Quelques questions de sélection de variables autour de l'estimateur LASSO

Mohamed Hebiri

Université Paris Diderot - Paris 7
UFR de Mathématiques

Thèse

pour obtenir le titre de

Docteur de l'université Paris Diderot - Paris 7

Spécialité : Mathématiques Appliquées

présentée par

Mohamed HEBIRI

Quelques Questions De Sélection De Variables Autour De L'Estimateur LASSO

Directeur de thèse : **Nicolas VAYATIS**

Soutenue publiquement le 30 Juin 2009, devant le jury composé de :

M.	**Francis**	**BACH**	INRIA	(Rapporteur)
M.	**Stéphane**	**CANU**	INSA – Rouen	
M.	**Arnak**	**DALALYAN**	CERTIS – Ecole des Ponts	
Mme	**Dominique**	**PICARD**	Université Paris Diderot – Paris 7	
M.	**Alexandre**	**TSYBAKOV**	CREST – ENSAE	
Mme	**Sara**	**VAN DE GEER**	ETH – Zürich	(Rapporteur)
M.	**Nicolas**	**VAYATIS**	ENS – Cachan	

Remerciements

Pour exprimer ma gratitude envers ceux qui m'ont aidé et encouragé durant ces quelques années, il m'aurait fallu plus que ces quelques lignes. C'est à regret, mais pour ne pas abuser de votre patience, que je me restreindrai.

Je remercie tout d'abord Nicolas Vayatis pour son encadrement et son soutien, en particulier dans les moments difficiles. Il a su me faire profiter de son expérience scientifique avec bienveillance.

Francis Bach et Sara van de Geer ont aimablement accepté d'être les rapporteurs de cette thèse. C'est pour moi un grand honneur d'être évalué par d'aussi remarquables chercheurs.

J'aimerais aussi faire part de ma fierté de compter Alexandre Tsybakov parmi les membres du jury. Ses cours, travaux et conseils m'ont continuellement guidés depuis la maîtrise. J'ai une grande estime pour ses qualités scientifiques et humaines.

Un autre membre du jury, qui est aussi un ami, Arnak Dalalyan, a toujours été présent pour me conseiller et m'éclairer. Je l'en remercie de tout coeur et espère que notre complicité durera.

Je suis honoré et me réjouis de la participation au jury de ma thèse de Dominique Picard, qui suit assidûment mon cheminement depuis le master. Je remercie chaleureusement Stéphane Canu d'avoir lui aussi accepté d'être dans ce jury. Ses qualités pédagogiques et sa gentillesse m'ont touché lors de nos trop rares rencontres.

Christophe Chesneau et Pierre Alquier sont deux collègues et amis qui, au moyen d'échanges, de réflexions et parfois d'affrontements, m'ont permis de développer mes modestes capacités de chercheur. De même, travailler avec Michèle Sebag a été aussi agréable qu'enrichissant. Nos longues discussions ont été sources d'inspiration. Je remercie amicalement Gilles Stoltz, sur qui j'ai toujours pu compter, pour le soutien qu'il m'a apporté ces dernières années.

Je tiens à exprimer ma gratitude envers les professeurs Thierry Berkover, Frédéric Praslon, Claudine Degand, René Guitart, avec lesquels il a été très plaisant d'enseigner. Merci à Séverine Cholewa, Romain Starck, Anne-Laure Jachiet, Mohammed Mikou et Martin Hils, pédagogues d'exception n'hésitant pas à s'impliquer et à se remettre en question, avec qui j'ai partagé d'agréables moments à Marne-la-Vallée et à Paris 7. Bien entendu, je n'oublie pas tous mes étudiants sans qui je n'aurais pas pu faire cette expérience et qui m'ont fait aimer l'enseignement.

Un grand merci à Zaïd Harchaoui, Abass Sagna, Karim Lounici, Adrien Saumart, Thomas Willer, Nicolas Verzelen, Erwan Le Pennec, Xavier Gendre, Guillaume Lecué, Etienne Roquain et Fanny Willer, Sylvain Delattre, Gérard Biau et Sophie Dédé. Amis de et pour la science, parfois philosophes, je leur dois de nombreux éclairements. Je remercie également Cristina Butucea, Stéphane Boucheron, Marc Hoffmann, Florence Merlevède et Gérard Kerkyacharian pour leurs suggestions et leur amabilité lors de ces années de thèse.

Je salue les thésards de Chevaleret et en particulier ceux occupant mes bureaux 5C9 et 5B1 : Luca, Julien, Christophe, François, Maxime, Pierre, Hubert, Idris, Marc, Thomas et tous les autres (ils sont trop nombreux).

Je remercie aussi l'ensemble de l'équipe administrative du LPMA, à commencer par Michèle Wasse, toujours à l'écoute et prête à nous aider, Jacques Portes, Pascal Chiettini

et Virginie Kuntzmann pour leur dévouement, Isabelle Mariage, Valérie Juvé et Véronique Carpentier pour leur efficacité.

Un grand merci au comité de relecture de cette thèse, en particulier à Sisi Yé qui a sacrifié quelques-unes de ses heures de sommeil, ainsi qu'à Joseph Salmon qui m'a héroïquement supporté. Ces remerciements ne seraient pas complets sans nommer ma collègue et amie Katia Meziani. Cette thèse et beaucoup d'autres choses n'auraient pas été les mêmes sans elle...

Enfin, je tiens à remercier mes proches, famille et amis, pour leur soutien, leur disponibilité, et pour m'avoir permis de décompresser lorsqu'il le fallait : Elodie ; mes frères et soeur Kaïs, Loubna et Hatem ; mes parents ; Katia, Baba et Imoux ; les familles Meziani, Muller, Perrin et Mery ; Sofiane, Ptit Moh, Sisi, Zaïd, Arno, Kamel la frontière, Bouzid, Khaled, Mihai, mes cousins Mohamed et Oussama, Jessy, Jérôme... bref, tout le monde...

TABLE DES MATIÈRES

Chapitre 1

Synthèse des travaux

Dans ce chapitre d'ouverture, nous présentons une méthode de sélection de variables pour la régression : le LASSO, et nous donnons une vue d'ensemble du travail réalisé.

1.1 Cadre général

Nous présentons ici les notations pour le modèle de régression linéaire. Puis, nous introduisons la procédure LASSO, qui permet de traiter certains problèmes d'inférence dans ce cadre.

1.1.1 Modèle de régression linéaire

Considérons que les données sont générées selon le modèle de régression linéaire (1.1) ; i.e., les données observées sont des couples $(x_i, y_i)_{i=1,\dots,n}$, où $y_i \in \mathbb{R}$ est la réponse telle que :

$$y_i = x_i\beta^* + \varepsilon_i, \quad i = 1, \dots, n. \tag{1.1}$$

Dans cette expression, $x_i = (x_{i,1}, \dots, x_{i,p})$ est un vecteur déterministe de taille p où chaque composante correspond à une variable explicative pour l'étude de y_i. Le paramètre $\beta^* = (\beta_1^*, \dots, \beta_p^*)'$ désigne un vecteur de taille p, où la composante β_j^* du vecteur β^* traduit l'influence de la variable $x_{i,j}$ sur la réponse y_i. Dans le modèle (1.1), les variables aléatoires $\{\varepsilon_i\}_{i=1,\dots,n}$ sont indépendantes et identiquement distribuées (i.i.d.). On supposera par la suite que les ε_i sont des variables aléatoires gaussiennes centrées, de variance σ^2 connue.

Dans le cadre que nous considérons, la dimension p peut être plus grande que la taille de l'échantillon n. Ces problèmes, où $p \geq n$, sont dits de *grande dimension*. Il est utile de supposer dans ce cas que le paramètre β^* est *sparse* (*parcimonieux* en français), i.e. peu de composantes β_j^*, $j \in \{1, \dots, p\}$ sont différentes de zéro. Définissons l'*ensemble de sparsité* $\mathcal{A}^*$ comme

$$\mathcal{A}^* = \{j : \beta_j^* \neq 0\}. \tag{1.2}$$

On s'intéressera à l'estimation du paramètre inconnu β^*, par le biais d'outils capables d'extraire un nombre restreint de variables explicatives fournissant une description quasi-complète du modèle considéré. L'objectif est de sélectionner les variables $X_j = (x_{1,j}, \dots, x_{n,j})'$ dont les indices j sont dans $\mathcal{A}^*$. En sélectionnant un sous-ensemble de variables explicatives, l'estimateur produit des conclusions facilement interprétables, et donc exploitables en pratique. Les méthodes statistiques classiques telles que l'estimateur des moindres carrés ou l'estimateur ridge ne permettent pas d'apporter des résultats satisfaisants pour ce type d'études. Des procédures de sélection de variables sont une possibilité pour répondre à ce problème. On peut par exemple considérer des critères d'information classiques tels que les critères C_p, AIC ou encore BIC. Toutefois, ces critères sont peu utilisables en pratique. En effet, la complexité algorithmique de telles méthodes est telle qu'elles sont difficiles à implémenter, même pour des dimensions p modestes.

Dans le cadre de la grande dimension, nous proposons dans ce manuscrit d'étudier des estimateurs capables de sélectionner les variables pertinentes, tout en étant facilement implémentables en pratique. L'estimateur LASSO est la méthode de référence considérée dans cette thèse.

1.1.2 Estimation

Cette section 1.1.2 est dédiée à une première description de la procédure LASSO introduite par Tibshirani [135] (voir aussi le *Basis Pursuit De-Noising* de Chen, Donoho, et Saunders [43]). Cette méthode d'estimation est définie comme étant un minimiseur du critère des moindres carrés pénalisés par la norme ℓ_1 du vecteur β. On parle de pénalité ℓ_1 ou de pénalité *LASSO*, i.e.

$$\hat{\beta}^L(\lambda) \in \underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}} \|Y - X\beta\|_n^2 + \lambda\|\beta\|_1. \tag{1.3}$$

Dans l'expression (1.3), Y et X sont tels que $Y = (y_1, \ldots, y_n)'$ et $X = (x_1', \ldots, x_n')'$. Les normes $\|a\|_n^2$ et $\|b\|_q^q$ renvoient respectivement à la norme empirique dans $\mathbb{R}^n$ ($\|a\|_n^2 = \frac{1}{n}\sum_{i=1}^n a_i^2$) et à la norme ℓ_q ($\|b\|_q^q = \sum_{j=1}^p |b_j|^q$) avec $a \in \mathbb{R}^n$, $b \in \mathbb{R}^p$ et $q \in \mathbb{N}^*$. Le paramètre $\lambda \in \mathbb{R}_+$ est le paramètre de lissage, son rôle ici est majeur.

Remarquons que, dans un premier temps, une forme duale de l'estimateur LASSO a été obtenue par Osborne, Presnell, et Turlach [118] et Alquier [4]. Elle s'énonce de la façon suivante :

$$\tilde{\beta}^L(\lambda) \in \begin{cases} \operatorname{argmin}_{\beta\in\mathbb{R}^p} \|X\beta\|_2^2 \\ \\ \text{s.c. } \left\|\Xi^{-1}X'(Y - X\beta)\right\|_\infty \leq \frac{n\lambda}{2}, \end{cases} \tag{1.4}$$

où Ξ est la matrice diagonale de taille p, dont le j-ième coefficient diagonal vaut $\|X_j\|_n = \sqrt{\frac{1}{n}\sum_{i=1}^n x_{i,j}^2}$. L'estimateur LASSO (1.3) et sa forme duale (1.4) sont tels que $\hat{\beta}^L = \tilde{\beta}^L$, dès lors que la matrice $\Psi^n = \frac{X'X}{n}$ est inversible. Lorsque ceci n'est plus vrai, par exemple dans le cas où $p > n$, seule la relation $X\tilde{\beta}^L(\lambda) = X\hat{\beta}^L(\lambda)$ reste vérifiée. Par la forme duale (1.4), l'estimateur Lasso peut être vu comme le projeté de $\mathbf{0_p}$, le zéro de $\mathbb{R}^p$, sur la région de confiance de niveau $\eta = \eta(\lambda) \in]0, 1[: \left\{\beta \in \mathbb{R}^p : \left\|\Xi^{-1}X'(Y - X\beta)\right\|_\infty \leq \frac{n\lambda}{2}\right\}$ (cf. Alquier [4]).

Notons également que l'estimateur LASSO sélectionne d'autant plus de variables explicatives que λ est petit. Différents algorithmes existent pour approximer l'ensemble des solutions du problème LASSO pour λ variant dans $[0, \infty[$. En particulier, l'algorithme LARS introduit par Efron, Hastie, Johnstone, et Tibshirani [61], apporte une réponse à ce problème d'optimisation. Cet algorithme a l'avantage d'être très rapide et a accru l'intérêt porté à la méthode LASSO. Cet algorithme est détaillé dans l'annexe B.

L'estimateur LASSO a été largement étudié, ses propriétés théoriques et ses limites algorithmiques sont à présent bien déterminées. Nous pouvons citer entre autres les contributions de Bickel, Ritov, et Tsybakov [17], Bunea [28], Chen, Donoho, et Saunders [43], Efron, Hastie,

Johnstone, et Tibshirani [61], Knight et Fu [90], Pötscher et Leeb [123] et Zhao et Yu [164]. Plus de détails peuvent être trouvés dans le chapitre 2.

1.1.3 Quelques résultats théoriques autour du LASSO

De nombreux résultats théoriques ont été établis ces dernières années dans la littérature statistique. Ces résultats ont été obtenus au prix d'hypothèses plus ou moins contraignantes, notamment sur la matrice de Gram $\Psi^n = \frac{X'X}{n}$. Les hypothèses faites sur la matrice de Gram n'autorisent qu'une faible corrélation entre les variables. Nous détaillerons plus loin les trois principales hypothèses imposées à la matrice Ψ^n. Au chapitre 2, nous nous attarderons plus longuement sur les différentes hypothèses rencontrées dans la littérature.

Introduisons quelques notations. Soit $\mathcal{A}$ et $\mathcal{B}$ deux sous-ensembles d'indices dans $\{1, \dots, p\}$. Étant donnée la matrice de Gram Ψ^n, on note

$$\Psi^n_{\mathcal{A},\mathcal{B}} = \frac{X'_{\mathcal{A}} X_{\mathcal{B}}}{n},$$

la matrice de taille $|\mathcal{A}| \times |\mathcal{B}|$ (où $|A|$ est le cardinal de l'ensemble $A \subset \{1, \dots, p\}$), restriction de Ψ^n aux lignes dont les indices appartiennent à l'ensemble $\mathcal{A}$ et aux colonnes dont les indices appartiennent à l'ensemble $\mathcal{B}$. De plus, pour un vecteur $a \in \mathbb{R}^q$ avec $q \in \mathbb{N}^*$, introduisons le vecteur signe $\mathrm{Sgn}(a) = (\mathrm{sgn}(a_1), \dots, \mathrm{sgn}(a_q))'$ où la fonction $\mathrm{sgn}(\cdot)$ est telle que pour tout $b \in \mathbb{R}$,

$$\mathrm{sgn}(b) = \begin{cases} 1 & \text{si } b > 0, \\ 0 & \text{si } b = 0, \\ -1 & \text{si } b < 0. \end{cases} \tag{1.5}$$

Les trois hypothèses principales posées dans la littérature sont les suivantes :

Hypothèse 1.1. (CI) Condition d'Irreprésentabilité

Il existe une constante $C > 0$ telle que

$$\|\Psi^n_{(\mathcal{A}^*)^C,\mathcal{A}^*} (\Psi^n_{\mathcal{A}^*,\mathcal{A}^*})^{-1} \mathrm{Sgn}(\beta^*_{\mathcal{A}^*})\|_\infty \leq 1 - C,$$

où $\|\cdot\|_\infty$ renvoie à la norme sup dans $\mathbb{R}^p$.

Hypothèse 1.2. (CM) Cohérence Mutuelle

Il existe une constante $C > 0$ telle que

$$\max_{j \in \mathcal{A}^*} \max_{\substack{k \in \{1,\dots,p\} \\ k \neq j}} |\Psi^n_{j,k}| \leq \frac{C}{|\mathcal{A}^*|},$$

où $|\mathcal{A}^*|$ est également vu comme l'indice de sparsité de β^*.

Hypothèse 1.3. (VPR) Valeurs Propres Restreintes

Soit $\alpha \in \mathbb{R}^p$, un vecteur non nul tel que

$$\sum_{j \notin \mathcal{A}^*} |\alpha_j| \leq 3 \sum_{j \in \mathcal{A}^*} |\alpha_j| \,. \tag{1.6}$$

Pour tout α vérifiant (1.6)*, il existe une constante $\kappa_1 > 0$ telle que :*

$$\frac{\alpha' \Psi^n \alpha}{\sum_{j \in \mathcal{A}^*} \alpha_j^2} \geq \kappa_1. \tag{1.7}$$

Les hypothèses (1.1) à (1.3) traduisent une faible corrélation entre les variables pertinentes et les autres. L'Hypothèse (CI) introduite par Zhao et Yu [164], est en général supposée dans le cadre dit de *problème de sélection*. Les Hypothèses (CM) et (VPR) introduites respectivement par Donoho, Elad, et Temlyakov [51] et Bickel, Ritov, et Tsybakov [17] sont quant à elles, plus souvent faites pour des *problèmes d'estimation et de prédiction*. Notons toutefois que l'Hypothèse (VPR) est une version plus faible, i.e., moins contraignante que l'Hypothèse (CM).

