



---

# Intervalles de confiance et tests d'hypothèses pour la régression en grande dimension

L'estimateur LASSO débiaisé

---

*Auteurs :*

Homer DURAND  
Loïs POURCHET

*Enseignant :*

Etienne ROQUAIN

*Projet réalisé dans le cadre du cours : Modèle Linéaire en Grande Dimension.*

15 janvier 2022

## Table des matières

|          |   |           |
|----------|---|-----------|
| <b>1</b> | <b>Introduction</b>                                     | <b>1</b>  |
| <b>2</b> | <b>Modèle</b>   | <b>1</b>  |
| <b>3</b> | <b>Résultat étudié</b>                                  | <b>2</b>  |
| 3.1      | Enonciation . . . . .                                   | 2         |
| 3.2      | Interprétation mathématique . . . . .                   | 3         |
| <b>4</b> | <b>Preuves</b>  | <b>3</b>  |
| 4.1      | Principal résultat à démontrer . . . . .                | 3         |
| 4.2      | Théorèmes auxiliaires et preuves . . . . .              | 3         |
| <b>5</b> | <b>Expérimentations pratiques</b>                       | <b>6</b>  |
| 5.1      | Adaptation de l'estimateur aux données . . . . .        | 7         |
| 5.2      | Tests d'hypothèse et intervalles de confiance . . . . . | 8         |
| 5.3      | Comparaison GAUSS-LASSO/DLASSO . . . . .                | 9         |
| <b>6</b> | <b>Conclusion</b>                                       | <b>10</b> |

## Table des figures

|   |  |   |
|---|--|---|
| 1 | Biais moyen et norme infini du biais de l'estimateur en fonction de la dimension des données . . . . . | 7 |
| 2 | Norme infini du biais en fonction du support de $\theta_0$ . . . . .                                   | 8 |
| 3 | Intervalle de confiance de $\hat{\theta}^u$ en comparaison avec $\theta_0$ . . . . .                   | 8 |
| 4 | Comparaison Gauss-LASSO et Debiased LASSO . . . . .  | 9 |

# 1 Introduction

Bien que les conditions nécessaires pour obtenir les bornes de convergence de l'estimateur LASSO soient essentiellement invérifiables, cette méthode d'estimation montre d'excellents résultats pratiques dans le cas de données sparses de grande dimension ce qui en fait une méthode de première importance pour la communauté du machine learning et des sciences des données en général. Mais comme cela est souligné par les auteurs de [Javanmard and Montanari \(2013\)](#), l'introduction d'un terme de régularisation dans le problème de minimisation résolu induit nécessairement un biais sur cet estimateur. Ils proposent donc une méthode permettant de compenser ce biais en ajoutant un terme proportionnel au sous gradient de la norme  $l_1$  sur l'estimateur LASSO. L'ordre du biais de ce nouvel estimateur est étudié en section (3.1).

Dans un second temps, les auteurs décrivent une procédure permettant d'effectuer des tests d'hypothèse multidimensionnels contrôlant le *Familywise Error Rate (FWER)* en proposant une méthode de calcul des p-valeurs. Nous n'étudions pas ici les propriétés théoriques relatives à cette procédure qui sont décrites dans [Javanmard and Montanari \(2013\)](#) mais nous l'avons testé sur des données numériques simulées.

## 2 Modèle

On a  $n$  paires de variables aléatoires i.i.d.  $(X_i, Y_i)_{1 \leq i \leq n}$  telles que :  $Y_i = \langle X_i, \theta_0 \rangle + \epsilon_i$  avec  $\theta_0$  et  $X_i \in \mathbb{R}^p$  ainsi que  $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$   
Sous forme vectorielle, le modèle de régression linéaire s'écrit :  $Y = X\theta + \varepsilon$  avec  $Y = (Y_1, \dots, Y_n)^T$ , la matrice de design  $X = (X_1, \dots, X_n)^T$  et  $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_{n \times n})$

**Définition 1.** L'estimateur  $\hat{\theta}^{MC}$  des moindres carrés correspond, par définition, à :

$$\hat{\theta}^{MC} \in \arg \min_{\theta \in \mathbb{R}^p} (\|Y - X\theta\|_2^2). \quad (1)$$

Lorsque  $p \gg n$ , l'estimateur  $\hat{\theta}^{MC}$  qui est non biaisé n'est pas défini puisque  $X$  n'est pas de plein rang. Par conséquent, cet estimateur  $\hat{\theta}^{MC}$  ne convient plus. On définit un nouvel estimateur nommé estimateur LASSO.

**Définition 2.** L'estimateur  $\hat{\theta}^n(\lambda)$  LASSO correspond à :

$$\hat{\theta}^n \in \arg \min_{\theta \in \mathbb{R}^p} \left( \frac{1}{2n} \|Y - X\theta\|_2^2 + \lambda \|\theta\|_1 \right) \quad (2)$$

Seulement, cet estimateur  $\hat{\theta}^n$  LASSO est biaisé. L'enjeu est donc de proposer une méthode pour débiaiser cet estimateur. Pour construire cet estimateur non biaisé, on ajoute un terme à cet estimateur. Ce terme supplémentaire dépend d'une matrice  $M$  que l'on définit grâce au paramètre de cohérence.

