \beta_j^* f_j(z). \tag{2.2}$$

If this property holds, introducing the notation

$$x_i = \begin{bmatrix} f_1(z_i) \\ \vdots \\ f_d(z_i) \end{bmatrix} \in \mathbb{R}^d \quad \text{and} \quad \beta^* = \begin{bmatrix} \beta_1^* \\ \vdots \\ \beta_d^* \end{bmatrix} \in \mathbb{R}^d, \tag{2.3}$$

equation (1.1) becomes

$$Y_i = x_i^\top \beta^* + \xi_i. \tag{2.4}$$

The input points $z_1, \dots, z_n$ and the dictionary functions $f_1, \dots, f_d$ being known and deterministic, the vectors $x_1, \dots, x_n$ are fixed quantities within the knowledge of the statistician during the analysis. The vectors $x_1, \dots, x_n$ will also be called the ***design points*** and the unknown vector $\beta^* \in \mathbb{R}^d$ the ***regression vector***. In many applications, the dictionary functions can be chosen quite naturally as described in the next example.

Example 2.1 (Image denoising)**.** *Consider the problem of recovering the true colors at n selected points in an image corrupted by noise (a blured image). To formalize the problem, let $q \geq 2$ be some (large) integer and let us represent, for simplicity, an image as the square $[0,1)^2$ constituted of the q^2 pixels*

$$[t_k, t_{k+1}) \times [t_\ell, t_{\ell+1}), \quad k, \ell \in \{0, \dots, q-1\},$$

where $t_j := j/q$. Suppose that we select points $z_1, \dots, z_n \in [0,1)^2$ and that we measure the color Y_i (coded as a number in $\mathbb{R}$ for simplicity) of the image at z_i. Saying that the image is corrupted by noise may be modelled, precisely, by saying that $Y_i = f^*(z_i) + \xi_i$ for some unknown $f^* : [0,1)^2 \to \mathbb{R}$ and random noise ξ_i. Here it is understood that f^* represents the underlying and uncorrupted image in the sence that $f^*(z)$ is the true color of the image at position z. Now, the choice of the dictionary functions may be done naturally by noticing that the image has uniform color on a given pixel so that f^* must be constant on each $[t_k, t_{k+1}) \times [t_\ell, t_{\ell+1})$. In other words, letting $f_{k,\ell} : [0,1)^2 \to \mathbb{R}$ be the function defined by

$$\forall z \in [0,1)^2 : \quad f_{k,\ell}(z) := \mathbf{1}_{[t_k, t_{k+1}) \times [t_\ell, t_{\ell+1})}(z),$$

it is clear that the unknown function f^* belongs to model

$$\mathcal{F} := \left\{ \sum_{0 \leq k,\ell \leq q-1} \beta_{k,\ell} f_{k,\ell} : \beta_{k,\ell} \in \mathbb{R} \right\}.$$

and identity (2.2) holds in this situation.

This last example shows that, in some applications, coming up with a dictionary may be done naturally and may lead to an automatically well-specified model. In more complex situations, another strategy is to select dictionary functions in a given functional basis.

Example 2.2 (Fourier basis)**.** *For some $T > 0$, suppose that $\mathcal{Z} = [0, T]$ and that the unknown $f^* : [0, T] \to \mathbb{R}$ is square-integrable. Next, define the functions $f_j, j \in \mathbb{Z}$, by $f_0(z) = 1$,*

$$f_{2k}(z) := \sqrt{\frac{2}{T}} \cos\left(k \frac{2\pi}{T} z\right),$$

and

$$f_{2k+1}(z) := \sqrt{\frac{2}{T}} \sin\left(k \frac{2\pi}{T} z\right).$$

Then the Fourier expansion of f^* is

$$f^* = \sum_{j \in \mathbb{Z}} c_j^* f_j \quad \textit{where} \quad c_j^* = \int_0^T f^*(z) f_j(z) \, \mathrm{d}z,$$

and where the first equality holds in $\mathbb{L}^2([0, T], \mathrm{d}z)$. In this context, taking $f_{-m}, \ldots, f_0, \ldots, f_m$ as dictionary functions (for a large value of m) is a reasonable choice. However, in full generality, there is no reason for model

$$\mathcal{F} := \left\{ \sum_{j=-m}^{m} c_j f_j : c_j \in \mathbb{R} \right\},$$

to be well-specified.