Les résultats théoriques escomptés, évaluant les performances de l'estimateur LASSO, peuvent être de trois types :

- **Inégalité de Sparsité (IS) en prédiction :** où l'objectif est de produire la meilleure approximation du vecteur $X\beta^*$. Sous les Hypothèses (CM) et (VPR), les IS recherchées sont de la forme :
$$\mathbb{E}\left[\|X(\hat{\beta}^L - \beta^*)\|_n^2\right] \leq C|\mathcal{A}^*|\frac{\log(p)}{n}. \tag{1.8}$$
- **Inégalité de Sparsité (IS) en estimation :** où l'objectif est de produire l'estimation du vecteur β^*. Sous les Hypothèses (CM) et (VPR), les IS recherchées sont de la forme :
$$\mathbb{E}\left[\|\hat{\beta}^L - \beta^*\|_1\right] \leq C|\mathcal{A}^*|\sqrt{\frac{\log(p)}{n}}. \tag{1.9}$$
- **Consistance en sélection :** où l'objectif est d'identifier le support $\mathcal{A}^*$ de β^* défini en (1.2), ou bien d'estimer le vecteur signe de β^* noté :
$$\mathrm{Sgn}(\beta^*) = (\mathrm{sgn}(\beta_1^*), \ldots, \mathrm{sgn}(\beta_p^*))',$$
où la fonction $\mathrm{sgn}(\cdot)$ est définie par (1.5). Pour ce type de problèmes, on se place en général sous l'Hypothèse (CI) et les résultats théoriques s'expriment en terme de probabilité, de la façon suivante :
$$\mathbb{P}\left(\forall j \in \{1, \ldots, p\},\ \mathrm{sgn}(\hat{\beta}_j^L) = \mathrm{sgn}(\beta_j^*)\right) \leq 1 - \eta_n, \tag{1.10}$$
où $(\eta_n)_{n \geq 1}$ est une suite décroissante de termes positifs.

Notons que les *Inégalité de Sparsité* (IS) sont établies en espérance sous la loi de ε, mais peuvent également être réécrites avec des inégalités probabilistes.

Dans les deux premières situations (*prédiction* et *estimation*), de nombreux résultats de type IS ont été prouvés en grande dimension ($p \geq n$), entre autres par Bickel, Ritov, et Tsybakov [17], Bunea, Tsybakov, et Wegkamp [33, 32]. La dimension p reste toutefois un $o(e^n)$.

Dans le cadre de la *sélection*, on peut citer Zhao et Yu [164], qui ont établi la consistance en signe de l'estimateur LASSO sous l'Hypothèse (CI) lorsque $p \leq n$; ils ont montré que la probabilité (1.10) est telle que η_n converge vers zéro avec n convergeant vers l'infini. Dans ce même cadre mais lorsque $p \geq n$, Zhao et Yu [164] et Bunea [28] ont prouvé ce type de résultat à échantillon n fini, avec grande probabilité.

1.1.4 Limites de l'estimateur LASSO

Les résultats théoriques relatifs à l'estimateur LASSO nécessitent une hypothèse sur la matrice de Gram Ψ^n. Cette hypothèse n'autorise que de faibles corrélations entre les variables. D'autre part, le LASSO n'intègre pas une connaissance a priori sur le modèle : il ne permet pas d'inclure la connaissance de structures particulières entre les variables, comme par exemple la prise en compte des corrélations connues entre certaines variables. Enfin, l'estimateur LASSO nécessite d'être adapté pour pouvoir prendre en compte des problèmes dans le cadre semi-supervisé ou transductif.

L'estimateur LASSO repose sur une hypothèse implicite sur la faible dépendance des variables explicatives. L'algorithme le plus populaire pour résoudre de critère de minimisation LASSO est le *LARS*, basé également sur ces corrélations. Ainsi, dans des problèmes d'estimation avec de fortes corrélations entre les variables, l'algorithme LARS échoue à reconstituer le modèle.

1.2 Présentation des résultats

Dans la littérature, de nombreuses méthodes sont proposées pour apporter des réponses aux problèmes rencontrés par le LASSO, dont une présentation exhaustive est faite dans le Chapitre 2. Chacune des parties de cette section décrit un axe d'étude développé dans la thèse.

1.2.1 Procédure S-Lasso

Les estimateurs considérés dans cette section proposent de modifier la pénalité LASSO, en ajoutant un deuxième terme de contrôle. Ils sont de la forme :

$$\hat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|Y - X\beta\|_n^2 + \operatorname{pen}(\beta), \tag{1.11}$$

où l'application pen : $\mathbb{R}^p \to \mathbb{R}^+$, est appelée *pénalité*. Leur introduction est motivée par la volonté de résoudre les problèmes algorithmiques de l'estimateur LASSO rencontrés face aux corrélations entre les variables. Le premier estimateur que nous considérons est l'estimateur *Elastic Net* introduit par Zou et Hastie [167], défini par la pénalité

$$\operatorname{pen}(\beta) = \lambda\|\beta\|_1 + \mu \sum_{j=1}^{p} \beta_j^2.$$

Le deuxième estimateur, *Fused Lasso*, introduit par Tibshirani, Saunders, Rosset, Zhu, et Knight [136], est défini par la pénalité

$$\operatorname{pen}(\beta) = \lambda\|\beta\|_1 + \mu \sum_{j=2}^{p} |\beta_j - \beta_{j-1}|.$$

Les quantités λ et μ sont des paramètres de lissage. Ces deux estimateurs présentent un premier terme de pénalité identique (la norme ℓ_1 de β) ; celui-ci assure la sélection de variables, c'est-à-dire la sparsité de la solution $\hat{\beta}$. Ils se différencient de par leur second terme de pénalité, qui permet de sélectionner les variables pertinentes, même lorsque celles-ci sont fortement corrélées.

Bien que similaire à bien des égards à l'estimateur Elastic Net, l'estimateur Fused Lasso répond de manière plus pointue à des situations où les variables sont ordonnées (cf. Tibshirani, Saunders, Rosset, Zhu, et Knight [136] pour des applications possibles). Les résultats théoriques de Rinaldo [125], relatifs à cet estimateur confirment l'importance de la structure de blocs des variables pertinentes. L'estimateur Elastic Net, quant à lui, s'adapte à tout type de corrélations. Contrairement aux résultats expérimentaux observés dans diverses références, les résultats théoriques présentés sur l'Elastic Net ne laissent pas penser qu'il améliore significativement les performances de l'estimateur LASSO (Bunea [26] et Hebiri [78]).

Contribution : *article de Hebiri [78].*

L'estimateur que nous présentons ici est l'estimateur *S-Lasso*, qui permet de procéder à la sélection de variables lorsque celles-ci sont ordonnées. Son spectre d'application est relativement large et recouvre, entre autres, les cas cités par Tibshirani, Saunders, Rosset, Zhu,

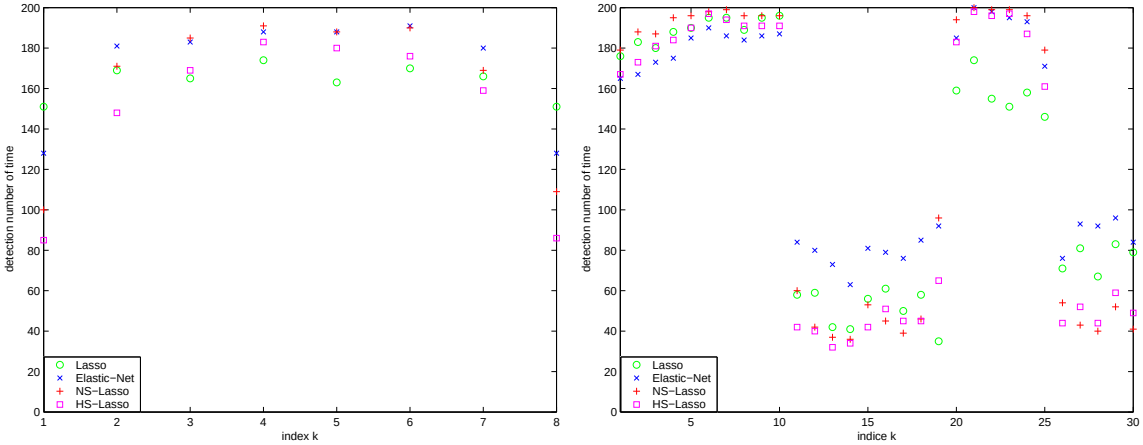

FIG. 1.1 – Nombre de détections par variable pour les procédures LASSO, Elastic Net et S-Lasso (2 versions : NS-Lasso et HS-Lasso) sur 200 itérations. Les indices des variables sont en abscisse.

et Knight [136].

L'estimateur *S-Lasso* est défini par (1.11) avec la pénalité

$$\text{pen}(\beta) = \lambda\|\beta\|_1 + \mu\sum_{j=2}^{p}(\beta_j - \beta_{j-1})^2.$$

Inspiré du Fused Lasso, cet estimateur capture les corrélations entre les variables successives. Il présente l'intérêt d'avoir un deuxième terme de pénalité strictement convexe, ce qui facilite la résolution du critère (1.11). Sur le plan pratique, le S-Lasso sera comparé à l'Elastic Net et au LASSO, méthodes de référence, sur différents types de problèmes. La Figure 1.1 illustre un cas particulier étudié dans le chapitre 3.

On constate que, dans les études où les variables sont ordonnées (structure de blocs entre variables pertinentes, successives et corrélées), l'estimateur S-Lasso offre de meilleures performances de *sélection*. Cela est particulièrement vrai pour les variables situées à l'intérieur des groupes (en terme d'indices), et cela, autant pour la sélection des variables pertinentes que pour l'éviction des variables non pertinentes. Toutefois, les performances sont légèrement dégradées lorsque la sélection concerne des variables au bord des blocs. Cet effet de bord est également rencontré par l'Elastic Net, bien que de façon moins marquée (Figure 1.1 - gauche).

Sur le plan théorique, considérons dans un premier temps les problèmes avec une dimension $p \leq n$ fixe. On suppose que $\lambda = \lambda_n \overset{n\to\infty}{\to} \lambda_0$ et que $\mu = \mu_n \overset{n\to\infty}{\to} \mu_0$, où λ_0 et μ_0 sont des paramètres positifs. Des résultats asymptotiques de consistance en *sélection* de variables

peuvent être établis sous une hypothèse inspirée de l'hypothèse 1.1 (CI). Avant de présenter cette hypothèse, introduisons les éléments suivants :

- soit la matrice $\widetilde{J}$ de taille $p \times p$, telle que $\widetilde{J}_{j,j} = 2$ pour $j \in \{2, \ldots, p-1\}$ et 1 sinon ; $\widetilde{J}_{j,j+1} = \widetilde{J}_{j+1,j} = -1$ pour $j \in \{1, \ldots, p-1\}$;
- soient les restrictions $\widetilde{J}_{\mathcal{A}^*,\mathcal{A}^*}$ et $\widetilde{J}_{(\mathcal{A}^*)^C,\mathcal{A}^*}$ de la matrice $\widetilde{J}$, introduites de manières analogues aux restrictions de Ψ^n ;
- soit $\Omega_{(\mathcal{A}^*)^C}$ tel que

$$\begin{aligned} \Omega_{(\mathcal{A}^*)^C} &= \Psi^n_{(\mathcal{A}^*)^C,\mathcal{A}^*}(\Psi^n_{\mathcal{A}^*,\mathcal{A}^*} + \mu_0 \widetilde{J}_{\mathcal{A}^*,\mathcal{A}^*})^{-1} \left(2^{-1} \operatorname{Sgn}(\beta^*_{\mathcal{A}^*}) + \frac{\mu_0}{\lambda_0} \widetilde{J}_{\mathcal{A}^*,\mathcal{A}^*} \beta^*_{\mathcal{A}^*} \right) \\ &\quad - \frac{\mu_0}{\lambda_0} \widetilde{J}_{(\mathcal{A}^*)^C,\mathcal{A}^*} \beta^*_{\mathcal{A}^*}. \end{aligned} \tag{1.12}$$

L'hypothèse que nous considérons est alors donnée par l'inégalité suivante :

$$\|\Omega_{(\mathcal{A}^*)^C}\|_\infty \leq 1 - C.$$

Dans le cas $p \leq n$, il existe toujours un couple de paramètres (λ_0, μ_0) satisfaisant cette hypothèse. Dans le cas $p \geq n$, les IS sont établies pour les erreurs de *prédiction* et d'*estimation* (en norme ℓ_1) sous l'hypothèse 1.2 (CM), avec grande probabilité. De plus, des résultats de consistance en *sélection* de variables du S-Lasso sont obtenus sous l'hypothèse suivante :

$$\max_{k \neq j} |(\Psi^n + \mu_n \tilde{J})_{j,k}| \leq \frac{C}{|\mathcal{A}^*|}, \tag{1.13}$$

lorsque les coefficients $|\beta^*_j|$ avec $j \in \mathcal{A}^*$ sont plus grands qu'un seuil de l'ordre de $\sqrt{\frac{\log p}{n}}$. Remarquons que l'estimateur LASSO est lui consistant en *sélection* de variables sous une hypothèse similaire correspondant à $\mu_n = 0$ (cf. Lounici [101] pour plus de détails). Dans le cas du S-Lasso, certaines valeurs du paramètre μ_n permettent de satisfaire l'hypothèse (1.13) pour des problèmes présentant des corrélations entre variables successives.

Par ailleurs, les résultats obtenus pour le S-Lasso sont également généralisables à tout estimateur solution du problème de minimisation suivant :

$$\hat{\beta} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_n^2 + \lambda \|\beta\|_1 + \mu \beta' M \beta,$$

où M est une matrice de taille $p \times p$. Un bon choix pour la matrice M permet d'introduire une information a priori sur le modèle. L'Elastic Net appartient à cette famille d'estimateurs. Toutefois, les résultats obtenus pour l'Elastic Net ne sont pas satisfaisants. En effet, pour cette méthode d'estimation, l'hypothèse 1.13 devient identique à l'hypothèse 1.2 (CM), utilisée pour le LASSO. Ces résultats vont dans le sens des recherches menées par Bunea [26] qui montrent que l'Elastic Net n'apporte qu'une très légère amélioration du LASSO sur le plan théorique.

1.2.2 Méthode Groupée

Lorsque les variables $X_1, \ldots, X_p$ relèvent d'une structure de groupe, il s'avère plus pertinent d'utiliser un estimateur prenant en compte cette caractéristique. Introduit par Yuan et Lin [159], l'estimateur *Group Lasso* est devenu rapidement populaire en permettant d'appréhender des groupes de variables. Il est défini par :

$$\hat{\beta}^G \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|Y - X\beta\|_n^2 + \lambda \sum_{l=1}^{L} \sqrt{\sum_{j \in G_l} \beta_j^2}. \tag{1.14}$$

où λ est le paramètre de régularisation, L est un entier dans $\{1, \ldots, p\}$, et $(G_l)_{l \in \{1,\ldots,L\}}$ est une partition de $\{1, \ldots, p\}$. Dans cette définition, L désigne le nombre de groupes et G_l est l'ensemble des indices des variables contenues dans le groupe l avec $l = 1 \ldots, L$.

Sous une hypothèse issue d'une version groupée de la Condition d'Irreprésentabilité, Bach [12] démontre que le Group Lasso est consistant en sélection de *groupes de variables* :

$$\mathbb{P}(\hat{\mathcal{A}}_G = \mathcal{A}_G^*) \to 1,$$

où $\mathcal{A}_G^* = \{l \in \{1, \ldots, L\} : \exists j \in G_l \text{ avec } \beta_j^* \neq 0\}$ et $\hat{\mathcal{A}}_G$ est son équivalent en considérant $\hat{\beta}_j^G$ à la place de β_j^*. Par analogie à la notion de *variable pertinente*, nous appelons *groupe pertinent* un groupe G_l contenant au moins une variable pertinente, i.e., pour $l \in \mathcal{A}_G^*$.
Aucun résultat de consistance en sélection de *variables individuelles* n'a encore été établi pour le Group Lasso.

Contribution : *article de Chesneau et Hebiri [44].*

Dans le chapitre 4, nous étudions les propriétés théoriques du Group Lasso en grande dimension, pour lequel nous établissons la première IS. Celle-ci est obtenue sous une version groupée de l'hypothèse 1.2 (CM), exprimée sous la forme :

$$\max_{l \in \mathcal{A}_G^*} \max_{m=1,\ldots,L} \sqrt{\sum_{j \in G_l} \sum_{\substack{k \in G_m \\ k \neq j}} (\Psi_{j,k}^n)^2} \leq \frac{C}{\operatorname{Card}(\mathcal{A}_G^*)}.$$

L'IS vérifiée par le Group Lasso est de la forme :

$$\|X\hat{\beta}^G - X\beta^*\|_n^2 \leq C|\mathcal{A}_G^*| \frac{\log p}{n},$$

où l'inégalité ci-dessus est vérifiée avec grande probabilité. L'intérêt de la borne ci-dessus repose sur l'introduction d'un indice de sparsité groupé $|\mathcal{A}_G^*|$, avec $|\mathcal{A}_G^*| \leq |\mathcal{A}^*|$. Ainsi, le

Group Lasso exploite mieux la sparsité du modèle que le LASSO.

Le Group Lasso se distingue d'autant mieux du LASSO que les groupes constitués sont homogènes en terme de pertinence et que le nombre des groupes est petit. Inversement, l'IS montre l'inefficacité du Group Lasso dans des situations contraires, lorsque les groupes sont constitués d'une faible proportion de variables pertinentes, ou lorsque le nombre de groupes est important. Un cas extrême non favorable au Group Lasso est le cas où une seule variable par groupe est pertinente.

Les considérations ci-dessus permettent d'affirmer que le Group Lasso répond essentiellement à des problèmes où les variables sont bien groupées voire même ordonnées. Dans ce cadre, des conclusions similaires ont récemment été apportées par Huang et Zhang [81]. En effet, les auteurs complètent les résultats que nous avons obtenus en considérant l'erreur d'estimation en norme ℓ_2. Huang et Zhang [81] démontrent une IS en *estimation* de la forme :

$$\|\hat{\beta}^G - \beta^*\|_2^2 \leq C \left(\frac{s_G + |\mathcal{A}_G^*| \log(L)}{n} \right),$$

où $s_G = |\bigcup_{l \in \mathcal{A}_G^*} G_l|$, est le nombre total de variables contenues dans les *groupes pertinents*. En comparant cette borne à celle obtenue pour le LASSO (1.9), on remarque que les cas favorables au Group Lasso sont ceux où les variables sont regroupées judicieusement (i.e., où les variables pertinentes et les variables non pertinentes sont séparées) et où les groupes sont de grandes tailles.