**Définition 3.** On se donne  $\mathbf{X} \in \mathbb{R}^{n \times p}$ ,  $M \in \mathbb{R}^{p \times p}$  et  $\hat{\Sigma} = \frac{X^T X}{n}$ . Le paramètre de cohérence généralisée noté  $\mu_*(X; M)$  est défini tel que :

$$\mu_*(X; M) \equiv |M\hat{\Sigma} - I|_\infty \quad (3)$$

On définit alors  $M_{\min}(X)$  telle que  $M_{\min}(X) = \arg \min_{M \in \mathbb{R}^{p \times p}} (\mu_*(X; M))$

Le paramètre de cohérence étant à présent défini, on peut définir à son tour l'estimateur  $\hat{\theta}^u(\lambda)$  LASSO débiaisé :

**Définition 4.** *L'estimateur  $\hat{\theta}^u(\lambda)$  LASSO débiaisé correspond à :*

$$\hat{\theta}^u = \hat{\theta}^n(\lambda) + \frac{1}{n}MX^T(Y - X\hat{\theta}^n(\lambda)) \quad (4)$$

### 3 Résultat étudié

Le résultat théorique au coeur de cette étude est le théorème 2.3 de l'article de (Javanmard and Montanari, 2013). En premier lieu, ce théorème sera énoncé dans la partie 3.1 puis les limites et une interprétation de ce théorème seront présentées dans la partie 3.2.

#### 3.1 Enonciation

Avant d'énoncer le théorème d'intérêt, il faut définir la condition de compatibilité faisant l'objet d'une hypothèse indispensable au résultat théorique en question.

**Définition 5.** *Soit une matrice symétrique  $\hat{\Sigma} \in \mathcal{M}(\mathbb{R})_{p \times p}$  et un ensemble  $S \subset [p]$ , la condition de compatibilité correspond à :*

$$\phi^2(\hat{\Sigma}, S) = \min \left\{ \frac{|S| \langle \theta, \hat{\Sigma} \theta \rangle}{\|\theta_S\|_1^2} : \theta \in \mathbb{R}^p, \|\theta_{S^c}\|_1 \leq 3\|\theta_S\|_1 \right\} \quad (5)$$

On dit que  $X$  vérifie la condition de compatibilité pour un ensemble  $S \subset [p]$  et une constante  $\phi_0$  si  $\phi_0 \leq \phi(\hat{\Sigma}, S)$  avec  $\hat{\Sigma} = \frac{X^T X}{n}$ .

#### Théorème 1.

Soient  $\mathbf{X} \in \mathbb{R}^{n \times p}$  une matrice de design et  $\hat{\theta}^* = \hat{\theta}^*(Y, \mathbf{X}; M, \lambda)$  un estimateur général débiaisé comme celui défini par l'équation 4. En posant  $Z = \frac{M\mathbf{X}^T W}{\sqrt{n}}$ , on a :

$$\sqrt{n}(\hat{\theta}^* - \theta_0) = Z + \Delta, \quad Z \sim \mathcal{N}(0, \sigma^2 M \hat{\Sigma} M^T), \quad \Delta = \sqrt{n}(M \hat{\Sigma} - I)(\theta_0 - \hat{\theta}^n). \quad (6)$$

De plus, supposons que :

- $X$  satisfait la condition de compatibilité pour un ensemble  $S = \text{supp}(\theta_0)$  avec comme cardinal  $|S| \leq s_0$  constant  $\phi_0$
- $X$  possède un paramètre de cohérence généralisé  $\mu_* = \mu_*(\mathbf{X}; M)$ .

En se donnant  $K \equiv \max_{1 \leq i \leq p} (\hat{\Sigma}_{ii})$  et  $\lambda = \sigma \sqrt{\frac{(c^2 \log(p))}{n}}$ , on a :

$$\mathbb{P} \left( \|\Delta\|_\infty \geq \frac{4c\mu_*\sigma s_0}{\phi_0^2} \sqrt{\log(p)} \right) \leq 2p^{-c_0}, \quad c_0 = \frac{c^2}{32K} - 1. \quad (7)$$

## 3.2 Interprétation mathématique

Ce théorème permet de contrôler le biais de l'estimateur  $\hat{\theta}^u$ , à savoir le terme  $\Delta$ , en décomposant l'erreur d'estimation en un terme de variance  $Z$  suivant une loi gaussienne centrée multivariée et un terme de biais  $\Delta$  dépendant de la matrice de design  $X$ . Le contrôle de  $\|\Delta\|_\infty$  dépend de plusieurs paramètres tels que la constante de compatibilité  $\phi_0$ , du paramètre de cohérence généralisé  $\mu_*$  et de la sparsité des données. Par ailleurs, le supremum du biais étant linéairement contrôlé par le support de  $\theta_0$ , cette estimateur sera d'autant plus performant que les données seront sparses. Répétition ? ("**sparsité des données et la dernière phrase ?**")