For simplicity, we consider in the sequel that the model $\mathcal{F}$ is well-specified, *i.e.* that assumption (2.2) holds.

3 The high-dimensional setup

The interest of the modeling step presented in the previous paragraph is obviously that, provided assumption (2.2) holds for a given set $f_1, \ldots, f_d$ of dictionary functions, the problem of finding an unknown function f^* is reduced to that of estimating a d-dimensional vector β^*. However, for a decomposition such as (2.2) to be realistic, the number d (corresponding to the dimension of both β^* and the x_i's) of dictionary 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 d is of small magnitude and the sample size n is supposed much larger. Here, statistical guarantees are required for fixed n and d with, possibly, $d \gg n$. In the sequel, the high-dimensional setup will refer to the situation where $d > n$.

4 The failure of ordinary least squares

In the context of representation (2.4), and given the data

$$(x_1, Y_1), \ldots, (x_n, Y_n) \in \mathbb{R}^d \times \mathbb{R},$$

a natural method to estimate β^*, and therefore μ^*, is to minimize the ordinary least squares (OLS) criterion, *i.e.* consider

$$\hat{\beta}^{\mathrm{ls}} \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^d} \mathcal{C}(\beta) \quad \text{where} \quad \mathcal{C}(\beta) := \frac{1}{n} \sum_{i=1}^{n} (Y_i - x_i^\top \beta)^2. \tag{4.1}$$

The related estimator $\hat{\mu}^{\mathrm{ls}}$ of μ^* is the vector with coordinates

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

As described in this section, the OLS strategy unfortunately fails in the high-dimensional context.

Matrix notation. In the sequel, it will be convenient to encode representation (2.4) in matrix notation. We denote

$$\mathbf{Y} = \mathbf{X}\beta^* + \boldsymbol{\xi},$$

where

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \in \mathbb{R}^n, \quad \mathbf{X} = \begin{bmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{bmatrix} \in M_{n,d}(\mathbb{R}), \quad \text{and} \quad \boldsymbol{\xi} = \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_n \end{bmatrix} \in \mathbb{R}^n.$$

In this context, the least squares estimator $\hat{\mu}^{\mathrm{ls}}$ and the unknown vector $\mu^* \in \mathbb{R}^n$ are given by

$$\hat{\mu}^{\mathrm{ls}} = \mathbf{X}\hat{\beta}^{\mathrm{ls}} \quad \text{and} \quad \mu^* = \mathbf{X}\beta^*,$$

where

$$\hat{\beta}^{\mathrm{ls}} \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^d} \mathcal{C}(\beta) \quad \text{and} \quad \mathcal{C}(\beta) = \frac{1}{n} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2. \tag{4.2}$$

In particular, the mean squared error writes

$$\mathrm{MSE}(\hat{\mu}^{\mathrm{ls}}) = \frac{1}{n} \|\hat{\mu}^{\mathrm{ls}} - \mu^*\|_2^2 = \frac{1}{n} \|\mathbf{X}(\hat{\beta}^{\mathrm{ls}} - \beta^*)\|_2^2.$$

We now review some basic facts.

Theorem 4.1. *The following statements hold.*

(1) *The function* $\beta \mapsto \mathcal{C}(\beta)$ *is convex and* $\nabla \mathcal{C}(\beta) = 2\mathbf{X}^\top(\mathbf{X}\beta - \mathbf{Y})/n$.

(2) *The properties of convex functions guarantee that*

$$\hat{\beta} \in \arg\min_{\beta \in \mathbb{R}^d} \mathcal{C}(\beta) \quad \Leftrightarrow \quad \nabla \mathcal{C}(\hat{\beta}) = 0$$
$$\Leftrightarrow \quad \mathbf{X}^\top \mathbf{X}\hat{\beta} = \mathbf{X}^\top \mathbf{Y}.$$

(3) *If* $\mathrm{rank}(\mathbf{X}) = d$, *then* $\mathbf{X}^\top\mathbf{X} \in M_{d,d}(\mathbb{R})$ *is invertible and* $\hat{\beta}^{\mathrm{ls}}$ *is uniquely defined by*

$$\hat{\beta}^{\mathrm{ls}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}.$$

(4) *If* $\mathrm{rank}(\mathbf{X}) < d$ *(which is necessarily the case if* $n < d$*), then a solution (not unique) of (4.2) is defined by*

$$\hat{\beta}^{\mathrm{ls}} = (\mathbf{X}^\top\mathbf{X})^{+}\mathbf{X}^\top\mathbf{Y},$$

where, for any matrix A, *we denote* A^+ *its pseudo inverse*[1].