Notons d'autre part que parmi les résultats établis dans le Chapitre 4, une borne supérieure sur la taille des groupes est fournie. La taille maximale est évaluée à $\lfloor \log(p) \rfloor$, où $\lfloor \cdot \rfloor$ désigne la partie entière. Ceci est en accord avec la théorie de l'estimateur de Stein par bloc (une autre méthode groupée utilisée dans le modèle de suite gaussienne), pour lequel la taille optimale des groupes est également de l'ordre de $\log(p)$ (Cavalier et Tsybakov [40]).

1.2.3 Méthode exploitant un type particulier de sparsité

Nous considérons le modèle de régression où le paramètre inconnu β^* présente une sparsité particulière. Le vecteur β^* n'est pas sparse au sens habituel (i.e., $\|\beta^*\|_0 = \sum_{j=1}^p \mathbb{I}(\beta_j^* \neq 0)$ n'est pas petite où $\mathbb{I}(\cdot)$ est la fonction indicatrice). Or, il a été prouvé que la sparsité du modèle est essentielle à la détermination des IS (cf. les résultats de bornes inférieures obtenues par Bunea, Tsybakov, et Wegkamp [32, Theorem 5.1]).

Avant de mener la présentation plus en avant, introduisons une notion plus large de la

sparsité : on suppose qu'il existe une matrice connue P de taille $p \times p$, telle que

$$\|P\beta^*\|_0 = \sum_{j=1}^{p} \mathbb{I}((P\beta)_j^* \neq 0)$$

est petit. Dans ce cadre, les méthodes présentées précédemment ne fournissent pas de bonnes performances théoriques. En effet, les bornes obtenues dans les IS (1.8) et (1.9) font intervenir la quantité $\|\beta^*\|_0$ qui peut être égale à p dans la situation de cette section. L'erreur n'est alors plus contrôlée. Avant d'introduire notre contribution, il est utile de considérer des techniques provenant de la théorie du LASSO.

Introduisons un estimateur proche de l'estimateur LASSO écrit sous sa forme duale (1.4). Il s'agit du Dantzig Selector (Candès et Tao [35]), qui est défini par :

$$\hat{\beta}^D(\lambda) \in \begin{cases} \operatorname{argmin}_{\beta \in \mathbb{R}^p} \|\beta\|_1 \\ \\ \text{s.c. } \left\|\Xi^{-1} X' (Y - X\beta)\right\|_\infty \leq \lambda, \end{cases}$$

où $\lambda \geq 0$ est le paramètre de régularisation et Ξ est la matrice diagonale de taille p, dont le j-ième coefficient diagonal vaut $\|X_j\|_n = \sqrt{\frac{1}{n}\sum_{i=1}^n x_{i,j}^2}$. Les propriétés de cet estimateur ont été considérées par de nombreux auteurs dont Bickel, Ritov, et Tsybakov [17], Candès et Tao [35], Koltchinskii [91] et Lounici [101]. En outre, les IS et les résultats de consistance en *sélection* de variables obtenus (cf. James, Radchenko, et Lv [84] et Bickel, Ritov, et Tsybakov [17]) témoignent de l'équivalence entre le Lasso et le Dantzig Selector.
Une interprétation géométrique semblable peut être proposée pour les estimateurs LASSO (sous sa forme duale (1.4)) et Dantzig Selector : *ces deux estimateurs assurent la projection au sens des distances ℓ_2 et ℓ_1, respectivement, de $\mathbf{0_p}$ sur la région de confiance* $DC(\lambda) = \left\{\beta : \left\|\Xi^{-1} X' (Y - X\beta)\right\|_\infty \leq \lambda\right\}$. Cette interprétation géométrique est mise en évidence par Alquier [4] et par Alquier et Hebiri [7].

Contribution : *article de Alquier et Hebiri [7].*

Dans ce cadre de l'estimation avec une sparsité particulière, nous proposons de modifier légèrement les définitions des estimateurs LASSO et Dantzig Selector. Nous étudions la famille d'estimateurs définis par projection sur $DC(\lambda)$, donnée par :

$$\textbf{Programme I :} \qquad \hat{\beta}^P = \operatorname*{argmin}_{\beta \in DC(\lambda)} \|Z\beta\|_2^2,$$

$$\textbf{Programme II :} \qquad \tilde{\beta}^P = \operatorname*{argmin}_{\beta \in DC(\lambda)} \|P\beta\|_1,$$

où Z est une matrice quelconque, de taille $m \times p$ avec $m \in \mathbb{N}^*$.
Le Dantzig Selector correspond au cas particulier où $P = \mathbf{I_p}$ dans le Programme II, où $\mathbf{I_p}$ est la matrice identité de taille p. Le LASSO correspond au cas où $Z = X$ dans le Programme I. D'une manière plus générale, et dans le but d'exploiter la sparsité de $P\beta^*$, la matrice Z s'écrit explicitement en fonction de X et P.

Pour les deux problèmes de minimisation précédents, une hypothèse supplémentaire est à considérer en grande dimension ($p \geq n$) :

– *Condition des noyaux : la matrice Z est telle que* $\ker Z = \ker X$.

Les résultats décrits au Chapitre 5 sont établis sous la condition des noyaux et sous une version de l'hypothèse (VPR), où l'inégalité (1.7) doit être vérifiée, avec une matrice Ω dépendant de $X'X$ et de $Z'Z$. Cette hypothèse peut dans certains cas être moins restrictive que l'hypothèse (VPR) originelle. En étudiant une forme duale de l'estimateur défini par le Programme I, nous obtenons pour l'estimateur $\hat{\beta}$, selon qu'il soit égal à $\hat{\beta}^P$ ou $\tilde{\beta}^P$, des IS de la forme :

$$\left\| Z(\hat{\beta} - \beta^*) \right\|_2^2 \leq C_1 \log(p) \left\| P\beta^* \right\|_0 ,$$

et

$$\left\| P\left(\hat{\beta} - \beta^*\right) \right\|_1 \leq C_2 \sqrt{\frac{\log(p)}{n}} \left\| P\beta^* \right\|_0 ,$$

pour une valeur habituelle de λ de l'ordre de $\sqrt{\frac{\log(p)}{n}}$. Dans le Chapitre 5, nous présentons plus en détails les résultats précédents et abordons en particulier les problèmes d'unicité.

Cette première contribution à l'estimation dans le cadre d'un modèle de régression linéaire avec une sparsité sur $P\beta^*$, avec P connue et quelconque, reste pour l'heure d'ordre purement théorique. En outre, certains cas particuliers du modèle ont été étudiés, qui connaissent une mise en application immédiate. Citons par exemple l'estimateur *Correlation Selector* introduit par Alquier [4], ou encore des versions transductives du LASSO et du Dantzig Selector, qui font l'objet d'une présentation dans la section suivante.

1.2.4 Méthodes transductives

On considère de nouveau le modèle de régression linéaire (1.1) avec β^* supposé sparse. Le statisticien observe $X = (x_1', \ldots, x_n')'$ et $Y = (y_1, \ldots, y_n)'$, où y_i est la réponse associée à x_i. Dans cette section, nous supposons que nous observons également $Z = (x_{n+1}', \ldots, x_{n+m}')'$ avec $m \in \mathbb{N}^*$. Nous disposons de ce fait de m nouveaux individus pour lesquels nous ne connaissons pas les valeurs des réponses associées $y_{n+1}, \ldots, y_{n+m}$. En se plaçant dans le cadre

de l'inférence *transductive* (Vapnik [145]), une estimation fine du vecteur $(y_{n+1}, \ldots, y_{n+m})'$ peut être produite en utilisant des arguments locaux pour l'estimation (cf. Györfi, Kohler, Krzyżak, et Walk [73]). Prenons le cas des enquêtes de sondage, domaine particulièrement adapté à l'inférence transductive où nous sommes souvent confrontés à des données X, Y et Z telles que $p \geq n$ et $m \geq n$. Pour des raisons pratiques, le sondage ne peut être effectué sur l'ensemble de la population. L'enquête est souvent réduite à un petit échantillon (produisant dans notre cadre les données (X, Y)). Le but est de généraliser la méthode pour les individus supplémentaires $Z = (x'_{n+1}, \ldots, x'_{n+m})'$.

En considérant les n premières observations, nous pouvons répondre aux trois objectifs classiques suivants :

- **Objectif 1** : **Estimation** $\longrightarrow$ l'estimation du paramètre β^*.
- **Objectif 2** : **Sélection** $\longrightarrow$ l'estimation du support de β^*.
- **Objectif 3** : **Prédiction** $\longrightarrow$ la reconstitution du signal $X\beta^*$.

Etant données les nouvelles observations Z non étiquetées (i.e., observées sans les valeurs des réponses associées), un autre objectif apparaît, qui est :

- **Objectif 4** : **Transduction** $\longrightarrow$ l'estimation du vecteur $Z\beta^*$.

Il existe plusieurs approches possibles pour répondre à cet objectif, nous en présentons ici deux :

- *Approche en deux étapes :* On estime β^* en appliquant une des méthodes de sélection de variables classique (LASSO, Dantzig, etc.) sur l'échantillon étiqueté (X,Y). On obtient alors l'estimateur de β^* noté $\hat{\beta}(X, Y)$. La prédiction $Z\hat{\beta}(X, Y)$ produit par la suite une estimation de $(y_{n+1}, \ldots, y_{n+m})'$.
- *Approche directe :* On construit un estimateur de β^* à l'aide de X, Y et de Z, noté $\hat{\beta} = \hat{\beta}(X, Y, Z)$. De la même façon que précédemment, $Z\hat{\beta}(X, Y, Z)$ produit une estimation de $(y_{n+1}, \ldots, y_{n+m})'$. Cette dernière approche présente l'avantage d'exploiter également les données Z pour construire l'estimateur $\hat{\beta}$.

Contribution : *article de Alquier et Hebiri [8].*

Dans le chapitre 6, nous apportons, entre autres, une réponse à l'objectif de transduction, pour lequel, il existe des méthodes de sélection de variables (cf. Catoni [37, 39] et Alquier [6] par exemple), à notre connaissance, aucune conjugue bonnes performances théoriques avec les facilités d'implémentation du LASSO. L'étude menée dans ce chapitre se place dans la continuité du travail réalisé dans le chapitre 5 (cf. résumé de la section 1.2.3). Les

estimateurs que nous étudions sont de la forme :

$$\textbf{Programme T1 :} \qquad \hat{\beta}^{LT} = \underset{\beta \in TDC(\lambda)}{\operatorname{argmin}} \|Z\beta\|_2^2,$$
$$\textbf{Programme T2 :} \qquad \hat{\beta}^{DT} = \underset{\beta \in TDC(\lambda)}{\operatorname{argmin}} \|\beta\|_1,$$

où la région de confiance $TDC(\lambda)$ est définie par :

$$TDC(\lambda) = \left\{\beta : \ \left\|\Xi^{-1}\frac{n}{m}Z'Z((\widetilde{X'X})^{-1}X'Y - \beta)\right\|_\infty \leq \frac{n}{2}\lambda\right\},$$

où :

- λ est le paramètre de régularisation,
- $(\widetilde{X'X})^{-1}$ est une pseudo-inverse de $X'X$,
- Ξ est la matrice diagonale dont le coefficient (j,j) est égale à $\xi_j^{\frac{1}{2}}$ avec $\xi_j = [\frac{Z'Z}{m}(\widetilde{\frac{X'X}{n}})^{-1}\frac{Z'Z}{m}]_{j,j}$, introduite à des fins de normalisation.

Par la suite, pour $\hat{\beta} = \hat{\beta}^{LT}$, ou $\hat{\beta}^{DT}$, nous démontrons les résultats suivants, obtenus avec grande probabilité :

- l'IS en *tranduction* : $\frac{1}{m}\left\|Z(\hat{\beta} - \beta^*)\right\|_2^2 \leq C_1\frac{\log(p)}{n}|\mathcal{A}^*|$,
- l'IS en *estimation* : $\left\|\hat{\beta} - \beta^*\right\|_1 \leq C_2\sqrt{\frac{\log(p)}{n}}|\mathcal{A}^*|$.

Ces résultats sont vrais sous la *Condition des noyaux* $\ker Z = \ker X$ et sous l'hypothèse 1.3 (VPR), où l'inégalité (1.7) est vérifiée pour la matrice $\frac{Z'Z}{m}$ (en lieu et place de la matrice $\Psi^n = \frac{X'X}{n}$). Il est intéressant de remarquer que, dans le cadre de la transduction, les résultats font intervenir des hypothèses sur la matrice des corrélations $\frac{Z'Z}{m}$ des nouvelles observations Z, sans pour autant en faire intervenir sur la matrice de Gram $\Psi^n = \frac{X'X}{n}$.

Quelques remarques :

– La région $TDC(\lambda)$ peut se réécrire sous la forme

$$TDC(\lambda) = \left\{\beta : \ \left\|\Xi^{-1}\bar{Z}'(\bar{Y} - \bar{Z}\beta)\right\|_\infty \leq \frac{n}{2}\lambda\right\}, \tag{1.15}$$

avec $\bar{Z} = \left(\sqrt{\frac{n}{m}}Z\right)$ et $\bar{Y} = \left(\sqrt{\frac{n}{m}}Z\right)(\widetilde{X'X})^{-1}X'Y$. Ainsi, les estimateurs $\hat{\beta}^{LT}$ et $\hat{\beta}^{DT}$ peuvent être vus comme des estimateurs LASSO et Dantzig définis à partir de données modifiées afin de répondre à l'objectif de transduction.

– D'un point de vue pratique, les solutions LASSO Transductif et Dantzig Selector Transductif peuvent être obtenues par le biais d'algorithmes fournissant des solutions approchées aux estimateurs LASSO et Dantzig Selector respectivement, à ceci près que l'on considère

ici les données $(\bar{Z}, \bar{Y})$ au lieu de (X, Y). En outre, une méthode de type LARS est parfaitement adaptée.

− Le Chapitre 6 considère le cas d'une relaxation de la *Condition des noyaux*. La nouvelle hypothèse s'énonce de la façon suivante :

Pour tout $u \in \mathbb{R}^p$ tel que $\|u\|_1 \leq \|\beta^*\|_1$, on suppose que

$$\left\| \left(\frac{X'X}{n} - \frac{Z'Z}{m} \right) u \right\|_\infty \leq \sigma \sqrt{\frac{2\log(p)}{n}}. \tag{1.16}$$

Sous cette condition, les IS décrites précédemment restent valides. Notons par ailleurs que cette hypothèse est simplement vérifiable dans notre cadre de travail (cf. Proposition 6.3 du chapitre 6).

En tenant compte des remarques ci-dessus, un résultat plus général peut également être établi. En effet, d'après l'écriture (1.15) de la contrainte, nous pouvons interpréter $\bar{Y}$ comme un estimateur préliminaire de $(y_{n+1}, \ldots, y_{n+m})'$. En outre, nous pouvons utiliser un estimateur préliminaire de quelque forme que ce soit, en remplacement de $\bar{Y}$. Dans le cadre de nos travaux, nous considérons les estimateurs LASSO et Dantzig Selector originels, i.e., nous remplaçons $\bar{Y}$ par $Z\hat{\beta}^L(\lambda)$ ou encore par $Z\hat{\beta}^D(\lambda)$. Sous l'hypothèse 1.16, les IS précédemment énoncées sont encore valides pour les nouvelles versions transductives du LASSO et Dantzig Selector.

Les performances pratiques de ces derniers estimateurs ont été évaluées, montrant que l'estimateur LASSO Transductif améliore l'estimateur LASSO originel dans plusieurs situations. Notons que le choix des paramètres de régularisation est également discuté au Chapitre 6.

1.2.5 Prédiction Conforme

Un cadre particulier de l'approche transductive est la prédiction séquentielle, où à chaque "instant", le statisticien obtient une nouvelle observation x_i. Son objectif est de fournir la réponse y_i associée avec la plus grande fiabilité. La notion de fiabilité peut se traduire en terme de prédiction avec intervalle de confiance. La taille de l'intervalle donne alors une notion de la qualité de la prédiction. Dans des travaux récents, Vovk, Gammerman, et Shafer [150] considèrent la construction d'intervalles de confiance, appelés *Prédicteurs Conformes* (ou *Conformal Predictors* en anglais) basée sur une approche séquentielle.

Pour une telle stratégie, les auteurs exploitent à chaque instant n les observations $\mathcal{E}_n = ((x_1, y_1), \ldots, (x_{n-1}, y_{n-1}), x_{new})$, où $x_{new} \in \mathbb{R}^p$, afin de fournir le meilleur intervalle de

confiance de y_{new}, le label inconnu de l'observation x_{new}. Cette méthode s'appuie sur des outils permettant d'évaluer le degré de conformité d'un label y au vu de $\mathcal{E}_n$.
Les outils visent à évaluer si le couple (x_{new}, y) est semblable aux $n-1$ précédents couples $(x_1, y_1), \ldots, (x_{n-1}, y_{n-1})$ avec un bon indice de confiance. Pour ce faire, les auteurs associent à chaque label y un vecteur de taille n, noté $\alpha(y) = (\alpha_1(y), \ldots, \alpha_n(y))'$, dont la i-ème composante est une mesure de la *non-conformité* du i-ème couple, lorsque y est associé à x_{new} dans la construction des estimations. On construit une *p-value* de y comme une quantification relative des α_i de la manière suivante :

$$p(y) = \frac{1}{n} \left| \{ i \in \{1, \ldots, n\} : \alpha_i(y) \geq \alpha_n(y) \} \right| . \tag{1.17}$$

Plus la p-value $p(y)$ est grande, plus la paire (x_{new}, y) est conforme aux autres couples. Le prédicteur conforme est alors donné par :

$$\Gamma^{\varepsilon} = \{ y \in \mathbb{R} : p(y) \geq \varepsilon \},$$

où $\varepsilon \in]0, 1[$. Cette notion de p-value diffère de la définition classique. Toutefois, en définissant les hypothèses suivantes :

$$\begin{cases} H_0 : & \text{la paire } (x_{new}, y) \text{ est conforme,} \\ H_1 : & \text{la paire } (x_{new}, y) \text{ n'est pas conforme,} \end{cases}$$

une connexion avec les tests d'hypothèses en statistique mathématique peut être établie. En effet, la fonction $p(y)$ définit une procédure de test statistique avec une région critique donnée par $\mathcal{R}_{\varepsilon} = \{ y : p(y) \leq \varepsilon \}$: on rejete H_0 si $y \in \mathcal{R}_{\varepsilon}$.

