# Chapter 3: linear regression

Edouard Pauwels

Statistics and optimization in high dimensions

**M2RI, Toulouse 3 Paul Sabatier**

# Main model

Generative model

$$Y_i = f^*(X_i) + \epsilon_i, \quad i = 1\dots, n$$

where $\epsilon = (\epsilon_1, \dots, \epsilon_n)^T \sim \mathrm{subG}(\sigma^2)$ and $\mathbb{E}[\epsilon] = 0$.

$f^* : x \mapsto \mathbb{E}[Y|X = x]$ of form $f^* : x \mapsto x^T\theta^*$ for an unknown $\theta^* \in \mathbb{R}^d$.

**Fixed design:** the design points $x_1, \dots, x_n \in \mathbb{R}^d$ are fixed and are given by the rows of $\mathbb{X} \in \mathbb{R}^{n\times d}$. We have the idendity in $\mathbb{R}^n$.

$$Y = \mathbb{X}\theta^* + \epsilon \tag{LM}$$

**Measure of performance:** the notion of risk is not meaningful here (no randomness on $X$), we use the mean squared error.

$$g : x \mapsto \theta^T x \qquad (\theta \in \mathbb{R}^d)$$

$$\mathrm{MSE}(g) = \frac{1}{n}\sum_{i=1}^{n}(g(x_i) - f^*(x_i))^2$$

$$\mathrm{MSE}(\theta) = \frac{1}{n}\|\mathbb{X}(\theta - \theta^*)\|_2^2.$$

# Least squares estimator

$$\hat{\theta}^{LS} \in \arg\min_{\theta \in \mathbb{R}^d} \|\mathbb{X}\theta - Y\|_2^2 \tag{3.1}$$

**Lemma (3.2.1)**

*We have*

$$\mathbb{X}^T \mathbb{X} \hat{\theta}^{LS} = \mathbb{X}^T Y$$

*and one solution is given by* $\hat{\theta}^{LS} = (\mathbb{X}^T \mathbb{X})^{\dagger} \mathbb{X}^T Y$, *where* $\dagger$ *denotes the Moore-Menrose pseudo inverse.*

Let $K$ denote a closed subset of $\mathbb{R}^d$, the $K$ constrained least squares estimator is given by

$$\hat{\theta}_K^{LS} \in \arg\min_{\theta \in K} \|\mathbb{X}\theta - Y\|_2^2 \tag{3.2}$$

**Lemma (3.3.1)**

*Let $K \subset \mathbb{R}^d$ be closed and $g : \mathbb{R}^d \mapsto \mathbb{R}$ denote any function. Assume that model* (LM) *holds and that $\theta^* \in K$, and set, assuming that the infimum is attained*

$$\hat{\theta}_{Kg}^{LS} \in \arg\min_{\theta \in K} \|\mathbb{X}\theta - Y\|_2^2 + g(\theta).$$

*Then, almost surely*

$$\|\mathbb{X}(\hat{\theta}_{Kg}^{LS} - \theta^*)\|_2^2 \leq 2\epsilon^T \mathbb{X}(\hat{\theta}_K^{LS} - \theta^*) + g(\theta^*) - g(\hat{\theta}_{Kg}^{LS}).$$

**Theorem (3.3.1)**

*Assume that* (LM) *holds with* $\epsilon \sim \mathrm{subG}(\sigma^2)$, *then*

$$\mathbb{E}\left[\mathrm{MSE}(\hat{\theta}^{LS})\right] \leq 16\sigma^2 \frac{r}{n}$$

*where* $r = \mathrm{rank}(\mathbb{X}^T\mathbb{X})$, *furthermore, for any* $\delta > 0$, *with probability at least* $1 - \delta$,

$$\mathrm{MSE}(\hat{\theta}^{LS}) \leq \frac{64\sigma^2 (2r + \log(1/\delta))}{n}$$

# Optimality and high dimensional setting

If $\mathbb{X}$ has full possible rank, then $r = \min(n, d) = d$ assuming $n \geq d$ and

$$\mathrm{MSE}(\hat{\theta}^{LS}) = (\hat{\theta}^{LS} - \theta^*)^T \frac{\mathbb{X}^T\mathbb{X}}{n}(\hat{\theta}^{LS} - \theta^*) \geq \lambda_{\min}\left(\frac{\mathbb{X}^T\mathbb{X}}{n}\right) \|\hat{\theta}^{LS} - \theta^*\|_2^2.$$