Proof. In seminar. □

We are now in position to describe the performance of the OLS estimator as a function of the sample size $n \geq 1$, the dimension $d \geq 1$ and the noise distribution. From now on, notation

$$A \lesssim B,$$

means that $A \leq CB$ for some numerical constant $C > 0$.

Theorem 4.2. *Let* $r = \mathrm{rank}(\mathbf{X})$. *Suppose that the noise vector* $\boldsymbol{\xi} \in \mathbb{R}^n$ *is sub-gaussian with variance proxy* $\sigma^2 > 0$. *Then the following statements hold.*

(1) *For all* $n \geq 1$,

$$\mathbb{E}[\mathrm{MSE}(\hat{\mu}^{\mathrm{ls}})] \lesssim \frac{r\sigma^2}{n}.$$

(2) *For all* $n \geq 1$ *and all* $\delta \in (0,1)$,

$$\mathrm{MSE}(\hat{\mu}^{\mathrm{ls}}) \lesssim \frac{r\sigma^2}{n} + \frac{\sigma^2}{n}\log\left(\frac{1}{\delta}\right),$$

with probability at least $1 - \delta$.

[1]The Moore-Penrose pseudo inverse of a matrix generalizes the notion of inverse for singular matrices. For any $A \in M_{p,q}(\mathbb{R})$ its pseudo inverse A^+ is a matrix in $M_{q,p}(\mathbb{R})$ such that $AA^+x = x$, $\forall x \in \mathrm{Im}(A)$, and such that $A^+Ay = y$, $\forall y \in \mathrm{Im}(A^\top)$. In particular $A^+ = A^{-1}$ when A is a square and invertible matrix.

Proof. In seminar. □

In the high-dimensional context where $d > n$, and if the design matrix $\mathbf{X}$ has full rank, i.e. $r = min\{n, d\} = n$, the above bound reduces to

$$\mathbb{E}[\mathrm{MSE}(\hat{\mu}^{\mathrm{ls}})] \lesssim \sigma^2.$$

This performance guarantee is bad for a large value of σ^2 and is of no interest from a statistical perspective. Hence, new estimation techniques are needed to deal efficiently with the high-dimensional nature of the problem.

5 Penalized least squares

One way around the problem is the penalized least squares approach. The method allows to encode formally some intuition about the solution β^* into the statistical procedure. Precisely, if the intuition on β^* can be formalized by saying that, for some given ***penalty function*** $\pi : \mathbb{R}^d \to [0, +\infty)$, the quantity $\pi(\beta^*)$ is likely to be small, a natural modification of the least squares is to select

$$\hat{\beta} \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^d} \{\, \mathcal{C}(\beta) + \lambda\, \pi(\beta) \,\}, \tag{5.1}$$

for some parameter $\lambda > 0$ (possibly depending on the sample size n) to be fixed by the statistician.

Remark 5.1 (MAP estimator)**.** *The principle of penalized least squares benefits from a natural bayesian interpretation when the noise vectors $\xi_1, \ldots, \xi_n$ are i.i.d. with distribution $\mathcal{N}(0, \sigma^2)$. Indeed, the intuition according to which the true parameter β^* is such that that $\pi(\beta^*)$ is small can be, in a bayesian framework, formalized by considering the prior distribution*

$$\frac{e^{-\pi(\beta)} \mathrm{d}\beta}{\int e^{-\pi(u)} \mathrm{d}u},$$

on the parameter set $\mathbb{R}^d$, provided $\beta \mapsto e^{-\pi(\beta)}$ is integrable. Note that, by construction, this distribution indeed favors parameters β such that $\pi(\beta)$ is small. From a bayesian perspective, a natural estimation technique, called ***maximum a posteriori*** *(or MAP), consists in considering the value of β maximizing the (posterior) density of β given the observations $Y_1, \ldots, Y_n$. Bayes formula indicates that this conditional density is precisely proportional (check it) to*

$$\exp\left(-\frac{n}{2\sigma^2}\left\{\mathcal{C}(\beta) + \frac{2\sigma^2}{n}\pi(\beta)\right\}\right),$$

so that the MAP estimator (maximizing the above expression) corresponds exactly to the penalized least squares estimator (5.1) *with* $\lambda = 2\sigma^2/n$. *Note, however, that this interpretation holds only in the context where the noise variables* $\xi_1, \ldots, \xi_n$ *are i.i.d. with distribution* $\mathcal{N}(0, \sigma^2)$.