Dans l'approche proposée de Vovk, Gammerman, et Shafer [150], la taille du prédicteur conforme augmente quand ε diminue, signifiant que l'on tolère une moins grande conformité de (x_{new}, y). En introduisant la suite d'*erreurs de validité*, $(\mathrm{err}_1^{\varepsilon}, \ldots, \mathrm{err}_n^{\varepsilon})$, où :

$$\mathrm{err}_i^{\varepsilon} = \mathbb{I}(y_i \notin \Gamma^{\varepsilon}(\mathcal{E}_i)),$$

Vovk [147] et Vovk, Gammerman, et Shafer [150, chapitres 2 et 8], montrent que les $\mathrm{err}_i^{\varepsilon}$ sont indépendantes entre elles et que chaque $\mathrm{err}_i^{\varepsilon}$ est une variable aléatoire dominée en loi par une Bernoulli de paramètre ε.
Pour de tels prédicteurs conformes, Vovk, Gammerman, et Shafer [150] étudient :

- *La validité.* L'erreur commise $\mathrm{err}_i^{\varepsilon}$, ou de manière plus générale l'erreur cumulée $\mathrm{Err}_n^{\varepsilon} = \sum_{i=1}^{n} \mathrm{err}_i^{\varepsilon}$.
- *La précision.* La taille du prédicteur conforme donne une indication sur sa fiabilité.

Vovk, Gammerman, et Shafer [150] considèrent les prédicteurs conformes dans le modèle de régression, en basant la construction des *scores de non-conformité* $\alpha(y)$ sur des estimateurs

tels que l'estimateur Ridge ou la méthode des plus proches voisins. Pour des problèmes en classification, les auteurs utilisent des méthodes à noyau.

Contribution : *article de Hebiri [77].*

Dans le chapitre 7, nous proposons une application des prédicteurs conformes au modèle de régression linéaire sous l'hypothèse que le paramètre de régression est sparse. A la manière de Vovk, Gammerman, et Shafer [150] qui basent la construction de leurs intervalles sur un estimateur (comme par exemple l'estimateur Ridge), nous suggérons le LASSO comme estimateur de référence pour ses qualités de sélection. La méthode de construction de l'intervalle de confiance de $y_n = y_{new}$ est la suivante :

- *Étape 1 (Sélection) :* On utilise l'algorithme LARS pour fournir K estimations (chacune correspondant à une étape de l'algorithme LARS, décrit dans l'annexe B), du vecteurs $\mathrm{Sgn}(\beta^*)$ pour K valeurs du paramètre de régularisation $\lambda^{(1)}, \ldots, \lambda^{(K)}$ (ces λ correspondent aux valeurs du paramètre de régularisation telles que l'algorithme s'actualise à chaque début d'étape).

- *Étape 2 (Construction) :* Pour chaque $\lambda^{(k)}$, $k = 1, \ldots, K$, on construit le prédicteur conforme associé. L'algorithme utilisé est décrit dans le chapitre 7. Il s'agit d'une adaptation de l'algorithme de Vovk, Gammerman, et Shafer [150] à l'utilisation de l'estimateur LASSO. Nous obtenons de ce fait, K prédicteurs conformes $\Gamma^{\varepsilon}_1, \ldots, \Gamma^{\varepsilon}_K$.

- *Étape 3 (Choix de λ_{opt}) :* Il s'agit ici de choisir le meilleur intervalle parmi ceux construits à l'étape précédente. Le prédicteur conforme final répondant au critère de *précision* est celui ayant la plus petite taille.

A l'*étape 3*, nous proposons une méthode de sélection du paramètre de régularisation optimal λ_{opt} dépendant seulement des données. Ceci augmente l'intérêt de cette méthode qui ne dépend pas d'un critère extérieur pour ce choix. La Figure 1.2 illustre un exemple d'intervalles construits sur des données simulées avec $p = 50$, $n = 300$ et $|\mathcal{A}^*| = 20$. Le prédicteur final obtenu en appliquant la règle définie à l'*étape 3* est désigné par une flèche. Au chapitre 7, une étude des performances de ce prédicteur final appelé *Conformal Lasso Predictor (ColP)*, notamment dans le cas $p \geq n$ est proposée. Pour s'adapter à ce dernier cas, une variante dans la construction de l'estimateur est apportée.

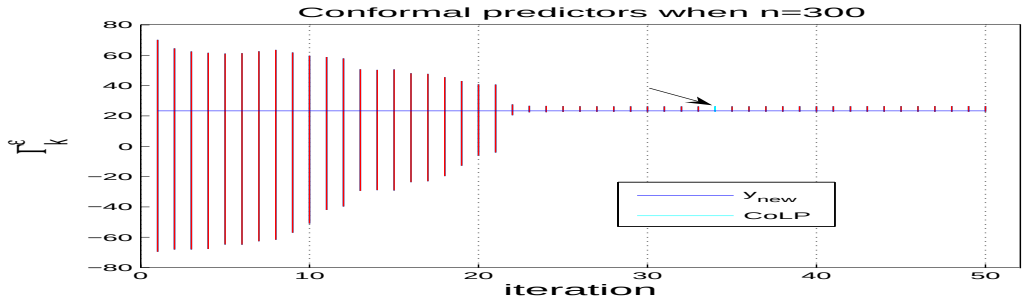

FIG. 1.2 – Évolution de prédicteurs conformes Γ_k^ε par rapport aux itérations de l'algorithme LARS. Le nombre d'observations est de $n = 300$ et le nombre de variables $p = 50$. (La première étape correspond à λ_{max} et la dernière correspond à λ_{min}). Le prédicteur sélectionné est tracé en cyan et est signalé par une flèche. La ligne horizontale bleue corresponds à la valeur de y_{new}.

Chapitre 2

Estimateur LASSO et ses variantes

Ce chapitre d'introduction présente le cadre général et l'estimateur LASSO comme exemple de référence. Il offre de plus une synthèse bibliographique des différentes méthodes de sélection de variables.

2.1 Cadre général

2.1.1 Modèle de régression linéaire

Nous considérons que les données collectées sont issues du modèle de régression linéaire multidimensionnel. On dispose d'un échantillon $(x_1, y_1), \ldots, (x_n, y_n)$ tel que

$$y_i = x_i\beta^* + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

où les ε_i sont des variables aléatoires mutuellement indépendantes de loi normale $\mathcal{N}(0, \sigma^2)$, avec σ connu. Dans cette expression, le vecteur $x_i = (x_{i,1}, \ldots, x_{i,p})$ est un vecteur p-dimensionnel, supposé déterministe et $\beta^* \in \mathbb{R}^p$ est un vecteur inconnu des paramètres de régression. Il est commode d'écrire le modèle de régression linéaire sous forme matricielle. Pour cela définissons X et X_j

$$X = \begin{pmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} X_1 & \cdots & X_p \end{pmatrix},$$

et définissons Y, ε et β^* tels que

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \quad \text{et } \beta^* = \begin{pmatrix} \beta_1^* \\ \vdots \\ \beta_p^* \end{pmatrix}.$$

Dans la littérature, on appelle la matrice X la *matrice des données* (matrice *design* en anglais), les variables X_j les *variables explicatives*, et Y la *réponse*.

Nous travaillons avec des données en grande dimension et nous autorisons même le nombre de variables p à dépendre de n. Nous avons donc $p = p(n)$. Dans la majorité des travaux présentés ici, nous supposons $p(n) \geq n$. Le modèle (2.1) peut être réécrit sous la forme matricielle suivante :

$$Y = \sum_{j=1}^{p} \beta_j^* X_j + \varepsilon = X\beta^* + \varepsilon. \tag{2.2}$$

En utilisant ces notations, soulignons que la composante β_j^* du vecteur β^* traduit l'influence de la variable X_j sur la réponse Y. Classiquement, nous nous intéressons à l'estimation de β^* et de $X\beta^*$, mais nous verrons plus loin qu'il peut y avoir d'autres objectifs.

2.1.2 Estimateur des moindres carrés et estimateur ridge

Nous présentons ci-après deux méthodes populaires pour estimer le paramètre β^*, l'estimateur des moindres carrés et l'estimateur ridge.

Estimateur des moindres carrés (EMC)

La méthode usuelle pour estimer le paramètre $\beta^* \in \mathbb{R}^p$ est celle des moindres carrés. Elle consiste à chercher une valeur $\hat{\beta}$ du paramètre qui minimise la somme des carrés des résidus :

$$\sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2 = \min_{\beta\in\mathbb{R}^p}\sum_{i=1}^{n}(y_i - x_i\beta)^2.$$

Le problème de minimisation de la somme des carrés des résidus peut s'écrire sous forme matricielle comme

$$\|Y - X\hat{\beta}\|_n^2 = \min_{\beta\in\mathbb{R}^p}\|Y - X\beta\|_n^2,$$

où $\|a\|_n$ dénote la norme empirique dans $\mathbb{R}^n$, $\|a\|_n^2 = \frac{1}{n}\sum_{i=1}^{n} a_i^2$. Il est facile de voir qu'il existe toujours une solution $\hat{\beta}$ de ce problème de minimisation que l'on appelle estimateur des moindres carrés (EMC) de β^* et noté $\hat{\beta}^{MC}$. On écrit alors

$$\hat{\beta}^{MC} \in \operatorname*{argmin}_{\beta\in\mathbb{R}^p}\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i\beta)^2 = \operatorname*{argmin}_{\beta\in\mathbb{R}^p}\|Y - X\beta\|_n^2.$$

Si la matrice $X'X$ inversible, il est établi que l'EMC est unique et s'écrit sous la forme

$$\hat{\beta}^{MC} = (X'X)^{-1}X'Y. \tag{2.3}$$

En revanche, la matrice $X'X$ n'est pas inversible

- pour $p > n$, la matrice $X'X$ étant au plus de rang n,
- lorsque les colonnes de X sont liées (e.g., $X_2 = 2X_1$).

Dans ces deux situations, le paramètre β^* n'est pas identifiable. Dans le cadre de ce manuscrit, nous nous intéressons plus particulièrement au cas où $p > n$. La matrice $X'X$ n'étant plus inversible, on utilise la pseudo-inverse (voir par exemple Penrose [120]) qui permet d'étendre à ce cas la formule de l'EMC. Cette extension n'est utile que pour l'estimation de $X\beta^*$. Pout tout (p, n) l'erreur de prédiction de l'EMC vaut :

$$\mathbb{E}\left[\|X\hat{\beta}^{MC} - X\beta^*\|_n^2\right] = \sigma^2\frac{\mathrm{rg}(X)}{n}, \tag{2.4}$$

où $\mathbb{E}$ est l'espérance sous la loi de probabilité de ε et $\mathrm{rg}(X)$ est le rang de la matrice X. L'estimateur $X\hat{\beta}^{MC}$ de $X\beta^*$, basé sur l'EMC, est alors consistant au sens ℓ_2 pour l'estimation de $X\beta^*$, quand n tend vers l'infini, dès lors que $\frac{\mathrm{rg}(X)}{n} = o(1)$. En particulier, cette consistance est vérifiée dès que $p(n) = o(n)$, ou lorsque p est fixe, étant donné que $\mathrm{rg}(X) \leq \min\{p, n\}$.

Inconvénients de l'EMC. Cet estimateur présente quelques inconvénients majeurs :

- D'un point de vue *statistique* : dans le cadre de la grande dimension ($p(n) \geq n$), l'EMC n'est pas défini de manière unique, et les résultats obtenus ne sont pas interprétables. Dans ce cadre, cet estimateur n'est en général pas consistant au sens ℓ_2 pour l'estimation de $X\beta^*$; excepté pour le cas où les colonnes/variables de X sont liées de telle sorte que $\mathrm{rg}(X)$ devienne asymptotiquement négligeable devant n.
- D'un point de vue *numérique* : pour p proche de n (et $p < n$), ou encore lorsque certaines variables X_j, $j \in \{1, \ldots, p\}$ sont corrélées, des valeurs propres de la matrice $X'X$ sont proches de zéro. La matrice $X'X$ est alors mal conditionnée et le calcul numérique de l'inverse de cette matrice est instable, rendant inexploitable l'EMC pour ce type de problèmes. Nous renvoyons le lecteur intéressé au livre de Ciarlet [45] pour plus de détails sur le conditionnement de matrices et au livre de Lawson et Hanson [96] pour un aperçu des méthodes de résolution du critère de l'EMC.

Estimateur ridge

Les inconvénients de l'EMC, énoncés précédemment, sont des raisons de l'introduction de certains M-estimateurs pénalisés, solutions du critère :

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|Y - X\beta\|_n + \mathrm{pen}(\beta, n), \tag{2.5}$$

où l'application pen : $\mathbb{R}^p \to \mathbb{R}^+$, est appelée *pénalité* (cf. Bickel et Li [18] pour plus de détails). Le choix de la pénalité pen est liée et dépend de la finalité de l'étude considérée. Bien choisie, elle permet d'éviter l'inversion de la matrice $X'X$ imposée par l'EMC (2.3) (cf. Witten et Tibshirani [155], pour plus de détails sur l'utilisation de l'estimation de l'inverse de la matrice $X'X$ dans le modèle de régression linéaire). Parmi la famille d'estimateurs pénalisés (2.5), l'estimateur ridge (Hoerl et Kennard [79]) est l'un des plus populaires. La pénalité ridge est définie comme suit :

$$\mathrm{pen}(\beta, n) = \lambda \sum_{j=1}^{p} \beta_j^2,$$

où $\lambda \geq 0$ est un paramètre de lissage qui dépend éventuellement de la taille de l'échantillon n. Contrairement à l'EMC, l'estimateur ridge $\hat{\beta}^R$ est défini de manière unique, même dans le cadre de la grande dimension, $p \geq n$, dès que $\lambda > 0$. Il s'écrit sous la forme

$$\hat{\beta}^R(\lambda) = (X'X + \lambda \mathbf{I_p})^{-1} X'Y, \tag{2.6}$$

où $\mathbf{I_p}$ est la matrice identité de taille p. D'un point de vue *numérique*, on s'aperçoit que l'estimateur ridge, écrit sous la forme (2.6), est plus stable que l'EMC $\hat{\beta}^{MC}$. En effet, un

choix suffisamment grand du paramètre λ permet de rendre les valeurs propres de la matrice $X'X + \lambda \mathbf{I_p}$ assez grandes pour que l'inverse de cette matrice soit bien conditionné.
En revanche, d'un point de vue *statistique*, tout comme l'EMC, l'estimateur ridge ne permet pas d'identifier la paramètre β^* quand $p > n$.

Vers la sélection de variables

Le paramètre $\beta^* = (\beta_1^*, \cdots, \beta_p^*)$ traduit le poids des variables explicatives $(X_1, \cdots, X_p)$ sur la réponse Y. Lorsque le nombre de variables explicatives est important, un objectif serait d'évaluer la contribution de chaque variable et d'éliminer les variables non pertinentes. Ce type d'approche fournit des estimateurs interprétables. Dans ce contexte, l'estimateur ridge et l'EMC sont inefficaces. Il est utile de considérer des méthodes capables de sélectionner le sous-ensemble des variables explicatives, offrant une représentation quasi-complète de la réponse Y. Diverses stratégies ont été proposées dans ce but. Une approche classique est la *sélection de sous-ensembles* (*Subset Selection* en anglais). Soit $\mathcal{B}_k$ un sous-ensemble variables explicatives de taille k, où $k \in \{1, \ldots, p\}$ est un entier. Cette méthode trouve le meilleur sous-ensemble $\mathcal{B}_k$ qui réduit au maximum la somme des carrés des résidus (cf. Hastie, Tibshirani, et Friedman [74]).

Une seconde stratégie pour sélectionner les variables pertinentes est le *seuillage*. Nous utilisons dans ce cas un estimateur préliminaire (e.g., l'EMC lorsque $p \leq n$), que nous exploitons pour écarter certaines variables de l'étude. Une variable n'est sélectionnée que si l'estimation du coefficient de régression correspondant, obtenue grâce à l'estimateur préliminaire, dépasse un certain seuil défini par le statisticien. Parmi ces méthodes de seuillages, nous pouvons compter les procédures de seuillage dur et de seuillage doux, illustrées dans la Figure 2.1 (cf. Donoho et Johnstone [54, 53] et la section 2.2.1).