## 4 Preuves

### 4.1 Principal résultat à démontrer

Le théorème énoncé en partie 3 doit être, à présent, démontré. Dans la démonstration figurant dans l'article ([Javanmard and Montanari, 2013](#)), une étape est à clarifier en particulier. Elle correspond au théorème suivant :

**Théorème 2.** *Pour tout  $\lambda \geq 4\sigma\sqrt{\frac{2K\log(pe^{\frac{t^2}{2}})}{n}}$  :*

$$\mathbb{P}(\|\hat{\theta}_n - \theta_0\|_1 \geq 4\lambda\frac{s_0}{\phi_0^2}) \leq 2e^{-\frac{t^2}{2}} \quad (8)$$

### 4.2 Théorèmes auxiliaires et preuves

Pour que la démonstration soit claire et structurée, nous allons, en complément, énoncer et démontrer trois théorèmes auxiliaires qui permettent de déduire ce théorème 2. La démarche générale sous-jacente de cette preuve s'inspire fortement des preuves des propositions 6 et 7 du cours de ([Roquain, 2021](#)) et des preuves des résultats théoriques (le théorème 6.1 et le lemme 6.2) de l'ouvrage ([Bhlmann and van de Geer, 2011](#)). La démonstration est construite sur l'idée que si l'on détermine un bon événement  $\Omega$  de probabilité suffisamment grande (théorème 4) telle que le complémentaire de cet événement implique une majoration de  $\|\hat{\theta}_n - \theta\|_1$  (théorème 3) alors on prouve, par contraposée, le résultat énoncé dans le théorème 2.

**Théorème 3.** *Pour tout  $\lambda \geq 4\sigma\sqrt{\frac{2K\log(pe^{\frac{t^2}{2}})}{n}}$ , avec  $x = \sqrt{2\log(pe^{\frac{t^2}{2}})}$  sur l'événement  $\Omega_0 = \left\{ \left\| \frac{X^T \epsilon}{\sqrt{n\sigma\sqrt{K}}} \right\|_\infty \leq x \right\}$ , on a :*

$$\frac{1}{n} \|X(\hat{\theta}_n - \theta_0)\|_2^2 + \frac{3\lambda}{2} \|\hat{\theta}_{n,S^c} - \theta_{0,S^c}\|_1 \leq \frac{5\lambda}{2} \|\hat{\theta}_{n,S} - \theta_{0,S}\|_1 \quad (9)$$

**Nota Bene :** Ce théorème 3 correspond à une version adaptée et plus générale de la proposition 7 du cours de ([Roquain, 2021](#)).

Preuve du Théorème 3. Soit  $\lambda \geq 4\sigma\sqrt{\frac{K\log(pe\frac{t^2}{2})}{n}}$  donc  $\lambda \geq 4\sigma\sqrt{\frac{K}{n}}x$  avec  $x = \sqrt{\log(pe\frac{t^2}{2})}$

Par définition de l'estimateur LASSO :

$$\frac{1}{2n}\|Y - X\hat{\theta}_n\|_2^2 + \lambda\|\hat{\theta}_n\|_1 \leq \frac{1}{2n}\|Y - X\theta_0\|_2^2 + \lambda\|\theta_0\|_1$$

$$\frac{1}{n}\|Y - X\hat{\theta}_n\|_2^2 + 2\lambda\|\hat{\theta}_n\|_1 \leq \frac{1}{n}\|Y - X\theta_0\|_2^2 + 2\lambda\|\theta_0\|_1$$

Or  $Y = X\theta_0 + \epsilon$

$$\|Y - X\hat{\theta}_n\|_2^2 = \|X(\hat{\theta}_n - \theta_0)\|_2^2 + 2\langle X(\hat{\theta}_n - \theta_0), \epsilon \rangle + \|\epsilon\|_2^2$$

Donc

$$\frac{1}{n}\|X(\hat{\theta}_n - \theta_0)\|_2^2 \leq \frac{2}{n}|\langle X(\hat{\theta}_n - \theta_0), \epsilon \rangle| + 2\lambda(\|\theta_0\|_1 - \|\hat{\theta}_n\|_1)$$

En prenant  $K = \max_{1 \leq i \leq p}(\hat{\Sigma}_{ii})$  et  $\hat{\Sigma} = \frac{X^T X}{n}$

$$\frac{1}{n}\|X(\hat{\theta}_n - \theta_0)\|_2^2 \leq \frac{2\sigma\sqrt{K}}{\sqrt{n}}|\langle \frac{X^T \epsilon}{\sqrt{n}\sigma\sqrt{K}}, (\hat{\theta}_n - \theta_0) \rangle| + 2\lambda(\|\theta_0\|_1 - \|\hat{\theta}_n\|_1)$$