**Theorem (3.3.2)**

*Suppose that $Y = \xi + \theta$ where $\theta \in \mathbb{R}^d$ and $\xi_i \sim \mathcal{N}(0, \sigma^2/n)$, $i = 1, \ldots, d$. Then, for any $\alpha \in (0, 1/4)$:*

$$\inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}^d} \mathbb{P}_\theta\left(\|\hat{\theta} - \theta\|_2^2 \geq \frac{\alpha}{256}\frac{\sigma^2 d}{n}\right) \geq \frac{1}{2} - 2\alpha$$

*where the infimum is taken over all measurable functions of $Y$.*

Reduction to finite hypothesis testing, information theoretic lower bounds, see chapter 4 of the lecture notes.

We let $B_1$ denote the unit ball of the $\ell_1$ norm in $\mathbb{R}^d$,

$$B_1 = \left\{ x \in \mathbb{R}^d, \sum_{i=1}^{d} |x_i| \leq 1 \right\}.$$

This is a polytope with $2d$ vertices given by the elements of the canonical basis and their oposite.

**Theorem (3.3.3)**

*Let $K = B_1$ and $d \geq 2$. Assume that model (LM) holds with $\epsilon \sim \mathrm{subG}(\sigma^2)$ and $\theta^* \in K$. Assume also that the columns of $\mathbb{X}$ are normalized such that $\|\mathbb{X}_j\| \leq \sqrt{n}$, $j =, 1 \dots, d$. Then, it holds that*

$$\mathbb{E}\left[\mathrm{MSE}(\hat{\theta}_K^{LS})\right] \leq \frac{4\sigma}{\sqrt{n}} \sqrt{2 \log(2d)}$$

*and for any $\delta > 0$, with probability at least $1 - \delta$, it holds that*

$$\mathrm{MSE}(\hat{\theta}_K^{LS}) \leq \sigma \sqrt{\frac{32 \log\left(2d/\delta\right)}{n}}.$$

# $\ell_0$ constrained least squares

$\ell_0$ pseudonorm: cardinality of the set of non zero coordinates of a vector. A vector with small $\ell_0$ norm is called sparse. For any $\theta \in \mathbb{R}^d$,

$$\|\theta\|_0 = \sum_{i=1}^{d} \mathbb{I}(\theta_j \neq 0)$$

$$\mathrm{supp}(\theta) = \{j \in \{1, \dots, d\}\,,\ \theta_j \neq 0\}\,,$$

$\|\theta\|_0 = \mathrm{card}(\mathrm{supp}(\theta))$ and for any $k = 1, \dots, d$, $B_0(k)$ denotes the set of $k$-sparse vectors.

## Theorem (3.3.4)

*For any $k \in \mathbb{N}^*$, $k \leq d/2$, let $K = B_0(k)$ and assume that model* (LM) *holds with $\epsilon \sim \mathrm{subG}(\sigma^2)$ and $\theta^* \in K$. Then, for any $\delta > 0$, with probability $1 - \delta$, it holds*

$$\mathrm{MSE}(\hat{\theta}_K^{LS}) \leq \frac{32\sigma^2}{n}\left(2k\log\left(\frac{ed}{2k}\right) + 2k\log(6) + \log(1/\delta)\right).$$

*Furthermore, we have*

$$\mathbb{E}\left[\mathrm{MSE}(\hat{\theta}_K^{LS})\right] \leq \frac{32\sigma^2}{n}\left(1 + 2k\log\left(\frac{ed}{2k}\right) + 2k\log(6)\right)$$

# Adaptivity

Require the knowledge of properties of the unknown $\theta^*$.

**Sub-gaussian sequence model:** $y = \theta^* + \xi \in \mathbb{R}^d$, where $\xi \sim \mathrm{subG}(\sigma^2/n)$. For any $\delta > 0$, with probability at least $1 - \delta$

$$\max_{1 \le i \le d} |\xi_i| \le \sigma \sqrt{\frac{2 \log(2d/\delta)}{n}} = \tau.$$

If $|y_j| \gg \tau$ for some $j$, then it must correspond to $\theta_j^* \neq 0$. If $|y_j| \le \tau$, then $|\theta_j^*| \le |y_j| + |\xi_j| \le 2\tau$ with high probability.