Pour réduire le nombre de variables explicatives, divers tests basés sur l'EMC ont été proposés pour tester la pertinence de chaque variable explicative X_j. Pour tout $j \in \{1, \ldots, p\}$, ces procédures testent sous l'hypothèse nulle $\beta_j^* = 0$ et sous l'alternative $\beta_j^* \neq 0$. Lorsque le bruit est gaussien, il est courant d'utiliser un test de Student ou encore un test de Fisher (cf. Tsybakov [138] et Hastie, Tibshirani, et Friedman [74]).
La sélection de variables par les tests a fait l'objet de nombreuses études sur l'identification du sous-ensemble des variables pertinentes. Ces méthodes de sélection de variables ne sont pas en lien direct avec celles considérées dans ce manuscrit, nous ne les détaillons donc pas davantage, nous référons le lecteur intéressé aux travaux de Birgé [19], Benjamini et Hoch-

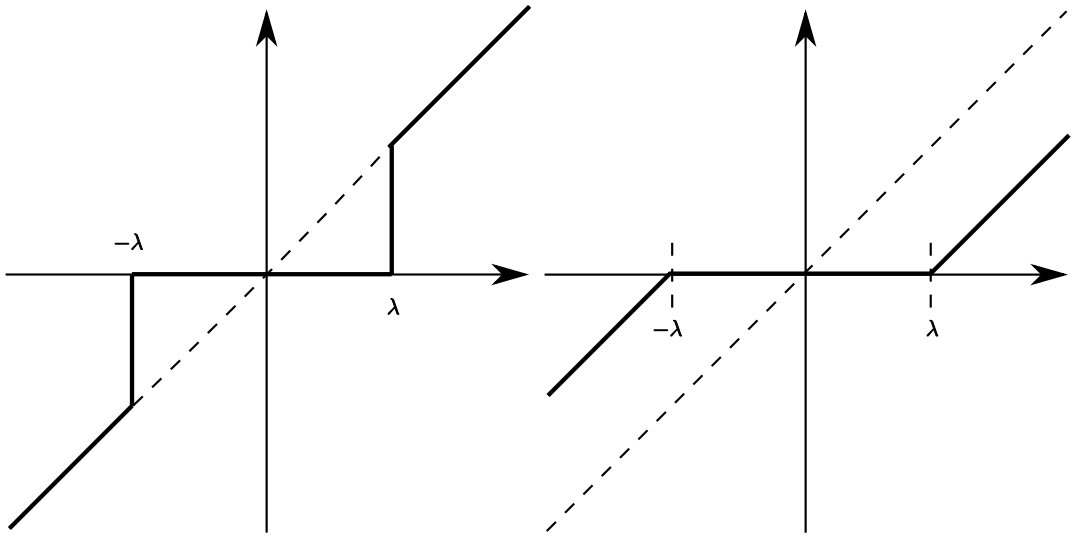

FIG. 2.1 – Fonctions de seuillage dur $v = u\mathbb{I}(u \geq \lambda)$ (à gauche) et de seuillage doux $v = \mathrm{sgn}(u)\,(|u| - \lambda)_+$ (à droite) pour $u \in \mathbb{R}$, où $(\cdot)_+$ dénote la partie positive, $\mathrm{sgn}(\cdot)$ est la fonction signe et $\lambda \geq 0$ est un paramètre de lissage qui détermine le seuil.

berg [16], Bauer, Pötscher, et Hackl [15], Pötscher [122] et Bunea, Wegkamp, et Auguste [31], issus de la littérature des tests multiples. Dans la prochaine section, nous présentons un M-estimateur important en sélection de variables.

2.1.3 Pénalisation ℓ_0

Notons $|\mathcal{A}|$ le cardinal de $\mathcal{A}$, un ensemble quelconque d'indice. Il est commode, pour l'étude de méthodes de sélection de variables, de définir l'*ensemble de sparsité* $\mathcal{A}^*$; c'est l'objet de la Définition 2.1.

Définition 2.1.

Soit le modèle de régression linéaire (2.1)*. On définit l'ensemble de sparsité associé au vecteur* β^* *par*

$$\mathcal{A}^* = \{j : \beta_j^* \neq 0\}. \tag{2.7}$$

On définit par la suite, l'indice de sparsité de β^* *par la quantité* $|\mathcal{A}^*|$*.*

La construction d'estimateurs interprétables est un enjeu important. Des estimateurs parmi ceux qui répondent à cette attente sont par exemple construits à partir de la pénalité ℓ_0, tels que les critères d'information $\mathrm{C_p}$ de Mallows, AIC (Akaike Information Criterion) ou encore BIC (Bayesian Information Criterion), aujourd'hui classiques et introduits respectivement par Mallows [105], Akaike [2] et Schwartz [127]. Ces critères sélectionnent parmi une collection de taille D d'estimateurs de β^*, notée

$$\widehat{\mathcal{F}} = \left\{\hat{\beta}_1, \cdots, \hat{\beta}_D\right\},$$

celui qui remplit au mieux le double objectif suivant : la bonne estimation de $X\beta^*$ et la bonne estimation de l'ensemble des variables pertinentes $\mathcal{A}^*$ définie en (2.7). On comprendra aisément l'importance du choix de cette famille $\widehat{\mathcal{F}}$. Ces critères sont construits à partir de la pénalité $\mathrm{pen}(\beta, n)$ qui fait intervenir la semi-norme ℓ_0 du vecteur β

$$\mathrm{pen}(\beta, n) = \lambda_n \|\beta\|_0, \tag{2.8}$$

où $\|\beta\|_0$ est définie par $\|\beta\|_0 = \sum_{j=1}^{p} \mathbb{I}(\beta_j \neq 0)$ et $\mathbb{I}(\cdot)$ dénote la fonction indicatrice. La définition et la valeur de λ_n sont propres au critère considéré. Dans le cadre de la régression linéaire (2.1) que nous considérons et sous l'hypothèse de bruit gaussien, les critères $\mathrm{C_p}$ et AIC sont confondus et la pénalité est définie comme

$$\mathrm{pen}^{\mathrm{AIC}}(\beta, n) = \mathrm{pen}^{\mathrm{C_P}}(\beta, n) := \frac{2\sigma^2 \|\beta\|_0}{n}. \tag{2.9}$$

Soit $\hat{\beta}$ un estimateur de β^*, définissons tout d'abord :

$$\mathrm{df}(\hat{\beta}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathrm{Cov}(x_i \hat{\beta}, y_i),$$

où Cov désigne la covariance sous la loi de ε. La quantité $\mathrm{df}(\hat{\beta})$ est le *degré de liberté* de l'estimateur $\hat{\beta}$ (cf. Efron [58] et Hastie et Tibshirani [75]). Le degré de liberté d'un estimateur est lié à la dimension de l'espace engendré par les variables explicatives X_j qui interviennent dans sa construction. Sous l'hypothèse du bruit gaussien, un point important est l'équation qui lie l'erreur de prédiction au degré de liberté $\mathrm{df}(\hat{\beta})$ d'un estimateur $\hat{\beta}$:

$$\mathbb{E}\left[\|X\hat{\beta} - X\beta^*\|_n^2\right] + \sigma^2 = \mathbb{E}\left[\|Y - X\hat{\beta}\|_n^2\right] + \frac{2\sigma^2}{n}\,\mathrm{df}(\hat{\beta}), \tag{2.10}$$

où $\mathbb{E}$ dénote l'espérance sous la loi de ε. On remarque que minimiser l'erreur de prédiction est équivalent à minimiser le membre de droite dans l'équation (2.10). Un estimateur de $\mathrm{df}(\hat{\beta})$ est donné par $\|\hat{\beta}\|_0$ (Mallows [105] et Hastie et Tibshirani [75]). De par sa définition, l'estimateur $\mathrm{C_p}$ et AIC, noté $\bar{\beta}$, est tel que

$$\|Y - X\bar{\beta}\|_n^2 + \frac{2\sigma^2 \|\bar{\beta}\|_0}{n}$$

donne une bonne approximation de l'erreur de prédiction. Les critères $\mathrm{C_p}$ et AIC fournissent des estimateurs performants du point de vue de la prévision de $X\beta^*$.

Le critère BIC est plus approprié pour l'estimation de l'ensemble $\mathcal{A}^*$ des composantes non nulles de β^*, en ce sens qu'il *surpénalise* les gros modèles ; i.e. ceux dont l'ensemble $\mathcal{A}^*$ associé est de cardinal grand. Le critère BIC est défini par (2.11) comme

$$\mathrm{pen}^{\mathrm{BIC}}(\beta, n) := \sigma^2 \log(p) \frac{\|\beta\|_0}{n}. \tag{2.11}$$

La pénalité BIC impose une plus forte contrainte aux estimateurs dont beaucoup de composantes sont différentes de zéro, dès que $p \geq 7$. En d'autres termes, le critère BIC tend à sélectionner des estimateurs plus parcimonieux (*sparse* en anglais) que ceux choisis par les critères C_p et AIC.

Notons que les performances théoriques de ces critères diffèrent selon la finalité de l'étude. L'estimateur BIC fournit en général de meilleures performances théoriques dans le cadre de l'estimation du support $\mathcal{A}^*$ de β^* (on parle de *sélection*). A l'inverse, les estimateurs C_p et AIC donnent de meilleurs résultats pour l'estimation de $X\beta^*$ (on parle de *prédiction*). Ces remarques ont fait l'objet de nombreux travaux ; citons ici Yang [156], Shibata [132], Foster et George [65], Leeb et Pötscher [97], McQuarrie et Tsai [108], Shao [129]. Yang [157] a montré qu'il est toutefois possible de combiner, dans certains cas, les avantages de ces différents critères.

Les critères bâtis sur la pénalité ℓ_0 ont été largement étudiés dans la littérature. Pour p fixé et n tendant vers l'infini, les critères C_p et AIC fournissent des estimateurs consistants au sens ℓ_2 si l'on considère l'erreur de prédiction. Les performances théoriques de ces estimateurs ont été établies par Shibata [131], Li [100], Polyak et Tsybakov [121], Baraud [13] et par Birgé et Massart [21] au prix d'un contrôle sur le cardinal D de la famille $\widehat{\mathcal{F}}$ considérée. Lorsque la finalité de l'étude est l'estimation du support $\mathcal{A}^*$, Haughton [76] et Guyon et Yao [72] ont obtenu de bons résultats théoriques en utilisant le critère BIC. Dans le modèle de régression semi-paramétrique, Bunea [27] montre que le critère BIC sélectionne le bon sous-ensemble des variables pertinentes avec une probabilité qui tend vers 1, lorsque n tend vers l'infini, en adaptant la pénalité au cadre semi-paramétrique par l'ajout d'un terme de correction.

Plus récemment et en grande dimension ($p \geq n$) et n fini, Barron, Birgé, et Massart [14], Birgé et Massart [21], Bunea [27] et Massart [106] se sont intéressés au contrôle non-asymptotique de l'erreur de prédiction d'estimateurs de type ℓ_0. Les auteurs montrent que des estimateurs définis avec une pénalité légèrement différente de (2.8) satisfont des inégalités pour l'erreur de prédiction dépendant de la dimension p de manière seulement logarithmique. Un point intéressant de ces travaux est qu'aucune hypothèse sur la matrice de *Gram*, $\Psi^n = \frac{X'X}{n}$, n'est nécessaire. En contrepartie, une hypothèse sur la taille D de la famille $\widehat{\mathcal{F}}$ doit être faite. Toutefois, Bunea, Tsybakov, et Wegkamp [32] et Birgé et Massart [20] se sont affranchis de cette hypothèse et ont obtenu des résultats similaires.

Inconvénient de ces critères. Les procédures définies à partir de la pénalité ℓ_0 aboutissent à de bonnes performances théoriques en s'affranchissant de toute hypothèse sur la matrice de Gram, mais sous la condition que le cardinal D de la famille $\widehat{\mathcal{F}}$ considérée soit suffisamment grand. La complexité algorithmique devient trop grande dès lors que D est grand, ou que p est grand. Des hypothèses sur le modèle sont alors nécessaires. Les hypothèses rencontrées dans la littérature sont du type : "supposer les variables ordonnées", ou encore "considérer une famille $\widehat{\mathcal{F}}$ d'estimateurs *sparses*".

2.1.4 Hypothèse fondamentale

Les méthodes de sélection de variables considérées précédemment ont été introduites dans un cadre classique d'estimation, où la dimension p est raisonnable ($p \leq n$). Dans ce manuscrit, nous considérons des problèmes en grande dimension ($p > n$). L'étude peut avoir différentes finalités : l'estimation de $X\beta^*$, de β^*, ou encore du support $\mathcal{A}^*$ de β^*. Les estimateurs présentés dans la section 2.1.2 (EMC et estimateur ridge) ne sont pas adaptés dans ce contexte. Pour répondre à ces difficultés, il est nécessaire d'exhiber de nouveaux estimateurs. Pour cela, nous introduisons dans cette section une notion fondamentale : la *sparsité*.

Hypothèse 2.1. ***Hypothèse (S) de sparsité.***
Soit $\mathcal{A}^*$ l'ensemble de sparsité défini en (2.7) *associé à β^*. Il existe un entier s, tel que*

$$|\mathcal{A}^*| \leq s, \tag{2.12}$$

et on suppose que $s = o(n)$.

L'hypothèse (S) traduit un contrôle sur la sparsité du modèle. Elle témoigne du fait qu'il y a peu de variables explicatives pertinentes pour l'étude.

Face à des problèmes *de grande dimension* (i.e., $p \geq n$), il s'avère nécessaire d'ajouter certaines hypothèses sur le modèle que nous considérons. Compte tenu des applications réelles, l'hypothèse (S) de sparsité est la plus adaptée et la plus naturelle. Celle-ci assure que seules très peu de coordonnées β_j^* du vrai vecteur de régression β^* sont différentes de zéro.

Estimateur des moindres carrés restreint

Dans le cadre idéal où l'ensemble de sparsité $\mathcal{A}^*$ est connu, un estimateur comblant nos attentes dans le contexte de la grande dimension est l'estimateur des moindres carrés restreint, noté $\hat{\beta}^{MCR}$. L'estimateur $\hat{\beta}^{MCR}$ est défini comme l'estimateur des moindres carrés,

restreint aux variables pertinentes X_j, i.e., les variables X_j dont les indices j sont dans $\mathcal{A}^*$. Introduisons la matrice $X_{\mathcal{A}^*}$, de taille $n \times |\mathcal{A}^*|$, qui désigne la restriction de la matrice X aux colonnes X_j avec $j \in \mathcal{A}^*$. L'estimateur $\hat{\beta}^{MCR}$ est de la forme :

$$\hat{\beta}^{MCR} = (X'_{\mathcal{A}^*} X_{\mathcal{A}^*})^{-1} X'_{\mathcal{A}^*} Y.$$

Si nous évaluons à présent l'erreur de prédiction commise par l'estimateur $\hat{\beta}^{MCR}$, il vient :

$$\mathbb{E}\left[\|X\hat{\beta}^{MCR} - X\beta^*\|_n^2\right] = \sigma^2 \frac{|\mathcal{A}^*|}{n}. \tag{2.13}$$

Si l'hypothèse de sparsité est satisfaite, l'estimateur $\hat{\beta}^{MCR}$ est consistant, son erreur ne dépendant pas de la dimension *effective* p du problème, mais de la dimension *réelle* $|\mathcal{A}^*|$. En pratique, l'ensemble de sparsité est rarement connu, donc l'estimateur $\hat{\beta}^{MCR}$ n'est pas exploitable en l'état. D'autres procédures qui exploitent la sparsité de β^* ont été étudiées. Dans la prochaine section, nous présentons une de ces méthodes, très populaire en sélection de variables : le LASSO.

2.2 Estimateur LASSO

Les problèmes de régression en grande dimension ont suscité la construction d'estimateurs à la fois fiables et facilement interprétables. Nous présentons dans cette section un exemple d'estimateur sparse : l'estimateur LASSO[1] (Least Absolute Shrinkage and Selection Operator). La procédure Lasso est la méthode de référence pour cette thèse.

2.2.1 Introduction de l'estimateur

Introduit pour la première fois sous l'appellation *Lasso* par Tibshirani [135], cet estimateur est défini comme l'estimateur des moindres carrés sous une contrainte de type ℓ_1 :

$$\bar{\beta}^L(t) = \begin{cases} \operatorname{argmin}_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_n^2, \\ \\ \text{s.c. } \|\beta\|_1 \leq t, \end{cases} \tag{2.14}$$

où t est un paramètre positif. Cet estimateur avait déjà été introduit en théorie du signal par Chen et Donoho [42] (voir également Chen, Donoho, et Saunders [43]), sous le nom de *Basis Pursuit De-Noising* et défini sous sa forme pénalisée :

$$\hat{\beta}^L(\lambda) \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|Y - X\beta\|_n^2 + \lambda \|\beta\|_1. \tag{2.15}$$

Une large discussion sur le Basis Pursuit De-Noising peut être trouvée dans Mallat [104, chapitre 12]. Notons que les estimateurs $\bar{\beta}^L$ et $\hat{\beta}^L$ sont équivalents point par point. A savoir,

[1] On notera dorénavant cet estimateur Lasso.

pour un t fixé, on peut trouver un λ (qui dépend des données) tel que $\bar{\beta}^L = \hat{\beta}^L$. Et inversement, pour un λ fixé, il existe un t tel que $\bar{\beta}^L = \hat{\beta}^L$ et vaut dans ce cas $t = \sum_{j=1}^{p} |\hat{\beta}_j^L(\lambda)|$. Dans la suite de ce manuscrit, nous appellerons ces deux estimateurs : estimateurs *Lasso*, $\hat{\beta}^L$.

Une forme duale de l'estimateur Lasso (2.15) a été proposée par Osborne, Presnell, et Turlach [118] et Alquier [4]. Elle se formule de la façon suivante :

$$\tilde{\beta}^L(\lambda) \in \begin{cases} \operatorname{argmin}_{\beta \in \mathbb{R}^p} \|X\beta\|_2^2 \\ \\ \text{s.c. } \left\|\Xi^{-1} X' (Y - X\beta)\right\|_\infty \leq \frac{n\lambda}{2}. \end{cases} \tag{2.16}$$

où Ξ est la matrice diagonale de taille p, dont le j-ième coefficient diagonal vaut $\|X_j\|_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_{i,j}^2}$. Une propriété importante de l'estimateur $\tilde{\beta}^L$ est qu'il est confondu avec l'estimateur Lasso ($\hat{\beta}^L = \tilde{\beta}^L$), dès lors que la matrice Ψ^n est inversible. Cela n'est plus vrai lorsque $p > n$, on a seulement $X\tilde{\beta}^L(\lambda) = X\hat{\beta}^L(\lambda)$.
La forme duale (2.16) de l'estimateur Lasso nous fournit une interprétation géométrique de celui-ci. En effet, Alquier [4] suggère que le Lasso peut être vu comme le projeté de $\mathbf{0_p}$, le zéro de $\mathbb{R}^p$, sur la région de confiance déterminée par la contrainte

$$\left\|\Xi^{-1} X' (Y - X\beta)\right\|_\infty \leq \frac{n\lambda}{2}.$$

Quelques remarques sur le Lasso.

