High-dimensional Statistical Methods

Skoltech

Q. Paris*

National Research University HSE
Faculty of Computer Science
Moscow, Russia

Chapter 1. High-dimensional regression

Lecture 1

Least squares and constrained least squares

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

The material discussed below is inspired from **????** and **?**. The proofs of the results gathered below are provided during the lectures.

1 Least squares

1.1 Basics

Let us first recall the general problem considered as well as some notation from the introduction. Let $n \geq 1$ and $z_1, \ldots, z_n$ be deterministic input points, known (or given) to the statistician, in some input space $\mathcal{Z}$. To each of the z_i's, corresponds an observation $Y_i \in \mathbf{R}$ of the form

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

where $\xi_1, \ldots, \xi_n$ are real-valued and independent random variables. Here, $f^\star : \mathcal{Z} \to \mathbf{R}$ denotes an unknown function and the goal is, based only on the observations $Y_1, \ldots, Y_n$, to recover the true vector

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

Let $\{\varphi_1, \ldots, \varphi_p\}$ be a collection of known functions $\varphi_j : \mathcal{Z} \to \mathbf{R}$ (referred to as the dictionary). Suppose that the unknown function $f^\star$ can be expanded in the dictionary in the sense that, for some unknown coefficients $\beta_1^\star, \ldots, \beta_p^\star \in \mathbf{R}$, we have

$$\forall i \in \{1, \ldots, n\} : \quad f^\star(z_i) = \sum_{j=1}^{p} \beta_j^\star \varphi_j(z_i). \tag{1.2}$$

With the notation

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

equation (1.1) therefore becomes $Y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}^\star + \xi_i$. In matrix form, we obtain

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

where we recall that

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

In this context, the least squares estimator $\hat{\mu}^{\mathrm{ls}}$ of the unknown vector $\mu^\star \in \mathbf{R}^n$ is defined by

$$\hat{\mu}^{\mathrm{ls}} = \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathrm{ls}},$$

where

$$\hat{\boldsymbol{\beta}}^{\mathrm{ls}} \in \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbf{R}^p} \mathcal{C}^{\mathrm{ls}}(\boldsymbol{\beta}) \quad \text{and} \quad \mathcal{C}^{\mathrm{ls}}(\boldsymbol{\beta}) = \frac{1}{n}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2. \tag{1.3}$$

We now review some basic facts about this estimation technique that should be familiar to the reader.

Theorem 1.1. *The following statements hold.*

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

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

$$\begin{aligned} \hat{\boldsymbol{\beta}} \in \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbf{R}^p} \mathcal{C}^{\mathrm{ls}}(\boldsymbol{\beta}) \quad &\Leftrightarrow \quad \nabla\mathcal{C}^{\mathrm{ls}}(\hat{\boldsymbol{\beta}}) = 0 \\ &\Leftrightarrow \quad \mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^\top\mathbf{Y}. \end{aligned}$$

(3) *If* $\mathrm{rk}(\mathbf{X}) = p$, *then* $\mathbf{X}^\top\mathbf{X} \in \mathrm{M}_p(\mathbf{R})$ *is invertible and* $\hat{\boldsymbol{\beta}}^{\mathrm{ls}}$ *is uniquely defined by*

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

(4) *If* $\mathrm{rk}(\mathbf{X}) < p$, *then a solution (not unique) of* (1.3) *is defined by*

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

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

[a] The Moore-Penrose pseudo inverse of a matrix generalizes the notion of inverse for singular matrices. For any $A \in \mathrm{M}_{p,q}(\mathbf{R})$ its pseudo inverse A^+ is a matrix in $\mathrm{M}_{q,p}(\mathbf{R})$ 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.

1.2 Performance bound

In this paragraph, we study the performance of the estimator $\hat{\mu}^{\mathrm{ls}}$ for any values of $n \geq 1$, the sample size, and $p \geq 1$, the dimension of the design points. In particular, the following results hold in the high-dimensional context, *i.e.* when p is much larger then $n \geq 1$.

Theorem 1.2. *Let $r = \mathrm{rk}(\mathbf{X})$. Suppose that the noise vector $\boldsymbol{\xi} \in \mathbf{R}^n$ is sub-gaussian with variance proxy $\sigma^2 > 0$. Then the following statements hold.*

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

$$\mathcal{E}(\hat{\mu}^{\mathrm{ls}}) \leq \frac{16\sigma^2}{n} \left\{ 2r + \log\left(\frac{1}{\delta}\right) \right\},$$

with probability at least $1 - \delta$.

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

$$\mathbf{E}\mathcal{E}(\hat{\mu}^{\mathrm{ls}}) \leq 8e^{\frac{2}{e}} \frac{\sigma^2 r}{n}.$$

Note that $c = 8e^{\frac{2}{e}} \leq 17$.

The previous theorem, valid for any values of $n \geq 1$, $p \geq 1$ and $r \geq 1$, highlights an important drawback of the least squares method in the high-dimensional context. Indeed, if $p > n$ and the design matrix $\mathbf{X}$ is of full rank n, then the previous upper bounds are of order σ^2, which may be quite large.