Hard-thresholding estimator:

$$\hat{\theta}_j^{HT} = y_j \mathbb{I}(|y_j| \ge 2\tau), \quad j = 1, \dots, d.$$

Conditioning on the event $\mathcal{A} = \{\max_i |\xi_i| \le \tau\}$, we have for all $j$, $|y_j| \ge 2\tau \Rightarrow |\theta_j^*| \ge \tau$ and $|y_j| \le 2\tau \Rightarrow |\theta_j^*| \le 3\tau$ and

$$\|\hat{\theta}^{RT} - \theta^*\|^2 \le \frac{32 \|\theta\|_0 \sigma^2 \log(2d/\delta)}{n}.$$

# Penalized least squares estimators

It turns out that

$$\hat{\theta}^{HT} = \arg\min_{\theta \in \mathbb{R}^d} \|y - \theta\|^2 + 4\tau^2 \|\theta\|_0.$$

Under model (LM), we set, for any $\lambda \geq 0$,

$$\hat{\theta}^{\ell_0} \in \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{2n} \|\mathbb{X}\theta - Y\|^2 + \lambda \|\theta\|_0$$

$$\hat{\theta}^{\ell_1} \in \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{2n} \|\mathbb{X}\theta - Y\|^2 + \lambda \|\theta\|_1$$

**Theorem (3.4.1)**

*Assume that model* (LM) *holds with* $\epsilon \sim \mathrm{subG}(\sigma^2)$ *then choosing* $\lambda = 8\log(6)\sigma^2/n + 16\sigma^2\log(ed)/n$, *we have for any* $\delta > 0$ *with probability at least* $1-\delta$,

$$\mathrm{MSE}(\hat{\theta}^{\ell_0}) \leq \frac{32\sigma^2\left(2\|\theta^*\|_0\left(\log(6)+\log(ed)\right)+\log(1/\delta)+\log(2)\right)}{n}$$

**Theorem (3.4.2)**

*Assume that model* (LM) *holds with* $\epsilon \sim \mathrm{subG}(\sigma^2)$. *Moreover assume that the columns of* $\mathbb{X}$ *have norm at most* $\sqrt{n}$. *Then, for any* $\delta > 0$, *choosing* $\lambda = \sigma/\sqrt{n}\left(\sqrt{2\log(2d)} + \sqrt{2\log(1/\delta)}\right)$, *we have for any* $\delta > 0$ *with probability at least* $1 - \delta$,

$$\mathrm{MSE}(\hat{\theta}^{\ell_1}) \leq \frac{4\|\theta^*\|_1 \sigma}{\sqrt{n}} \left(\sqrt{2\log(2d)} + \sqrt{2\log(1/\delta)}\right).$$

# Incoherence, random matrices and cone condition

## Definition (3.5.1)

A matrix $\mathbb{X} \in \mathbb{R}^{n \times d}$ is said to have incoherence $k \in \mathbb{N}^*$, if

$$\left\| \frac{\mathbb{X}^T \mathbb{X}}{n} - I_d \right\|_\infty \leq \frac{1}{32k},$$

where $\| \cdot \|_\infty$ denotes the largest absolute value of a matrix.

## Proposition (3.5.1)

*Let $\mathbb{A} \in \mathbb{R}^{n \times d}$ be a random matrix which entries are independent Rademacher variables ($\pm 1$ with probability $1/2$). Then, for any $\delta > 0$, if $n \geq 2^{11} k^2 \log(1/\delta) + 2^{13} k^2 \log(d)$, with probability $1 - \delta$ over the random draw of its entries, $\mathbb{A}$ has incoherence $k$.*

For any $\theta \in \mathbb{R}^d$, $S \subset \{1, \ldots, d\}$, $\theta_S \in \mathbb{R}^d$ is the vector which entries agree with those of $\theta$ on $S$ the others beeing 0.