– Par le biais de la pénalisation ℓ_1, nous pouvons établir des liens entre le Lasso et d'autres méthodes statistiques. Citons par exemple le *boosting sous contrainte* ℓ_1 (Friedman [68], Lugosi et Vayatis [103] et Bühlmann et Yu [25]), généralement utilisé en classification. Citons également le *Basis Pursuit* introduit pour la reconstitution d'un signal non bruité, i.e., $\varepsilon = \mathbf{0_n}$ dans le modèle de régression (2.2), où $\mathbf{0_n}$ est le zéro dans $\mathbb{R}^n$ (Chen et Donoho [42] et Chen, Donoho, et Saunders [43]).

– Notons que l'estimateur Lasso et l'estimateur des moindres carrés sont confondus ($\hat{\beta}^L = \hat{\beta}^{MC}$) pour $\lambda = 0$, i.e., la procédure Lasso sélectionne toutes les variables explicatives. En revanche, si $\lambda \to \infty$, la procédure Lasso ne sélectionne aucune variable explicative ($\hat{\beta}^L = \mathbf{0_p}$). Entre ces deux valeurs extrêmes, la procédure Lasso sélectionne d'autant plus de variables explicatives que λ est petit, i.e., plus λ est grand, plus la contrainte sur les coefficients de β l'est également.

– L'estimateur Lasso peut s'interpréter en terme d'estimation dans le cadre de l'inférence bayésienne. En effet, nous pouvons déduire l'estimateur Lasso en considérant le modèle de régression avec bruit gaussien, et en prenant comme loi a priori pour le paramètre β^* la *loi*

de Laplace qui est également appelée *loi double exponentielle*; i.e., β_j^* admet pour densité a priori par rapport à la mesure de Lebesgue $\frac{\lambda}{2}\exp\Big(-\lambda|\beta_j^*|\Big)$.

– L'estimateur Lasso est linéaire par morceaux comme fonction de λ.

• Lorsque les colonnes de la matrice X sont orthogonales (ou le cas trivial où $p=1$), résoudre le problème Lasso est équivalent à trouver la solution de p problèmes de seuillage doux (Donoho et Johnstone [53, 54]). Plus précisément, chaque composante du vecteur $\hat{\beta}^L = (\hat{\beta}_1^L, \ldots, \hat{\beta}_p^L)'$ s'écrit en fonction de l'estimateur des moindres carrés $\hat{\beta}^{MC} = (X'X)^{-1}X'Y$:

$$\hat{\beta}_j^L(\lambda) = \mathrm{sgn}(\hat{\beta}_j^{MC})(|\hat{\beta}_j^{MC}| - \lambda/2)_+, \ \forall j \in \{1, \ldots, p\}, \tag{2.17}$$

où $(\alpha)_+ = \max\{\alpha, 0\}$ pour α un réel et la fonction signe $\mathrm{sgn}(\cdot)$ est définie pour tout réel b comme

$$\mathrm{sgn}(b) = \begin{cases} 1 & \text{si } b > 0, \\ 0 & \text{si } b = 0, \\ -1 & \text{si } b < 0. \end{cases} \tag{2.18}$$

L'optimalité de cette méthode a été assurée dans le cadre du seuillage d'ondelettes par Donoho, Johnstone, Kerkyacharian, et Picard [56]. Sous la forme (2.17), l'estimateur Lasso est linéaire par morceaux comme fonction de λ. On peut sans peine trouver p valeurs du paramètre de régularisation $\{\lambda^{(1)}, \ldots, \lambda^{(p)}\}$ telles que la linéarité change. Dans ce cas $\lambda^{(k)} = |\hat{\beta}_k^{MC}|$, $k = 1, \ldots, p$.

• Dans le cadre plus général où X n'est pas une matrice orthogonale (et $p \neq 1$), la linéarité par morceaux de la solution Lasso comme fonction de λ a été établie par Efron, Hastie, Johnstone, et Tibshirani [61] et Zou, Hastie, et Tibshirani [169]. De plus, les valeurs du paramètre de régularisation telles que la linéarité change peuvent être trouvées en considérant les conditions d'optimalité du premier ordre (2.15). Ces conditions sont également connues sous le nom de conditions *Karush-Kuhn-Tucker (KKT)*.
Définissons l'*ensemble de sparsité ou ensemble actif* d'un estimateur $\hat{\beta}$ par

$$\hat{\mathcal{A}} = \{j \in \{1, \ldots, p\} : \hat{\beta}_j \neq 0\}. \tag{2.19}$$

Par suite, notons $\hat{\mathcal{A}}_L = \hat{\mathcal{A}}_L(\lambda)$, l'ensemble actif de l'estimateur Lasso $\hat{\beta}^L(\lambda)$. Les conditions KKT pour cet estimateur s'écrivent :

$$\begin{cases} 2X_j'\Big(Y - X\hat{\beta}^L(\lambda)\Big) = \lambda\,\mathrm{sgn}(\hat{\beta}_j^L) & \forall j \in \hat{\mathcal{A}}_L, \\ 2\Big|X_j'\Big(Y - X\hat{\beta}^L(\lambda)\Big)\Big| < \lambda & \forall j \notin \hat{\mathcal{A}}_L. \end{cases} \tag{2.20}$$

Ces conditions peuvent se reformuler de la façon suivante

$$\Big\|X'\Big(Y - X\hat{\beta}^L(\lambda)\Big)\Big\|_\infty \leq \lambda/2,$$

où $\|a\|_\infty = \sup_j a_j$ est la norme-sup du vecteur $a \in \mathbb{R}^p$. De manière analogue à l'estimateur de seuillage doux, nous pouvons définir l'ensemble des valeurs du paramètre de régularisation $\{\lambda^{(1)}, \ldots, \lambda^{(K)}\}$ pour lequel la linéarité de $\hat{\beta}^L(\lambda)$ change, et K est un entier positif. Nous pouvons trouver ces valeurs particulières de λ en considérant les conditions non saturées dans les conditions KKT (i.e., la deuxième ligne de (2.20)). En effet, pour tout $j \in \{1, \ldots, p\}$, la quantité $X_j' \left(Y - X\hat{\beta}^L(\lambda)\right)$ évolue en fonction de λ. En particulier, pour tout $j \notin \hat{\mathcal{A}}_L$, il existe un $\lambda^{(k)}$ tel que $\left|X_j' \left(Y - X\hat{\beta}^L(\lambda^{(k)})\right)\right| = \lambda^{(k)}/2$. Pour la valeur $\lambda = \lambda^{(k)}$, la contrainte correspondant à la variable X_j devient donc saturée et la variable X_j devient active (i.e., l'indice j "passe" de $\hat{\mathcal{A}}_L^C$ à $\hat{\mathcal{A}}_L$).

– Un avantage non négligeable de l'estimateur Lasso est qu'une solution approchée peut être obtenue rapidemment en utilisant, par exemple, l'algorithme LARS introduit par Efron, Hastie, Johnstone, et Tibshirani [61].

Remarque 2.1.
Dans l'annexe B, nous considérons l'estimateur Lasso d'un point de vue algorithmique. En particulier, nous détaillons un algorithme d'homotopie appelé le LARS, permettant de calculer l'ensemble des solutions Lasso $\hat{\beta}^L(\lambda)$ lorsque λ varie dans $[0, \infty[$. L'algorithme base sa construction sur les conditions KKT , i.e., sur les corrélations entre le résidu $\left(Y - X\hat{\beta}^L(\lambda)\right)$ et les variables explicatives $(X_1, \ldots, X_p)$. Comme la majeure partie des algorithmes utilisés pour approximer les solutions Lasso, le LARS échoue à construire de bonnes évaluations de l'estimateur lorsque de fortes corrélations interviennent entre les variables. L'algorithme tend à sélectionner une variable explicative par groupe de variables corrélées, écartant ainsi de l'étude les autres variables du groupe.

2.2.2 Nature des résultats

Dans le cadre de la régression linéaire en grande dimension, avec le paramètre β^* vérifiant l'hypothèse (S) de sparsité (2.12), trois objectifs naturels peuvent être définis :
– fournir la meilleure approximation du vecteur $X\beta^*$: on parle de **prédiction**,
– estimer le vecteur β^* : on parle d'**estimation**,
– identifier le support $\mathcal{A}^*$ de β^* défini en (2.7), ou encore le vecteur signe de β^* : on parle de **sélection**.

Dans ce cadre, les résultats théoriques espérés sont du type *Inégalité de Sparsité*, notée IS dans la suite (Définition 2.2). Les IS ont été introduites par Bunea, Tsybakov, et Wegkamp [33] dans le cadre de la régression non-paramétrique (i.e., le paramètre estimé est un paramètre fonctionnel f) sous le nom de *Sparsity Oracle Inequalities*, les inégalités faisant

intervenir un indice de sparsité dans la borne.

Définition 2.2.
*Soit l'***hypothèse (S)** *de sparsité vérifiée et s l'indice défini en* (2.12)*. On dit qu'un estimateur $\hat{\beta}$ vérifie une Inégalité de Sparsité (IS)*

$$\begin{aligned}
&\underline{\textit{en prédiction}}\ \textit{si :} && \mathbb{E}\left[\|X(\hat{\beta}-\beta^*)\|_n^2\right] \le Cs\frac{L(n,p)}{n};\\
&\underline{\textit{en estimation en norme } \ell_1}\ \textit{si :} && \mathbb{E}\left[\|\hat{\beta}-\beta^*\|_1\right] \le Cs\sqrt{\frac{L(n,p)}{n}};\\
&\underline{\textit{en estimation en norme-}\sup}\ \textit{si :} && \mathbb{E}\left[\|\hat{\beta}-\beta^*\|_\infty\right] \le C\sqrt{\frac{L(n,p)}{n}},
\end{aligned}$$

où C est une constante positive et $L(n,p)$ est un terme décrivant une dépendance au plus logarithmique en n et p. Toutes les IS peuvent être également définies à l'aide d'inégalités probabilistes.

Dans le cadre de la *sélection* du support $\mathcal{A}^*$, il est d'usage d'utiliser l'IS en norme-sup. Dans la Définition 2.2, l'IS en *estimation* est définie à l'aide de la norme ℓ_1, mais d'autres normes ℓ_q tel que $q \ge 1$ ont été considérées dans la littérature. Notons que la notion d'IS est particulièrement intéressante lorsque nous considérons des problèmes en grande dimension ($p \ge n$). En effet, dans de telles situations, la dépendance en la dimension $p = p(n)$ est un élément majeur à considérer.

Apportons à présent quelques précisions sur l'ordre de grandeur de la quantité $L(n,p)$ introduite dans la Définition 2.2.

– Un cadre plus général que celui que nous avons supposé dans ce manuscrit, est celui de la régression non-paramétrique à *design* fixe, où f est le paramètre fonctionnel inconnu. Une approche possible est l'aggrégation dans laquelle l'estimateur $\hat{f}$ de f est défini en agrégeant plusieurs fonctions d'un dictionnaire $\{f_1 \dots, f_p\}$ donné. Ce problème a largement été étudié ces dernières années (Birgé [19], Catoni [38], Györfi, Kohler, Krzyżak, et Walk [73], Juditsky et Nemirovski [86], Tsybakov [139], Wegkamp [154] et Yang [158]). D'un point de vue minimax, Bunea, Tsybakov, et Wegkamp [32, Théorème 5.1] et Bunea, Tsybakov, et Wegkamp [29] ont étudié les bornes inférieures de l'erreur de prédiction et ont montré que la *vitesse optimale* φ_n de convergence pour ce problème est

$$\varphi_n = \frac{|\mathcal{A}^*|}{n}\log\left(\frac{p}{|\mathcal{A}^*|}+1\right). \tag{2.21}$$

On constate une dépendance en l'indice de sparsité $|\mathcal{A}^*|$ et un terme logarithmique.

– L'EMC restreint est construit en supposant connu l'ensemble de sparsité $\mathcal{A}^*$ et son erreur de prédiction est de l'ordre $\frac{|\mathcal{A}^*|}{n}$. Dans notre cadre, cet ensemble est toujours inconnu. Ainsi, au vu de (2.21), un facteur au minimum logarithmique est à prévoir dans l'erreur de prédiction d'un estimateur $\hat{\beta}$ qui s'affranchit de la connaissance de $\mathcal{A}^*$. Les bornes des IS des estimateurs $\hat{\beta}$ qui s'adaptent au paramètre de sparsité inconnu $\mathcal{A}^*$ sont à comparer avec (2.21) comme borne optimale.

Les IS sont généralement difficiles à obtenir et il est nécessaire d'ajouter une hypothèse, souvent restrictive, sur la matrice de Gram $\Psi^n = \frac{X'X}{n}$. Dans la littérature, un "bon" estimateur est un estimateur qui répond au mieux aux exigences suivantes :

- *Optimalité :* Les résultats d'optimalité sont de deux types. D'un point de vue asymptotique ($n \to \infty$), on s'intéresse à la consistance de la procédure d'estimation. D'un point de vue non-asymptotique, on recherche des IS du même type que celles décrites dans la Définition 2.2.

- *Hypothèses :* Les hypothèses, en particulier sur la matrice de Gram Ψ^n, doivent être aussi faibles que possible.

- *Complexité algorithmique :* Outre les aspects théoriques, il est important qu'une méthode d'estimation soit utilisable en pratique ; i.e., implémentable en un temps raisonnable même lorsque p est plus grand que n.

- *Interprétabilité :* Dans les problèmes en *grande dimension*, il est souhaitable que l'estimateur construit soit *sparse*. De cette façon, il devient possible d'extraire l'information utile et d'interpréter les résultats.

2.3 Résultats théoriques pour l'estimateur Lasso

De nombreux résultats théoriques ont été obtenus ces dernières années dans la littérature statistique. Nous discutons certains d'entre eux dans les sections 2.3.2 et 2.3.3. Ces résultats ont été obtenus au prix d'hypothèses plus ou moins contraignantes, notamment sur la matrice de Gram Ψ^n. C'est l'objet de cette première section 2.3.1.

2.3.1 Hypothèses sur la matrice de Gram

Les hypothèses qui ont été faites sur la matrice de Gram Ψ^n n'autorisent qu'une faible corrélation entre les variables. Nous détaillons dans cette section, les principales hypothèses

imposées à la matrice Ψ^n. Au préalable, introduisons quelques notations. Étant donnée la matrice de Gram

$$\Psi^n = \left[\Psi^n_{j,k}\right]_{1\leq j,k\leq p} = \frac{X'X}{n},$$

et $\mathcal{A}$ un ensemble d'indice, on note

$$\Psi^n_{\mathcal{A},\mathcal{A}} = \frac{X'_{\mathcal{A}}X_{\mathcal{A}}}{n},$$

la restriction de Ψ^n aux lignes et colonnes dont les indices appartiennent à l'ensemble $\mathcal{A}$. De plus, on introduit

$$\Psi^n_{\mathcal{A}^C,\mathcal{A}} = \frac{X'_{\mathcal{A}^C}X_{\mathcal{A}}}{n},$$

la matrice de taille $|\mathcal{A}^C| \times |\mathcal{A}|$ (où $|B|$ est le cardinal de l'ensemble B), restriction de Ψ^n aux lignes dont les indices appartiennent à l'ensemble $\mathcal{A}^C$ et aux colonnes dont les indices appartiennent à l'ensemble $\mathcal{A}$.

Soit un entier m tel que $m \leq p$, on définit les valeurs propres minimale et maximale m-sparses de Ψ^n respectivement par

$$\phi_{\min}(m) = \min_{u\in\mathbb{R}^p:\, \|u\|_0\leq m} \frac{u'\Psi^n u}{u'u}, \quad \text{et} \quad \phi_{\max}(m) = \max_{u\in\mathbb{R}^p:\, \|u\|_0\leq m} \frac{u'\Psi^n u}{u'u}. \tag{2.22}$$

La notion de valeurs propres m-sparse a été introduite dans le cadre de la sélection de variables par Donoho [52]. Ces quantités sont les valeurs propres minimale et maximale sur l'ensemble des sous-matrices de la matrice Ψ^n, de tailles inférieures ou égales à $m \times m$. Dans le cas $p > n$, la valeur propre minimum $\phi_{\min}(m)$ est nulle pour tout $m > n$.

Présentons à présent, les hypothèses communes concernant la matrice de Gram. Elles traduisent une faible corrélation entres les variables pertinentes et les autres.

Hypothèse 2.2. Cohérence Mutuelle (CM).
Il existe une constante $C > 0$ telle que

$$\max_{k\neq j} |\Psi^n_{j,k}| \leq \frac{C}{s}, \tag{2.23}$$

où s, défini en (2.12)*, est tel que $|\mathcal{A}^*| \leq s$.*

Dans la littérature sur le Lasso, cette hypothèse est en général utilisée sous une version moins restrictive. Dans cette version, seules les corrélations des variables X_j pertinentes avec les autres variables sont restreintes. L'hypothèse s'écrit alors :
il existe une constante $C > 0$ telle que

$$\max_{j\in\mathcal{A}^*} \max_{\substack{k\in\{1,\dots,p\}\\ k\neq j}} |\Psi^n_{j,k}| \leq \frac{C}{|\mathcal{A}^*|}. \tag{2.24}$$

L'hypothèse (CM), sous sa forme la moins restrictive (2.24), est entre autres utilisée par Bunea [28], Bunea, Tsybakov, et Wegkamp [33] et Bunea, Tsybakov, et Wegkamp [32]. Notons que dans sa version forte (2.23), cette condition de cohérence mutuelle implique l'unicité du vecteur β^* lorsque l'hypothèse (S) de sparsité est vérifiée. Le paramètre β^* est alors identifiable (Lounici [101]).