2 Constrained least-squares

This section investigates the favorable case where some preliminary, or apriori, information on the unknown $\boldsymbol{\beta}^\star$ can be implemented in the statistical procedure in the form of an explicit constraint, *i.e.* when one considers

$$\hat{\mu}^{\mathrm{ls}}_{\mathcal{K}} = \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathrm{ls}}_{\mathcal{K}} \quad \text{where} \quad \hat{\boldsymbol{\beta}}^{\mathrm{ls}}_{\mathcal{K}} \in \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathcal{K}} \mathcal{C}^{\mathrm{ls}}(\boldsymbol{\beta}),$$

for some explicit $\mathcal{K} \subset \mathbf{R}^p$. Below, we focus on two specific examples for the constraint $\mathcal{K}$, of specific importance in the high-dimensional setting.

2.1 ℓ_0-constrained least-squares

This paragraph studies the case where the unknown $\boldsymbol{\beta}^\star$ is apriori known to be s-sparse. Recall that $\boldsymbol{\beta} \in \mathbf{R}^p$ is said s-sparse if $\|\boldsymbol{\beta}\|_0 \leq s$ where

$$\|\boldsymbol{\beta}\|_0 = \sum_{j=1}^{p} \mathbf{1}\{\beta_j \neq 0\}.$$

This information corresponds obviously to the constraint

$$\mathcal{K} = \{\boldsymbol{\beta} \in \mathbf{R}^p : \|\boldsymbol{\beta}\|_0 \leq s\}.$$

It should be noted that, for this constraint, the computation of $\hat{\boldsymbol{\beta}}^{\mathrm{ls}}_{\mathcal{K}}$ is unrealistic in practice for large values of p. Indeed, it requires to minimizes $\binom{p}{s}$ least-squares criterions in dimension s. However, the results presented below stand as a benchmark for the next paragraph.

Theorem 2.1. *Let $s \geq 1$ be a integer smaller than $p/2$. Suppose that $\boldsymbol{\beta}^\star \in \mathcal{K}$ and the noise vector $\boldsymbol{\xi} \in \mathbf{R}^n$ is sub-gaussian with variance proxy $\sigma^2 > 0$. Then the following statements hold.*

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

$$\mathcal{E}(\hat{\mu}_{\mathcal{K}}^{\text{ls}}) \leq \frac{8\sigma^2}{n} \left\{ 2s \log\left(\frac{3ep}{s}\right) + \log\left(\frac{1}{\delta}\right) \right\},$$

with probability at least $1 - \delta$.

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

$$\mathbf{E}\mathcal{E}(\hat{\mu}_{\mathcal{K}}^{\text{ls}}) \leq \frac{8\sigma^2}{n} \left\{ 1 + 2s \log\left(\frac{3ep}{s}\right) \right\}.$$

Contrary to the global LS estimator, the ℓ_0-constrained LS exhibits remarkable properties. First, its (theoretical) performance is not affected by the rank of the design matrix $\mathbf{X}$ and depends on the dimension p only through a log term. As mentioned above, the computation of this estimator is however computationally unrealistic.

2.2 ℓ_1-constrained least-squares

This paragraph studies a computationally friendly alternative to the ℓ_0-constrained LS. If the unknown $\boldsymbol{\beta}^\star$ is supposed sparse and bounded, the inequality

$$\|\boldsymbol{\beta}\|_1 \leq \min\left\{ \|\boldsymbol{\beta}\|_2 \sqrt{\|\boldsymbol{\beta}\|_0}\,,\, \|\boldsymbol{\beta}\|_\infty \|\boldsymbol{\beta}\|_0 \right\},$$

suggests that $\boldsymbol{\beta}^\star$ lies in an ℓ_1-ball of small radius. Next, consider, therefore the constraint

$$\mathcal{K} = \{\boldsymbol{\beta} \in \mathbf{R}^p : \|\boldsymbol{\beta}\|_1 \leq \lambda\},$$

for some $\lambda > 0$. Contrary to the previous paragraph, the ℓ_1-constrained LS can be computed very efficiently (see below). The next result, provides upper bounds on its theoretical performance.

Theorem 2.2. *Suppose that* $\boldsymbol{\beta}^\star \in \mathcal{K}$ *and the noise vector* $\boldsymbol{\xi} \in \mathbf{R}^n$ *is subgaussian with variance proxy* $\sigma^2 > 0$. *Denote*

$$\varkappa := \sup_{1 \le j \le p} \|\mathbf{x}^j\|_2,$$

where $\mathbf{x}^1, \dots, \mathbf{x}^p \in \mathbf{R}^n$, *denote the columns of the design matrix* $\mathbf{X}$. *Then the following statements hold.*

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

$$\mathcal{E}(\hat{\mu}^{\mathrm{ls}}_{\mathcal{K}}) \le \frac{4\sigma\varkappa}{n}\sqrt{2\log\left(\frac{2p}{\delta}\right)},$$

with probability at least $1-\delta$.

(2) *For all* $n \ge 1$,

$$\mathbf{E}\mathcal{E}(\hat{\mu}^{\mathrm{ls}}_{\mathcal{K}}) \le \frac{4\sigma\varkappa\sqrt{2\log(2p)}}{n}.$$

In practice, normalize the design matrix such that $\varkappa \le \sqrt{n}$.