Sur l'évènement  $\Omega_0 = \left\{ \left\| \frac{X^T \epsilon}{\sqrt{n}\sigma\sqrt{K}} \right\|_\infty \leq x \right\}$ , il en découle :

$$\frac{1}{n}\|X(\hat{\theta}_n - \theta_0)\|_2^2 \leq \frac{2\sigma\sqrt{K}}{\sqrt{n}}x\|\hat{\theta}_n - \theta_0\|_1 + 2\lambda(\|\theta_0\|_1 - \|\hat{\theta}_n\|_1)$$

Or on a  $\|\hat{\theta}_n\|_1 = \|\hat{\theta}_{n,S}\|_1 + \|\hat{\theta}_{n,S^c}\|_1$  donc  $\|\hat{\theta}_n\|_1 \geq \|\theta_{0,S}\|_1 - \|\hat{\theta}_{n,S} - \theta_{0,S}\|_1 + \|\hat{\theta}_{n,S^c}\|_1$   
Puisque  $\theta_{0,S^c} = 0$ , on en déduit que  $\|\hat{\theta}_n\|_1 \geq \|\theta_0\|_1 - \|\hat{\theta}_{n,S} - \theta_{0,S}\|_1 + \|\hat{\theta}_{n,S^c} - \theta_{0,S^c}\|_1$

On obtient

$$\frac{1}{n}\|X(\hat{\theta}_n - \theta_0)\|_2^2 \leq \frac{2\sigma\sqrt{K}}{\sqrt{n}}x\|\hat{\theta}_n - \theta_0\|_1 + 2\lambda(\|\hat{\theta}_{n,S} - \theta_{0,S}\|_1 - \|\hat{\theta}_{n,S^c} - \theta_{0,S^c}\|_1)$$

De même, puisque  $\|\hat{\theta}_n - \theta_0\|_1 = \|\hat{\theta}_{n,S} - \theta_{0,S}\|_1 + \|\hat{\theta}_{n,S^c} - \theta_{0,S^c}\|_1$

$$\frac{1}{n}\|X(\hat{\theta}_n - \theta_0)\|_2^2 \leq (2\lambda + \frac{2\sigma\sqrt{K}}{\sqrt{n}}x)\|\hat{\theta}_{n,S} - \theta_{0,S}\|_1 - (2\lambda - \frac{2\sigma\sqrt{K}}{\sqrt{n}}x)\|\hat{\theta}_{n,S^c} - \theta_{0,S^c}\|_1$$

Puisque  $2\lambda \geq 2(\frac{4\sigma\sqrt{K}}{\sqrt{n}}x) \geq \frac{2\sigma\sqrt{K}}{\sqrt{n}}x$  Avec  $\frac{\lambda}{2} \geq \frac{2\sigma\sqrt{K}}{\sqrt{n}}x$ , on obtient :

$$\frac{1}{n}\|X(\hat{\theta}_n - \theta_0)\|_2^2 + \frac{3\lambda}{2}\|\hat{\theta}_{n,S^c} - \theta_{0,S^c}\|_1 \leq \frac{5\lambda}{2}\|\hat{\theta}_{n,S} - \theta_{0,S}\|_1$$

□

**Théorème 4.** Pour tout  $\lambda \geq 4\sigma\sqrt{\frac{2K\log(pe\frac{t^2}{2})}{n}}$ , sur l'événement  $\Omega_0 = \left\{ \left\| \frac{X^T\epsilon}{\sqrt{n}\sigma\sqrt{K}} \right\|_\infty \leq x \right\}$ , on a :

$$\|\hat{\theta}_n - \theta_0\|_1 \leq 4\lambda \frac{s_0}{\phi_0^2} \quad (10)$$

**Nota Bene :** Ce théorème 4 correspond à la majoration du biais en norme 1 sur le bon évènement en question du théorème 4 du cours de (Roquain, 2021).

*Preuve du Théorème 4.* Tout d'abord :

$$\frac{1}{n} \|X(\hat{\theta}_n - \theta_0)\|_2^2 + \frac{3\lambda}{2} \|\hat{\theta}_n - \theta_0\|_1 = \frac{1}{n} \|X(\hat{\theta}_n - \theta_0)\|_2^2 + \frac{3\lambda}{2} \|\hat{\theta}_{n,S^c} - \theta_{0,S^c}\|_1 + \frac{3\lambda}{2} \|\hat{\theta}_{n,S} - \theta_{0,S}\|_1$$

Or, d'après le théorème 1, avec  $\frac{\lambda}{2} \geq \frac{2\sigma\sqrt{K}}{\sqrt{n}}x$ , on a :

$$\frac{1}{n} \|X(\hat{\theta}_n - \theta_0)\|_2^2 + \frac{3\lambda}{2} \|\hat{\theta}_{n,S^c} - \theta_{0,S^c}\|_1 \leq \frac{5\lambda}{2} \|\hat{\theta}_{n,S} - \theta_{0,S}\|_1$$