Hypothèse 2.3. Condition d'Irreprésentabilité (CI).
Il existe une constante $C > 0$ telle que

$$\|\Psi^n_{(\mathcal{A}^*)^C,\mathcal{A}^*}(\Psi^n_{\mathcal{A}^*,\mathcal{A}^*})^{-1}\operatorname{Sgn}(\beta^*_{\mathcal{A}^*})\|_\infty \leq 1 - C, \tag{2.25}$$

où $\|\cdot\|_\infty$ dénote la norme sup dans $\mathbb{R}^p$ et la fonction signe Sgn est définie $\forall x \in \mathbb{R}^p$ par

$$\operatorname{Sgn}(x) = (\operatorname{sgn}(x_1), \ldots, \operatorname{sgn}(x_p))',$$

et la fonction $\operatorname{sgn}(\cdot)$ *est définie en* (2.18).

L'hypothèse (CI) a été introduite dans le modèle de regression linéaire par Yuan et Lin [160], Zhao et Yu [164] et Zou [166] et dans le modèle graphique gaussien par Meinshausen et Bühlmann [112] (cf. Lauritzen [95] pour des détails sur le modèle graphique gaussien). Cette condition d'Irreprésentabilité (CI) est appellée *condition de stabilisation de voisinage* par Meinshausen et Bühlmann [112]. Une interprétation claire de cette hypothèse demeure toutefois difficile, puisqu'elle fait intervenir le signe du vecteur β^* dans son énoncé (2.25).

Hypothèse 2.4. Valeurs Propres Restreintes (VPR).
Soit l'ensemble $\Delta(\mathcal{A})$ défini tel que

$$\Delta(\mathcal{A}) = \left\{ \alpha \in (\mathbb{R}^*)^p \,:\, \sum_{j\in\mathcal{A}^C} |\alpha_j| \leq 3 \sum_{j\in\mathcal{A}} |\alpha_j| \right\}, \tag{2.26}$$

il existe une constante $\kappa_1(s) > 0$ telle que

$$\kappa_1(s) = \min_{\mathcal{A}\subset\{1,\ldots,p\}:\,|\mathcal{A}|\leq s} \quad \min_{\alpha\in\Delta(\mathcal{A})} \quad \frac{\alpha'\Psi^n\alpha}{\sum_{j\in\mathcal{A}}\alpha_j^2} > 0, \tag{2.27}$$

où s est défini en (2.12).

Notons que l'hypothèse (VPR) introduite par Bickel, Ritov, et Tsybakov [17], est une version plus faible, i.e. moins contraignante que l'hypothèse (CM). Une large discussion autour de cette hypothèse se trouve dans le papier de Bickel, Ritov, et Tsybakov [17]. Il y est établi que si l'hypothèse (CM) est vérifiée pour la constante $C \geq \frac{\phi_{\min}(s)}{6}$ alors l'hypothèse (VRP) est satisfaite, où $\phi_{\min}(s)$ est la valeur propre minimale m-sparse donnée par (2.22). De même, l'hypothèse (CM) dans sa version faible (2.24) implique l'hypothèse (VPR) dès lors que l'on restreint les sommes aux coordonnées dans $\mathcal{A}^*$, dans cette dernière condition.

Hypothèse 2.5. Designs Incohérents (DI)).
Il existe une suite de réels positifs e_n, appelée séquence de multiplicateurs de sparsité, telle que

$$\liminf_{n\to\infty} \frac{e_n \phi_{\min}(e_n^2 s)}{\phi_{\max}(s + \min\{n, p\})} \geq C, \tag{2.28}$$

où $C > 0$ est une constante et s est défini en (2.12).

Cette hypothèse (DI) est une hypothèse sur la décroissance de la valeur propre minimale m-sparse (2.22). Elle a été introduite par Meinshausen et Yu [114] et sous le nom de *condition sparse de Riesz* par Zhang et Huang [163]. Sous certaines conditions, Bickel, Ritov, et Tsybakov [17] montrent que l'hypothèse (VRP) est moins restrictive que l'hypothèse (DI) (voir Bickel, Ritov, et Tsybakov [17, Assumption 2]). Pour cela, ils introduisent une hypothèse du type (DI) moins restrictive. Ils supposent que :

$$m\phi_{\min}(s + m) > \overline{c} s \phi_{\max}(m), \tag{2.29}$$

où $\overline{c} > 0$ est une constante, $1 \leq s \leq p/2$, $m \geq s$ et $s + m \leq p$.
Pour $s + m = s \log n$, les auteurs montrent alors que l'hypothèse 2.29 est moins restrictive que l'hypothèse (DI). De plus, ils prouvent que l'hypothèse 2.29 implique l'hypothèse (VPR) pour $\kappa_1(s)$ telle que

$$\kappa_1(s) = \sqrt{\phi_{\min}(s + m)} \left(1 - \overline{c} \sqrt{\frac{s \phi_{\max}(m)}{m \phi_{\min}(s + m)}} \right).$$

2.3.2 Résultats dans le cas $p < n$

Dans notre cadre (section 2.1.1) pour $p < n$ avec p fixe, les résultats sur l'estimateur Lasso $\hat{\beta}^L(\lambda_n)$ portent sur la consistance en *estimation* et en *sélection*. Dans ce cadre on suppose que

$$\Psi^n \to \Psi, \quad \text{quand} \quad n \to \infty,$$

où Ψ une matrice définie positive.

Pour le cas trivial où $\lambda_n \overset{n\to\infty}{\to} \infty$, l'estimateur Lasso $\hat{\beta}^L(\lambda_n)$ converge vers zéro dans $\mathbb{R}^p$. L'estimateur $\hat{\beta}^L(\lambda_n)$ n'est pas consistant.

Pour $\lambda_n \overset{n\to\infty}{\to} \lambda_0 \geq 0$, où λ_0 est une constante finie, Knight et Fu [90] montrent que l'estimateur Lasso $\hat{\beta}^L(\lambda_n)$ converge vers l'unique minimum de la fonction

$$\beta \in \mathbb{R}^p \mapsto (\beta - \beta^*)' \Psi (\beta - \beta^*) + \lambda_0 \|\beta\|_1. \tag{2.30}$$

L'estimateur Lasso est consistant en *estimation* si et seulement si $\lambda_0 = 0$. En revanche, la consistance en *sélection* est toujours assurée si $\lambda_0 = 0$ et dépend du minimiseur global de (2.30) sinon.

Pour $\lambda_n \overset{n\to\infty}{\to} 0$ et $\sqrt{n}\lambda_n \overset{n\to\infty}{\to} \infty$, Zou [166] montre que l'estimateur Lasso $\hat{\beta}^L$ est tel que $\lambda_n^{-1}(\hat{\beta}^L - \beta^*)$ converge vers l'unique minimum de la fonction

$$u \in \mathbb{R}^p \mapsto u'\Psi u + u'_{\mathcal{A}^*} \operatorname{Sgn}(\beta^*_{\mathcal{A}}) + \|u_{(\mathcal{A}^*)^C}\|_1.$$

L'estimateur Lasso est biaisé et la consistance en *estimation* n'est plus assurée. En contrepartie, la consistance en *signe* de l'estimateur Lasso est établie avec une probabilité qui tend vers 1, si et seulement si l'hypothèse 2.3 (CI) est satisfaite. Ce résultat est également prouvé indépendamment par Yuan et Lin [160], Zhao et Yu [164], Zou [166] et Meinshausen et Bühlmann [112].

Pour $\sqrt{n}\lambda_n = \lambda_0 > 0$, λ_0 une constante finie, $\hat{\beta}^L$ est biaisé et non consistant en *estimation* (Knight et Fu [90]). En effet, $\sqrt{n}(\hat{\beta}^L - \beta^*)$ converge en loi vers une variable aléatoire d'espérance non nulle. Bach [10] démontre que la probabilité d'obtenir le vecteur signe de $\hat{\beta}^L$ égal à celui de β^* tend vers une limite dans $]0, 1[$. Ce résultat est en accord avec celui obtenu par Zou [166]. Notons que pour cette valeur du paramètre λ_0, le niveau de pénalisation est devenu trop faible pour éliminer les variables non pertinentes de l'étude.

Pour $\sqrt{n}\lambda_n \overset{n\to\infty}{\to} 0$, la vitesse optimale de convergence en *estimation* est $\sqrt{n}$ dans notre cadre de travail (Knight et Fu [90]). Plus précisément, $\sqrt{n}(\hat{\beta}^L - \beta^*)$ converge en loi vers $\mathcal{N}(\mathbf{0}_p, \sigma^2\Psi^{-1})$. Remarquons que l'estimateur Lasso converge vers l'estimateur des moindres carrés, le terme de pénalisation n'intervenant plus. Ainsi, toutes les variables sont sélectionnées et l'estimateur Lasso n'est pas consistant en *sélection* de variables.

2.3.3 Résultats dans le cas $p \geq n$

Dans cette section, nous énumérons les résultats principaux pour $p \geq n$ selon l'objectif de l'étude : *prédiction*, *estimation* et *sélection*. Les résultats obtenus sont en général des IS présentées par la Définition 2.2. Notons toutefois que la dimension p est telle que $p = o(e^n)$.

Prédiction.

Pour une valeur du paramètre de régularisation $\lambda_n = A\sigma\sqrt{\frac{\log(p)}{n}}$ et la constante C dépendant seulement de la norme empirique des X_j et du niveau de bruit σ^2, des IS sont obtenues par Bunea, Tsybakov, et Wegkamp [32], Bunea, Tsybakov, et Wegkamp [33] et Bickel, Ritov, et Tsybakov [17] sous l'hypothèse 2.24 (CM) ou encore l'hypothèse 2.4 (VPR). Dans ces

divers travaux, le *design* est considéré déterministe sauf dans le papier de Bunea, Tsybakov, et Wegkamp [32].

Dans la cadre de la régression linéaire où l'indice de sparsité s est tel que $s = o(\log(p)/n)$, Greenshtein et Ritov [71] montrent que l'erreur de prédiction au sens ℓ_2 de l'estimateur Lasso tend vers zéro. Les techniques utilisées sont proches de celles considérées pour l'obtention des IS en *prédiction*.
Pour le modèle de régression linéaire, Zhang et Huang [163] obtiennent des IS sous l'hypothèse 2.5 (DI) avec λ_n de l'ordre de $\sqrt{\frac{\log(p)}{n}}$, C dépendant de σ^2 et de constantes (valeurs propres) intervenant dans la condition (2.28).

Estimation.
La plupart des IS en *estimation* sont obtenues pour les normes ℓ_1 et ℓ_2. Sous l'hypothèse (CM), Bunea [28] et Bunea, Tsybakov, et Wegkamp [33] obtiennent une IS en norme ℓ_1 pour $\lambda_n = A\sigma\sqrt{\frac{\log(p)}{n}}$ et la constante C dépendant seulement de la norme empirique des X_j et du niveau de bruit σ^2.

Sous l'hypothèse (2.4) moins restrictive (VPR), Bickel, Ritov, et Tsybakov [17] obtiennent une IS en norme ℓ_1 pour une constante C inversement proportionnelle à $\kappa_1^2(s)$, défini en (2.27). De même en norme ℓ_q avec $1 < q \leq 2$ et sous des hypothèses légèrement plus restrictives, ils établissent des IS en *estimation*. Les auteurs prouvent qu'avec une probabilité proche de 1,

$$\|\hat{\beta}^L - \beta^*\|_q^q \leq Cs\left(\sqrt{\frac{\log(p)}{n}}\right)^q,$$

où C est, ici, inversement proportionnelle à $\kappa_1^{2/q}(s)$.

Des IS en *estimation* sont obtenues par Meinshausen et Yu [114] en norme ℓ_2 et par Zhang et Huang [163] en norme ℓ_q avec $q \geq 1$, sous l'hypothèse 2.28 (DI).

Sélection.
La sélection de variables est l'objectif qui a suscité le plus d'intérêt ces dernières années. La question peut se formuler de la façon suivante :
Sous quelles conditions le problème de minimisation sous contrainte ℓ_0 et ℓ_1 coincident-ils ?
Les résultats obtenus dans la littérature sont de trois types étroitement liés.

1. Lorsque des IS sont obtenues en norme-sup (cf. Définition 2.2), on dit que l'estimateur est *consistant en norme-sup*.

2. Le deuxième type de résultats visent à obtenir une borne à échantillon fini de la forme

$$\mathbb{P}(\hat{\mathcal{A}}_L \neq \mathcal{A}^*) \leq \eta_n, \tag{2.31}$$

où η_n est une suite qui tend vers zéro avec n tendant vers l'infini. Si une telle borne est obtenue, on dit que l'estimateur est *consistant en sélection*.

3. On dit qu'un estimateur vérifie la propriété de consistance en *signe* si l'inégalité (2.31) est satisfaite pour $\mathrm{Sgn}(\hat{\beta}^L) \neq \mathrm{Sgn}(\beta^*)$. L'intérêt d'une telle propriété est qu'elle apporte une interprétation claire de la pertinence de chaque variable, mais également du signe de la corrélation entre chaque variable et la réponse Y.

Pour la grande dimension $p \geq n$, les premiers résultats en sélection de variables sont ceux de Donoho [52], Donoho et Tanner [55] et Tropp [137]. Dans le cadre de la régression linéaire *sans bruit*, les auteurs fournissent une condition nécessaire et suffisante pour que les problèmes de minimisation (2.32) et (2.33) soient équivalents

$$\|\beta\|_1 \quad \text{sc.} \quad Y = X\beta, \tag{2.32}$$

$$\|\beta\|_0 \quad \text{sc.} \quad Y = X\beta. \tag{2.33}$$

Dans leur cadre (*sans bruit*), ce résultat permet de reconstituer parfaitement l'ensemble de sparsité $\mathcal{A}^*$. En présence de bruit, une hypothèse supplémentaire est nécessaire pour obtenir des résultats en *sélection*. Cette hypothèse (Hypothèse 2.6) impose à la plus petite coordonnée non nulle du vecteur β^* d'être supérieure à un certain seuil. Soit $\beta_{\min}$ cette quantité, définie telle que

$$\beta_{\min} = \min\{|\beta_j^*| : j \in \mathcal{A}^*\}. \tag{2.34}$$

Hypothèse 2.6. ***Hypothèse (A).***
Il existe une suite de nombres positifs υ_n qui tend vers zéro quand n tend vers l'infini, telle que $\forall n \in \mathbb{N}^*$, $\beta_{\min}$ que vérifie

$$\beta_{\min} \geq \upsilon_n.$$

En présence de bruit, un compromis doit être réalisé entre l'hypothèse (A) et l'hypothèse sur la matrice de Gram Ψ^n considérée (CM, CI, VRP ou DI). Une hypothèse restrictive sur la matrice Ψ^n rend l'estimation du support de β^* aisée et une hypothèse forte sur $\beta_{\min}$ n'est plus nécessaire. L'avantage, d'une hypothèse (A) faible ; i.e. la restriction sur $\beta_{\min}$ est petite, est qu'un signal de faible intensité est plus facilement détectable.

Les premiers résultats en *sélection* de variables, dans le cadre de la grande dimension sont ceux de Meinshausen et Bühlmann [112], Zhao et Yu [164], où la dimension p est supposée

polynomiale en n ($p = \mathcal{O}(n^\gamma)$, $\gamma > 0$) et $\beta_{\min}$ de l'ordre de $n^{-\delta/2}$, $0 < \delta < 1$. Meinshausen et Bühlmann [112] démontrent la consistance en *sélection* de variables dans le modèle graphique gaussien, sous l'hypothèse 2.3 (CI). Les auteurs supposent le vrai paramètre β^* unique. Pour le modèle de régression linéaire, Zhao et Yu [164] démontrent l'*existence* d'une solution vérifiant la consistance en *signe*.

En exploitant une borne supérieure sur l'erreur d'estimation en norme ℓ_2 et sous l'hypothèse 2.28 (DI), Meinshausen et Yu [114], Zhang et Huang [163] établissent la consistance en *norme-sup* de $\hat{\beta}^L$. Ils en déduisent la consistance en *signe* d'une version de l'estimateur Lasso, obtenue par une règle de seuillage dur et notée $\hat{\beta}^{L-seuil} \in \mathbb{R}^p$:

$$\hat{\beta}_j^{L-seuil} = \hat{\beta}_j^L \mathbb{I}(|\hat{\beta}_j^L| \geq C\upsilon_n), \tag{2.35}$$

où υ_n est la vitesse (ordre de grandeur de la borne) intervenant dans l'hypothèse (A). La vitesse υ_n doit être du même ordre que celle obtenue dans l'IS en *norme-sup*. Dans les travaux de Meinshausen et Yu [114], Zhang et Huang [163], cette vitesse vaut respectivement

$$\upsilon_n = \mathcal{O}\left(\sqrt{s\,\tfrac{\log(p)}{n}}\right) \text{ et } \upsilon_n = \mathcal{O}\left(\left(s\,\tfrac{\log(p)}{n}\right)^{1/4}\right).$$

Comparé à l'ordre de grandeur attendu pour l'IS en *norme-sup* (Définition 2.2), cette méthode est, en un sens sous-optimale.

Une autre approche pour établir la consistance en *signe* et qui évite l'apparition de l'indice de sparsité s dans la borne l'IS en *norme-sup*, est de travailler directement en norme ℓ_∞. Sous l'hypothèse (2.23) (CM) et avec cette approche, Lounici [101] obtient la consistance en *norme-sup* de l'estimateur Lasso avec $\upsilon_n = \sqrt{\frac{\log(p)}{n}}$. Il en déduit la consistance en *signe* de $\hat{\beta}^{L-seuil}$ défini en (2.35). Sous ces conditions, le modèle est identifiable ; i.e., le vecteur β^* est unique.

Avant lui et sous l'hypothèse 2.24 (CM) moins restrictive que l'hypothèse (CM), Bunea [28] démontre la consistance en *sélection* de $\hat{\beta}^L$. Elle obtient une vitesse υ_n dans la borne en norme ℓ_∞, de l'ordre de $\sqrt{\frac{\log(np)}{n}}$.

Sous l'hypothèse 2.24 (CM), Wainwright [152] montre qu'*il existe* une solution Lasso qui vérifie la consistance en *signe* dès lors que $\sqrt{\frac{\log(p)}{n}} = o(\upsilon_n)$ et qu'il existe un lien entre n, la taille de l'échantillon et (p, s) le nombre de variables et l'indice de sparsité respectivement.