## Lemma (3.5.1)

*For any $k \leq d$ and $\mathbb{X}$ having incoherence $k$, any $S$ with $|S| \leq k$ and any $\theta \in \mathbb{R}^d$ satisfying the cone condition: $\|\theta_{S^c}\|_1 \leq 3\|\theta_S\|_1$, we have $\|\theta\|_2^2 \leq 2\frac{\|\mathbb{X}\theta\|_2^2}{n}$.*

# Fast rate for the Lasso estimator

**Theorem (3.5.1)**

*For $n \neq 2$, assume that model LM holds with $\epsilon \sim \mathrm{subG}(\sigma^2)$. Assume that $\|\theta^*\|_0 \leq k$ and that $\mathbb{X}$ has incoherence $k$. Then, for any $\delta > 0$, the Lasso estimator $\hat{\theta}^{\ell_1}$ with $\lambda = 8\sigma/n(\sqrt{\log(2d)} + \sqrt{\log(1/\delta)})$ satisfies with probability $1 - \delta$*

$$\mathrm{MSE}(\hat{\theta}^{\ell_1}) \leq (2^{12}) \frac{k\sigma^2 \log(2d/\delta)}{n}$$

$$\|\hat{\theta}^{\ell_1} - \theta^*\|_2^2 \leq (2^{13}) \frac{k\sigma^2 \log(2d/\delta)}{n}$$

The signal to be recovered is $\theta^* \in \mathbb{R}^d*$ which is unknown and assumed to be sparse, that is $\|\theta^*\|_0 = k < d$. $\mathbb{X} \in \mathbb{R}_{n \times d}$ is a sensing matrix which will result in the following measurements:

$$\mathbb{X}\theta^* = y \tag{3.10}$$

How many measurements are required so that $\theta^*$ can be infered accurately only from the knowledge of $y$ and $\mathbb{X}$?

We introduce the estimator

$$\hat{\theta}_{CS}^{\ell_0} \in \min_{\theta \in \mathbb{R}_d} \quad \|\theta\|_0 \quad \text{s.t.} \quad \mathbb{X}\theta = y. \tag{3.11}$$

under mild assumption on the sensing matrix $\mathbb{X}$, this estimator deterministically recovers the unknown signal $\theta^*$.

**Proposition (3.6.1)**

*Given $k \in \mathbb{N}$, $k \leq d/2$, assume that $\|\theta^*\|_0 \leq k$, and assume that for any $S$, $|S| \leq 2k$, that $\mathbb{X}_S$ has full column rank. Then, the solution of (3.11) is unique and is equal to $\theta^*$.*

We introduce an estimator.

$$\hat{\theta}^{\ell_1}_{CS} \in \min_{\theta \in \mathbb{R}_d} \quad \|\theta\|_1 \quad \text{s.t.} \quad \mathbb{X}\theta = y. \tag{3.12}$$

**Corollary (3.6.1)**

*Given $k \in \mathbb{N}$, $k \leq d$, and $\delta > 0$, assume that $\mathbb{X}$ is a Rademacher matrix with $n \geq 2^{11}k^2 \log(1/\delta) + 2^{13}k^2 \log(d)$. Assume furthermore that $\|\theta^*\|_0 \leq k$ in (3.10). Then with probability $1 - \delta$ over the random draw of $\mathbb{X}$, the solution of (3.12) is unique and is equal to $\theta^*$.*

# Recap on complexity for linear regression

| Least squares estimator | Mean squared error | Assumptions |
|---|---|---|
| Unconstrained/unpenalized | $\frac{\sigma^2 d}{n}$ | Design full column rank |
| $\ell_1$ constrained | $\frac{\sigma \log(d)}{\sqrt{n}}$ | $\|\theta^*\|_1 \leq 1$, $\|\mathbb{X}_j\|_2 \leq \sqrt{n}$ |
| $\ell_0$ constrained | $\frac{\sigma^2 k \log(d)}{n}$ | $\|\theta^*\|_0 \leq k$ |
| $\ell_1$ penalized | $\frac{\sigma \log(d)}{\sqrt{n}}$ | $\|\mathbb{X}_j\|_2 \leq \sqrt{n}$ |
| $\ell_0$ penalized | $\frac{\sigma^2 \|\theta^*\|_0 \log(d)}{n}$ | |
| $\ell_1$ penalized | $\frac{\sigma^2 k \log(d)}{n}$ | $\|\theta^*\|_0 \leq k$, $\mathbb{X}$ incoherence $k$ |

**General conclusion:**

- In high dimension, prior knowledge on $\theta^*$ is required to obtain meaningful bounds.
- For sparisity, $\ell_0$ pseudo norm has more favorable statistical properties than $\ell_1$ norm.
- Penalized estimators are adaptive to unknown properties of $\theta^*$, contrasting with constrained estimators.