Par conséquent :

$$\frac{1}{n} \|X(\hat{\theta}_n - \theta_0)\|_2^2 + \frac{3\lambda}{2} \|\hat{\theta}_n - \theta_0\|_1 \leq \frac{5\lambda}{2} \|\hat{\theta}_{n,S} - \theta_{0,S}\|_1 + \frac{3\lambda}{2} \|\hat{\theta}_{n,S} - \theta_{0,S}\|_1 = 4\lambda \|\hat{\theta}_{n,S} - \theta_{0,S}\|_1$$

Or, d'après le théorème 2, on remarque que :

$$\|\hat{\theta}_{n,S^c} - \theta_{0,S^c}\|_1 \leq \frac{5}{3} \|\hat{\theta}_{n,S} - \theta_{0,S}\|_1 \leq \frac{9}{3} \|\hat{\theta}_{n,S} - \theta_{0,S}\|_1 = 3 \|\hat{\theta}_{n,S} - \theta_{0,S}\|_1$$

donc que  $\hat{\theta}_n - \theta_0$  appartient au cône  $\{\|u_{S^c}\|_1 \leq \|u_S\|_1\}$  sur l'événement  $\Omega_0$ . Sur cet évènement, on a donc par définition de la condition de compatibilité :

$$\|\hat{\theta}_{n,S} - \theta_{0,S}\|_1 \leq \frac{\sqrt{s_0} \|X(\hat{\theta}_n - \theta_0)\|_2}{\sqrt{n}\phi_0}$$

On en déduit :

$$\frac{1}{n} \|X(\hat{\theta}_n - \theta_0)\|_2^2 + \frac{3\lambda}{2} \|\hat{\theta}_n - \theta_0\|_1 \leq 4\lambda \frac{\sqrt{s_0} \|X(\hat{\theta}_n - \theta_0)\|_2}{\sqrt{n}\phi_0}$$

Puisque que  $\forall (u, v) \in \mathbb{R}^2, uv \leq \frac{1}{2}(u^2 + v^2)$ , on a également :

$$4\lambda \frac{\sqrt{s_0} \|X(\hat{\theta}_n - \theta_0)\|_2}{\sqrt{n}\phi_0} \leq \frac{1}{2} \left( \left( \frac{\sqrt{2}}{\sqrt{n}} \|X(\hat{\theta}_n - \theta_0)\|_2 \right)^2 + \left( 2\sqrt{2} \frac{\lambda\sqrt{s_0}}{\phi_0} \right)^2 \right) = \frac{1}{n} \|X(\hat{\theta}_n - \theta_0)\|_2^2 + 4\lambda^2 \frac{s_0}{\phi_0^2}$$

On en conclut que :

$$\|\hat{\theta}_n - \theta_0\|_1 \leq \frac{8}{3} \lambda \frac{s_0}{\phi_0^2} \leq 4\lambda \frac{s_0}{\phi_0^2}$$

□

**Théorème 5.** On pose :  $\Omega_0 = \left\{ \left\| \frac{X^T \epsilon}{\sqrt{n\sigma\sqrt{K}}} \right\|_\infty \leq x \right\}$

On montre que :

$$\mathbb{P}(\Omega_0^c) \leq 2e^{-\frac{t^2}{2}} \quad (11)$$

**Nota Bene :** Ce théorème 4 correspond à la minoration de la probabilité de l'évènement d'intérêt à l'instar du théorème 4 du cours de (Roquain, 2021).

*Preuve du Théorème 5.* Premièrement,  $\Omega_0^c = \left\{ \left\| \frac{X^T \epsilon}{\sqrt{n\sigma\sqrt{K}}} \right\|_\infty \geq x \right\}$

$$\text{Donc } \Omega_0^c = \left\{ \max_{1 \leq i \leq p} \left| \frac{X_i^T \epsilon}{\sqrt{n\sigma\sqrt{K}}} \right| \geq x \right\} \subset \left\{ \exists 1 \leq i \leq p \left| \frac{X_i^T \epsilon}{\sqrt{n\sigma\sqrt{\hat{\Sigma}_{ii}}}} \right| \geq x \right\} \subset \bigcup_{i=1}^p \left\{ \left| \frac{X_i^T \epsilon}{\sqrt{n\sigma\sqrt{\hat{\Sigma}_{ii}}}} \right| \geq x \right\}$$

$$\text{Ainsi } \mathbb{P}(\Omega_0^c) \leq \mathbb{P}\left(\bigcup_{i=1}^p \left\{ \left| \frac{X_i^T \epsilon}{\sqrt{n\sigma\sqrt{\hat{\Sigma}_{ii}}}} \right| \geq x \right\}\right) \leq 2p\mathbb{P}\left(\frac{X_i^T \epsilon}{\sqrt{n\sigma\sqrt{\hat{\Sigma}_{ii}}}} \geq x\right)$$