6 Sparsity

In the high-dimensional framework, a possible geometric characteristic of the high-dimensional vector β^*, known to be of crucial impact on the performance of statistical methods, is sparsity. The vector β^* is said to be sparse if only a few of its components are non-zero. Formally, β^* is said to be sparse if its ℓ_0-norm [2] $\|\beta^*\|_0$, defined by

$$\|\beta^*\|_0 = \sum_{j=1}^{d} \mathbf{1}\{\beta_j^* \neq 0\},$$

is small.

7 The BIC estimator

When it is supposed that the unknown β^* is sparse, then the choice of the penalty term $\pi(\beta) = \|\beta\|_0$ in (5.1) leads to the BIC [3] estimator,

$$\hat{\beta}^{\text{bic}} \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^d} \{\, \mathcal{C}(\beta) + \lambda^2 \|\beta\|_0 \,\}. \tag{7.1}$$

(The form of the parameter λ^2 is here conventional). As described next, the theoretical performance of the BIC estimator is remarkable as it adapts to the unknown sparsity of β^* and, contrary to the LS estimator, is much less affected from the dimensionality d of the problem.

Theorem 7.1. *Suppose that the noise vector* $\boldsymbol{\xi} \in \mathbb{R}^n$ *is sub-gaussian with variance proxy* $\sigma^2 > 0$. *Then, there exists a universal constant* $c > 0$ *such that if*

$$\lambda^2 \geq c \frac{\sigma^2 \log d}{n},$$

then, for all $n \geq 1$ *and all* $\delta \in (0, 1)$,

$$\text{MSE}(\hat{\mu}^{\text{bic}}) \lesssim \frac{\sigma^2 \|\beta^*\|_0}{n} \log\left(\frac{d}{\delta}\right),$$

[2] This quantity is abusively called a norm by convention but is not a norm in the mathematical sense.

[3] BIC stands for Bayes Information Criterion

with probability at least $1-\delta$.

Proof. In seminar. □

From a theoretical point of view, the performance of the BIC estimator is optimal. Roughly speaking, one can show that no estimator can provide a better performance whenever the noise is sub-gaussian. Unfortunately, from a numerical point of view, computing the BIC estimator is usually unrealistic due in particular to the non-convexity of the objective function in (7.1). In fact, no known computational method does significantly better than an exhaustive search among the 2^d possible sparsity patterns of a d-dimensional vector. The next question is therefore to find an estimator that recovers similar statistical performance while being computationally realistic.

8 The LASSO estimator

An important alternative to the BIC estimator is the LASSO [4] estimator. One way to motivate the construction of the LASSO estimator is to see it as a convex relaxation of the BIC criterion. To that aim, note that for all $\beta \in \mathbb{R}^d$,

$$\lim_{p \to 0+} \|\beta\|_p^p = \|\beta\|_0,$$

where

$$\|\beta\|_p^p = \sum_{j=1}^{d} |\beta_j|^p.$$

Note then that the smallest value of $p > 0$ for which the map $\beta \mapsto \|\beta\|_p^p$ is convex is $p = 1$. The LASSO estimator is obtained by

$$\hat{\beta}^{\text{lasso}} \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^d} \{ \, \mathcal{C}(\beta) + 2\lambda \|\beta\|_1 \, \}, \tag{8.1}$$

where

$$\|\beta\|_1 := \sum_{j=1}^{d} |\beta_j|,$$

The next lectures will show that the LASSO estimator is both computationally realistic and statistically performant in the high-dimensional setting, with a performance close to that of the BIC estimator.

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

9 Recommended literature

For reading on the subject of High-dimensional statistics, we recommended [2, 3, 1, 4, 5, 6]

References

[1] A. Bandeira. Ten lectures and forty-two open problems in the mathematics of data science. Lecture notes, 2016.

[2] D. Chafaï, O. Guedon, G. Lecue, and A. Pajor. *Interactions between compressed sensing, random matrices and high-dimensional geometry*. Unpublished lecture notes, 2009.

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

[4] P. Rigollet and J.-C. Hütter. High-dimensional statistics. Lecture notes, 2017.

[5] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2018.

[6] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2019.