$$\text{Or } \frac{X_i^T \epsilon}{\sqrt{n\sigma\sqrt{\hat{\Sigma}_{ii}}}} \sim \mathcal{N}(0, 1) \text{ donc } \mathbb{P}(\Omega_0^c) \leq 2pe^{-\frac{x^2}{2}}$$

$$\text{Alors } \mathbb{P}(\Omega_0^c) \leq 2pe^{-\frac{x^2}{2}} \text{ si } \sqrt{2\pi}x \geq 1 \text{ ce qui est vrai si } p \geq 2.$$

$$\text{Enfin, } \mathbb{P}(\Omega_0^c) \leq 2e^{-\frac{t^2}{2}} \text{ puisque } x = \sqrt{2\log(pe^{\frac{t^2}{2}})} = \sqrt{t^2 + 2\log(p)}$$

□

*Preuve du Théorème 2.* On va utiliser les théorèmes démontrés précédemment.

$$\text{D'après la contraposée du théorème 4, pour tout } \lambda \geq 4\sigma\sqrt{\frac{2K\log(pe^{\frac{t^2}{2}})}{n}} : \|\hat{\theta}_n - \theta_0\|_1 \geq 4\lambda\frac{s_0}{\phi_0^2} \implies \Omega_0^c$$

$$\text{Le théorème 3 donne ensuite : } \mathbb{P}(\|\hat{\theta}_n - \theta_0\|_1 \geq 4\lambda\frac{s_0}{\phi_0^2}) \leq \mathbb{P}(\Omega_0^c) \leq 2e^{-\frac{t^2}{2}}$$

□

## 5 Expérimentations pratiques

Nous analysons dans cette section les résultats pratiques obtenus concernant l'estimateur  $\hat{\theta}^u$  calculé par l'algorithme (1). Les tests effectués dans cette section sont fait sur des données simulées telles que

$$Y = X\theta_0 + \epsilon$$

avec  $X \sim \mathcal{N}(0, \Sigma)$  et  $\epsilon \sim \mathcal{N}(0, 1)$ . La matrice de covariance  $\Sigma$  est, à moins que cela ne soit indiqué, la matrice identité de dimension  $p \times p$ . La matrice  $X$  est toujours de dimension  $n \times p$  avec  $n = 100$  et  $p = 80$  à moins que cela ne soit indiqué. Nous sommes conscients que nous ne sommes pas ici dans un cas de très grande dimension, mais nous sommes limités par le coût de calcul de l'algorithme qui eplose rapidement dans sa version naïve implémentée. Le vecteur à estimer  $\theta_0$  est de support  $s_0 = 5$  à moins que cela ne soit indiqué et ses valeurs non-nulles sont toujours les  $s_0$  premières valeurs. Nous sélectionnons les paramètres  $\mu$  et  $\lambda$  comme décrit dans Javanmard and Montanari (2013), à savoir  $\mu = 2\sqrt{\log(p)/n}$  et  $\lambda = 4\hat{\sigma}\sqrt{\log(p)/n}$  avec  $\hat{\sigma}$  l'estimation du bruit par *scaled LASSO* décrit dans Sun and Zhang (2012) avec  $\tilde{\lambda} = 10\sqrt{2\log(p)/n}$ .



---

**Algorithm 1** Estimateur débiaisé de  $\theta_0$  pour régression linéaire en grande dimension

---

**Require:** Vecteur de mesure  $y$ , matrice de design  $X$  et paramètres  $\lambda, \mu$

Soit  $\hat{\theta}^n$  l'estimateur LASSO

$$\hat{\Sigma} = (X^T X)/n$$

**for**  $i = 1, \dots, p$  **do**

Soit  $m_i$  une solution de

$$\begin{aligned} & \text{minimize} && m^T \hat{\Sigma} m \\ & \text{s.c.} && \|\hat{\Sigma} m - e_i\|_\infty \leq \mu \end{aligned}$$

où  $e_i \in \mathbb{R}^p$  est le vecteur valant 1 à la  $i$ -ème position et 0 partout ailleurs.

**end for**

Soit  $M = (m_1, \dots, m_p)^T$ . Si un des problème précédent n'est pas résoluble alors on prend  $M = I_{p \times p}$ .

L'estimateur  $\hat{\theta}^u$  de la façon suivante :

$$\hat{\theta}^u = \hat{\theta}^n(\lambda) + (1/n)MX^T(Y - X\hat{\theta}^n(\lambda))$$

---

## 5.1 Adaptation de l'estimateur aux données

Nous nous intéressons dans un premier temps à la capacité de cet estimateur à s'adapter à la dimension des données. Comme cela est montré dans le théorème (1), le terme de biais évolue avec le logarithme de la dimension de la matrice de design. Pour cela, nous générons des données de dimension  $p$  de plus en plus grande et observons l'évolution du biais de l'estimateur. Nous fixons le nombre d'observations à  $n = 100$  et faisons évoluer la dimension  $p \in [25, 200]$  en considérant que, pour  $0 \leq p \leq 50$ , nous sommes en petite dimension et que, au-delà, nous sommes en grande dimension.

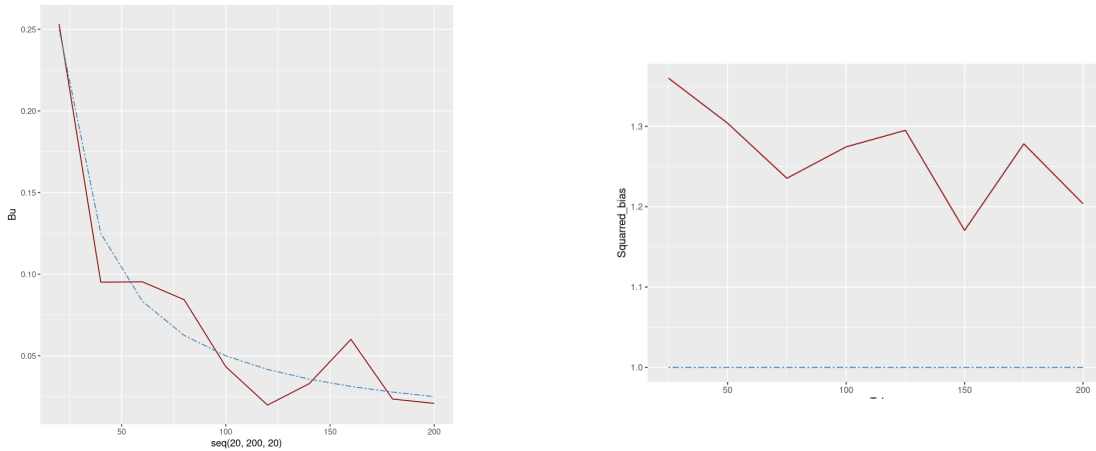


FIGURE 1 – Biais moyen et norme infini du biais de l'estimateur en fonction de la dimension des données

Notre estimation du biais moyen semble diminuer à mesure que la dimension augmente, en revanche, la norme infini du biais semble, elle, relativement stable (voir figure (1)).

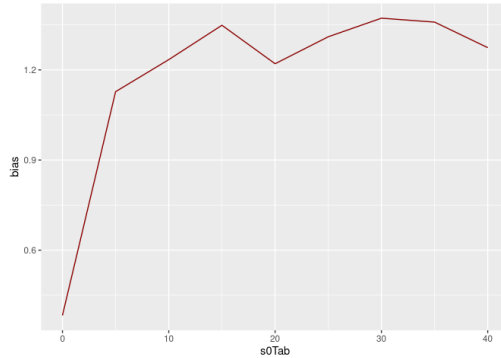


FIGURE 2 – Norme infini du biais en fonction du support de  $\theta_0$

Le théorème (1) montre que la norme infini du biais est majorée linéairement par le support de  $\theta_0$ . Comme on l’observe sur la figure (2), la norme infini du biais est dans un régime sous linéaire et semble plutôt évoluer comme le logarithme du support de  $\theta_0$  dans ce cas-ci, ce qui pourrait venir du fait que la matrice de design est générée à partir d’une simple matrice identité comme matrice de covariance.

## 5.2 Tests d’hypothèse et intervalles de confiance

Dans un second temps, nous avons évalué la procédure permettant de déterminer un intervalle de confiance sur l’estimateur et celle permettant d’effectuer les tests d’hypothèse proposés dans l’article, à savoir  $H_{0,i} : \hat{\theta}_i^u = 0$  contre  $H_{1,i} : \hat{\theta}_i^u \neq 0$ . Cependant, comme on peut le voir sur la figure (3), l’intervalle de confiance estimé par la procédure décrite dans l’article ne semble pas fonctionner dans notre cas et de plus le test d’hypothèse au niveau  $\alpha = 0.05$  ne rejette pas une des hypothèses de test sur cinq à tort, ce qui semble relativement élevé puisque la procédure contrôle le FWER au niveau  $\alpha$ . Nous envisageons que cela puisse venir du choix du paramètre  $\mu$  que nous avons choisi en accord avec les données de l’article.

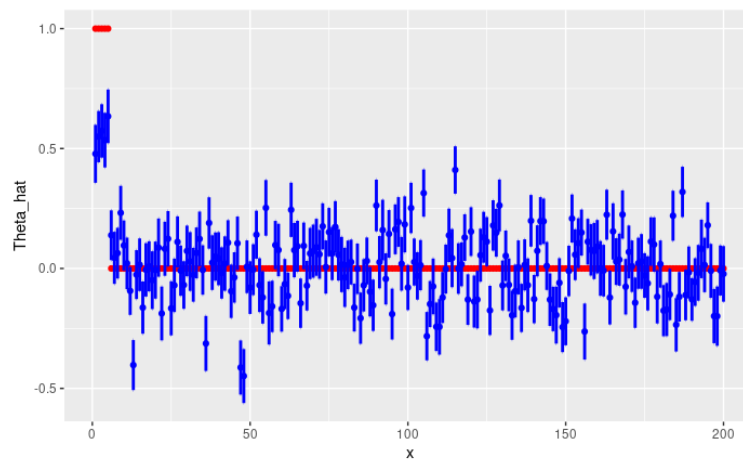


FIGURE 3 – Intervalle de confiance de  $\hat{\theta}^u$  (bleue) en comparaison avec  $\theta_0$  (rouge).

### 5.3 Comparaison GAUSS-LASSO/DLASSO

Nous proposons ici de comparer l'estimateur Gauss-LASSO étudié dans le cours [Roquain \(2021\)](#) à l'estimateur débiaisé étudié dans ce rapport. L'estimateur Gauss-LASSO consiste simplement à effectuer une estimation des moindres carrés sur le modèle sélectionné par le LASSO. Les données utilisées sont une concaténation de données de la base de Haar et de la base cosinus, de dimension  $n = 256$  et  $p = 200$ .

Comme on le voit sur la figure (4), il semblerait que le LASSO débiaisé compense le fait de diminuer le biais de l'estimateur en augmentant sa variance.

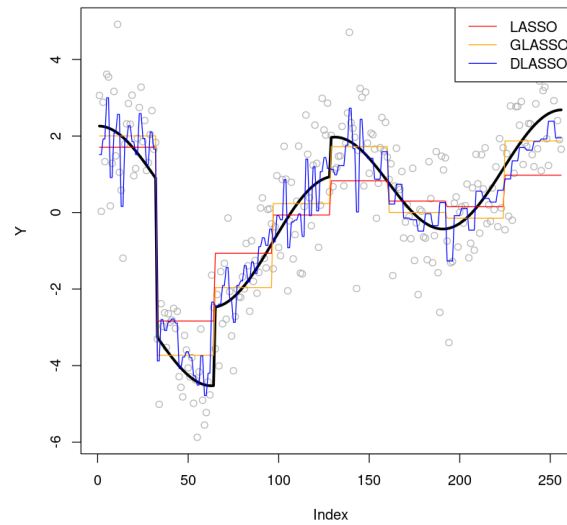


FIGURE 4 – Comparaison Gauss-LASSO et Debiased LASSO pour les données du TP2 du cours [Roquain \(2021\)](#)

## 6 Conclusion

Nous avons dans ce rapport étudié la procédure de compensation du biais de l'estimateur LASSO proposé par les auteurs de [Javanmard and Montanari \(2013\)](#). Dans un premier temps nous avons montré certains résultats théoriques importants de cet estimateur permettant de borner son biais. Nous avons vu que cette borne dépendait de la dimension des données mais également de leur sparsité et de leur constante de compatibilité. Par ailleurs, nous avons vu que cette procédure permettait également de fournir un intervalle de confiance asymptotique sur  $\hat{\theta}^u$  et de calculer les p-valeurs associées au test d'hypothèse, contrôlant ainsi le FWER à un niveau  $\alpha$ .

En revanche, cette méthode de compensation du biais implique un surcoût en terme de temps de calcul qui peut vite devenir très important pour des données de grande dimension. En effet l'implémentation naïve de la procédure est de l'ordre de  $O(p^2)$  ce qui devient vite impraticable lorsque  $p$  augmente. Nous avons donc dû nous restreindre dans nos tests à des données de dimension moyenne avec  $p$  du même ordre que  $n$  ce qui n'est pas un cadre extrême de grande dimension comme on peut en rencontrer lors d'études biologiques par exemple. Notons aussi que le stockage de la matrice  $M$  peut poser problème pour ce type de données.

Il pourrait être intéressant d'approfondir en cherchant une méthode d'estimation de  $M$  qui soit moins coûteuse en temps de calcul et en espace de stockage afin que la méthode soit utilisable en très grande dimension.

## Références

- Bhlmann, P. and van de Geer, S. (2011) *Statistics for High-Dimensional Data : Methods, Theory and Applications*. Springer Publishing Company, Incorporated, 1st edn.
- Javanmard, A. and Montanari, A. (2013) Confidence intervals and hypothesis testing for high-dimensional statistical models. **26**. URL : <https://proceedings.neurips.cc/paper/2013/file/b73dfe25b4b8714c029b37a6ad3006fa-Paper.pdf>.
- Roquain, E. (2021) Modèle linéaire en grande dimension : Chapitre 2 - Estimation dans le modèle linéaire gaussien. <https://etienneroquain-81.websself.net/home>.
- Sun, T. and Zhang, C.-H. (2012) Scaled sparse linear regression.