2.3.4 Limites et critiques de l'estimateur Lasso

En reprenant les résultats énoncés précédemment, nous exposons ici brièvement les inconvénients de l'estimateur Lasso (2.15) d'un point de vue théorique.

– *Technique.* L'ensemble des résultats relatifs à l'estimateur Lasso, que ce soit en termes d'IS ou de consistance en sélection de variables, font intervenir une hypothèse sur la matrice de Gram Ψ^n, qui impose de faibles corrélations entre les variables. Celle-ci limite de fait le champ d'application de l'estimateur Lasso.
– *Modélisation.* Dans de nombreuses applications, une information a priori sur les variables est à la disposition du statisticien, qu'il est intéressant d'exploiter. Or l'estimateur Lasso ne permet pas de prendre en compte cette information a priori, comme par exemple les corrélations entre les variables.
– *Cadre supervisé.* L'estimateur Lasso est introduit pour répondre à des problèmes statistiques dans le cadre supervisé ; il est en outre construit sur l'échantillon étiqueté (X, Y). Cette méthode n'est en revanche pas adaptée à d'autres types d'échantillonnage, comme dans le cadre semi-supervisé ou le cadre transductif selon Vapnik [145]. Nous verrons plus loin que d'autres méthodes peuvent, dans le cadre semi-supervisé, intégrer la connaissance des (X, Y) et celle des x_{n+i}, $i \geq 1$.

2.4 Extensions du Lasso

Les limites théoriques (cf. section 2.3.4) et pratiques (cf. remarque 2.1 et annexe B) de l'estimateur Lasso ont motivé l'étude d'extensions et de généralisations de la méthode. Dans cette section, nous présentons une sélection de méthodes developpées dans la littérature.

2.4.1 Variantes du Lasso : Dantzig Selector et pertes robustes

Nous présentons ici, le *Dantzig Selector* et d'autres méthodes combinant la pénalité ℓ_1 à des pertes robustes.

Dantzig Selector :
Introduit par Candès et Tao [35], le Dantzig selector est défini par :

$$\hat{\beta}^D(\lambda) \in \begin{cases} \operatorname{argmin}_{\beta \in \mathbb{R}^p} \|\beta\|_1 \\ \\ \text{s.c. } \left\|\Xi^{-1} X' (Y - X\beta)\right\|_\infty \leq \lambda, \end{cases}$$

où $\lambda_n \geq 0$ est le paramètre de régularisation et Ξ est la matrice diagonale de taille p, dont le j-ième coefficient diagonal vaut $\|X_j\|_n = \sqrt{\frac{1}{n}\sum_{i=1}^n x_{i,j}^2}$. La similitude du Dantzig Selector

avec la forme duale du Lasso (2.16), permet de les classer dans une même famille d'estimateurs définis par projection de $\mathbf{0_p}$ sur la région de confiance de niveau $1-\eta$, avec $\eta = \eta(\lambda)$, définie par la contrainte $\left\|\Xi^{-1}X'(Y - X\beta)\right\|_\infty \leq \lambda$ (Alquier et Hebiri [7]).

James, Radchenko, et Lv [84] considèrent des cas pour lesquels le Lasso et le Dantzig Selector sont équivalents d'un point de vue pratique. Ils ont également prouvé cette similitude dans le cas de corrélations particulières entre les variables. Cette étude est par ailleurs appuyée par Bickel, Ritov, et Tsybakov [17], qui ont prouvé sous l'hypothèse 2.4 (VPR), que ces deux estimateurs ont des erreurs de *prédiction* en norme ℓ_2 équivalentes, à un terme résiduel près. Notons également que plusieurs IS ont été établies pour le Dantzig Selector entre autres par Bickel, Ritov, et Tsybakov [17], Candès et Tao [35] et Koltchinskii [91] qui témoignent de ses bonnes propriétés. La consistance en *signe* de l'estimateur $\hat{\beta}^D$ sous l'hypothèse (2.23) (CM) a éte prouvé par Lounici [101].

LAD-Lasso et pertes Lipschitz :

Le Lasso (2.15) est défini avec une perte ℓ_2, à savoir $\|Y - X\beta\|_n^2$. Une amélioration peut lui être apportée dans le cas d'erreurs ε_i, $i = 1, \ldots, n$ à queues lourdes, ou en cas de présence de points aberrants, grâce à l'introduction de pertes plus robustes associées à la pénalité ℓ_1. Ainsi, Wang, Li, et Jiang [153] ont introduit le *LAD-Lasso (Least Absolute Deviation-Lasso)* défini sur la perte $n^{-1}\|Y - X\beta\|_1$, qui combine sélection et robustesse. Les auteurs ont étudié les performances théoriques et pratiques de cet estimateur dans le cas $p \leq n$.

Nous pouvons également citer le travail de Rosset et Zhu [126] qui combine la pénalité Lasso à la perte Huber : $n^{-1}\sum_{i=1}^n \ell(y - x_i\beta)$, où pour un noeud fixé t, la fonction ℓ est définie par :

$$\ell(u) = \begin{cases} u^2, & \text{si } |u| \leq t, \\ 2t|u| - t^2, & \text{sinon.} \end{cases}$$

En grande dimension, van de Geer [143, 142] déterminent des IS pour des pertes de type Lipschitz qui peuvent être utilisées pour produire des estimateurs consistants en *sélection*. Koltchinskii [93] associe la pénalité ℓ_1 à des pertes Lipschitz, continues et deux fois différentiables. L'auteur établit plusieurs IS proches de celles considérées pour le Lasso. Il fournit également des résultats sur l'erreur de prédiction au sens ℓ_2, sans hypothèse sur la matrice de Gram Ψ^n.

Dans un travail plus récent, Lounici, Pontil, Tsybakov, et van de Geer [102] ont étudié le

problème de régression linéaire à réponses multiples en considérant la perte ℓ_2. Les auteurs démontrent des IS, ainsi que la consistance en *sélection* de l'estimateur Lasso. Notons que les vitesses de convergence obtenues sont alors meilleures que celles obtenues en appliquant la procédure Lasso séparément sur chaque dimension de la réponse.
En outre, la pénalité ℓ_1 a également fait ses preuves dans le cadre de la classification, pour lequel nous pouvons citer Tarigan et van de Geer [134], van de Geer [143, 142].

2.4.2 Versions adaptatives du Lasso

De nombreux auteurs se sont inspirés des trauvaux autour de l'estimateur Lasso pour apporter des améliorations sur les aspects théoriques et pratiques. Les méthodes présentées ici visent à améliorer l'estimateur Lasso en vue de la consistance en *sélection*.

- D'un point de vue théorique, il s'agit de relâcher les hypothèses originelles du Lasso, tout en satisfaisant à la consistance en *sélection* et aux Inégalités de Sparsité.
- D'un point de vue pratique, il s'agit d'améliorer les performances sélectives et prédictives du Lasso, en respectant un temps de calcul raisonnable.

Dans cette section nous présentons quelques une de ces méthodes.

Adaptive Lasso :

Cette méthode, introduite par Zou [166], est définie en deux étapes. Dans un premier temps, le statisticien calcule un estimateur préliminaire $\tilde{\beta} \in \mathbb{R}^p$, pouvant être l'EMC, ou l'estimateur ridge ou tout autre estimateur. Dans un second temps, cet estimateur préliminaire $\tilde{\beta}$ est utilisé pour ajuster la pénalité imposée sur chacun des coefficients du paramètre de régression de l'Adaptive Lasso de la façon suivante :

$$\hat{\beta}^{AdapL}(\lambda, \tilde{\beta}) \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|Y - X\beta\|_n^2 + \lambda \sum_{j=1,\dots,p} \frac{1}{|\tilde{\beta}_j|} |\beta_j|. \tag{2.36}$$

Les pondérations $\frac{1}{|\tilde{\beta}_j|}$ permettent ici de réduire la pénalisation des coefficients β_j lorsque la valeur de $|\tilde{\beta}_j|$ est grande et de la renforcer dans le cas contraire.

Dans le cas où l'estimateur préliminaire $\tilde{\beta}$ est asymptotiquement consistant pour l'erreur d'estimation, l'Adaptive Lasso est consistant en *sélection* de variables et en *estimation*, avec p fixé. Ces propriétés restent vérifiées même dans des situations où le Lasso échoue.

Ces résultats sont affinés, et parfois controversés par Pötscher et Schneider [124] qui fournissent une analyse plus précise de la consistance de l'estimateur Adaptive Lasso. Ils prouvent

entre autres que la vitesse de convergence de cet estimateur est plus faible que $1/\sqrt{n}$ et déterminent sa distribution limite (voir également Pötscher et Leeb [123]).

Dans le cas $p \geq n$, sous l'hypothèse 2.4 (VPR) et l'hypothèse 2.6 (A) avec $\upsilon_n = \sqrt{s\log(p)/n}$, Zhou, van de Geer, et Bühlmann [165] montrent la consistance en *sélection* de l'Adaptive Lasso construit avec de l'estimateur Lasso comme estimateur préliminaire. Les auteurs généralisent également l'étude au modèle graphique gaussien.
D'autres adaptations du Lasso à la sélection de variables ont également été considérées par Meinshausen [111], sous le nom de *Relaxed Lasso*.

Bolasso :
Bach [10, 11] a introduit un estimateur consistant en *sélection* : l'estimateur Bolasso (*Bootstrapped Lasso*). Pour une dimension raisonnable $p \leq n$ et le paramètre de régularisation λ_n tel que $\sqrt{n}\lambda_n = \lambda_0$, avec λ_0 une constante strictement positive (section 2.3.2), l'auteur est parti du constat suivant : le Lasso sélectionne l'ensemble des variables pertinentes avec une probabilité qui tend vers 1 à une vitesse exponentielle et chacune des autres variables non pertinentes avec une probabilité comprise entre 0 et 1 strictement. Comme nous l'avons noté dans la section 2.3.2, le régime en question est tel que $\sqrt{n}\lambda_n = \lambda_0$, avec λ_0 une constante strictement positive.

L'estimateur Bolasso est construit à partir de M échantillons bootstraps indépendants (conditionnellement à $\{(x_1, y_1), \ldots, (x_n, y_n)\}$) ; i.e., M réplications bootstraps de l'échantillon de départ $\{(x_1, y_1), \ldots, (x_n, y_n)\}$. Pour une valeur $\lambda_n = \lambda_0/\sqrt{n}$ du paramètre de régularisation, M estimateurs $\left\{\hat{\mathcal{A}}^m\right\}_{m=1,\ldots,M}$ de l'ensemble de sparsité $\mathcal{A}^*$ sont construit à partir des M échantillons bootstraps. L'estimateur Bolasso est défini comme l'intersection de ces M ensembles

$$\hat{\mathcal{A}}_{Bolasso} = \bigcap_{m\in\{1,\ldots,M\}} \hat{\mathcal{A}}^m.$$

Avec une grande probabilité, l'auteur prouve que $\hat{\mathcal{A}}_{Bolasso} \supseteq \mathcal{A}^*$. Pour une dimension p telle que $p^6 \leq Cn$, où C est une constante positive et sous des conditions sur le nombre M d'échantillons bootstraps il établit la consistance en *sélection* de l'estimateur Bolasso.

En pratique, cet estimateur témoigne d'une bonne capacité à sélectionner le bon ensemble de sparsité même pour de plus grandes dimensions. Notons que cette méthode est peu sensible à des variations du paramètre de régularisation. Il suffit de prendre un λ_n relativement faible, ce qui permet de sélectionner un nombre élevé de variables par itération.

Sélection par Stabilisation :

Dans la méthode Bolasso, l'estimateur est construit par randomisation sur les observations (lignes de X). Meinshausen et Bühlmann [113] proposent de définir un nouvel estimateur par randomisation sur les variables (colonnes de X). Les auteurs considèrent des sous-échantillons de taille $\lfloor n/2 \rfloor$ pour chaque réplication Bootstrap, où $\lfloor \cdot \rfloor$ désigne la partie entière. Ils introduisent des poids aléatoires et indépendants $W_j \in]0,1]$ associés aux coefficients du vecteur β dans la définition de la pénalité Lasso. La pénalité s'écrit alors :

$$\text{pen}(\beta) = \lambda \sum_{j=1}^{n} \frac{|\beta_j|}{W_j}.$$

Cette méthode, appellée *Lasso randomisé avec stabilisation*, permet de sélectionner les variables pertinentes pour un nombre raisonnable de randomisations. Sous l'hypothèse 2.6 (A), avec υ_n de l'ordre de $s^{5/2}\sqrt{\log(p)/n}$, où s est l'indice de sparsité, les auteurs établissent des résultats de consistance en *sélection*. Notons que cette méthode est peu sensible au choix de λ. Des simulations montrent que, lorsque le paramètre de régularisation λ est déterminé par validation croisée, l'estimateur par stabilisation sélectionne beaucoup moins de variables non pertinentes que le vrai Lasso.

2.4.3 Méthodes de type Lasso

Suite aux difficultés pratiques rencontrées par l'estimateur Lasso dans divers problèmes présentant de fortes corrélations entre les variables (cf. remarque 2.1 et l'annexe B), de nombreux travaux réajustant la forme de la pénalité ont vu le jour.

Elastic Net :

Une méthode phare est l'*Elastic Net* de Zou et Hastie [167], défini par :

$$\hat{\beta}^{EN}(\lambda,\mu) \in \underset{\beta\in\mathbb{R}^p}{\text{argmin}} \, \|Y - X\beta\|_n^2 + \lambda\|\beta\|_1 + \mu\|\beta\|_2, \tag{2.37}$$

où λ et μ sont deux paramètres de régularisation. La pénalité est définie ici comme la somme d'une pénalité de type Lasso (norme ℓ_1) et d'un deuxième terme de type ridge (norme ℓ_2). Chaque terme de cette pénalité joue un rôle spécifique : la pénalité Lasso assure la sparsité, tandis que la pénalité ridge permet de capturer les corrélations entre les variables. Un algorithme de type LARS (algorithme itératif détaillé dans l'annexe B) est utilisé pour construire le chemin de régularisation associé à cette méthode. Cette version modifiée de l'algorithme LARS, permet de prendre en compte les corrélations entres les variables lors de la séléction.

Sur le plan théorique, Zou et Hastie [167] donnent une borne sur la distance $|\hat{\beta}_j^{EN} - \hat{\beta}_k^{EN}|$ entre deux coefficients du vecteur $\hat{\beta}^{EN}$. Cette distance dépend de la corrélation entre les variables X_j et X_k correspondantes. La consistance en *sélection* de variables de l'estimateur Elastic Net a également été étudiée par De Mol, De Vito, et Rosasco [50] lorsque $p \leq n$.
En grande dimension ($p \geq n$), Bunea [26] a étudié les performances de $\hat{\beta}^{EN}$ sous l'hypothèse 2.4 (VPR), où les sommes sont considérées uniquement sur l'ensemble $\mathcal{A}^*$. Elle fournit une IS en *estimation* avec grande probabilité, de la forme :

$$\|\hat{\beta} - \beta^*\|_1 \leq \frac{C}{\kappa_1(|\mathcal{A}^*|) + \mu} s\sqrt{\frac{\log(p)}{n}}, \tag{2.38}$$

où $\kappa_1(|\mathcal{A}^*|)$ est définie dans l'hypothèse (VRP). Remarquons, que pour $\mu = 0$ dans l'expression (2.38), l'estimateur $\hat{\beta}$ n'est autre que l'estimateur Lasso. Pour l'Elastic Net, la quantité μ est définie par $\mu = \frac{\lambda}{2\beta_{\max}}$, où $\beta_{\max} \geq \max_j |\beta_j^*|$.

Dans l'IS (2.38), il ressort que plus μ est grand, meilleure est la borne. Toutefois, ce gain reste relativement faible puisque μ est proportionnel à λ qui est petit et inversement proportionnel à $\beta_{\max}$ qui peut être grand selon l'intensité du signal. D'autre part, les hypothèses nécessaires pour établir cette IS sont, a peu de choses près, les mêmes pour l'estimateur Lasso et l'Elastic Net. L'Elastic Net n'apporte pas d'améliorations significatives au Lasso.

Sous une version légèrement plus restreinte de l'hypothèse 2.4 (VRP) et sous l'hypothèse 2.6 (A), Bunea [26] déduit de l'IS (2.38), des résultats de consistance en *sélection* de variables. Dans l'hypothèse (A), elle considère $\upsilon_n = \frac{|\mathcal{A}^*|}{\kappa_1(\mathcal{A}^*)}\sqrt{\log(p)/n}$. Sous l'hypothèse (2.24) (CM) et pour une hypothèse (A) allégée avec $\upsilon_n = \sqrt{\log(p)/n}$, elle obtient des résultats similaires.

Par ailleurs, Jia et Yu [85] démontrent l'existence d'une solution $\hat{\beta}^{EN}$ au problème de minimisation (2.37) consistante *en signe*. Ce résultat est établi sous une version de l'hypothèse 2.3 (CI) adaptée à l'estimateur Elastic Net.

Enfin, nous pouvons citer des extensions de la méthode Elastic Net à la régression logistique par Bunea [26] et dans le cadre de la classification par Bühlmann et Hothorn [24].

Fused Lasso :

Par une approche analogue à celle considérée par Zou et Hastie [167] qui ont introduit l'Elastic Net, Tibshirani, Saunders, Rosset, Zhu, et Knight [136] ont construit l'estimateur

Fused Lasso comme solution au problème suivant :

$$\hat{\beta}^{FL}(\lambda,\mu) \in \underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}} \|Y - X\beta\|_n^2 + \lambda\|\beta\|_1 + \mu\sum_{j=2}^{p}|\beta_j - \beta_{j-1}|, \tag{2.39}$$

où λ et μ sont deux paramètres de régularisation. Les auteurs ont étudié la distribution asymptotique de cet estimateur dans le cas $p \leq n$.

Cet estimateur intègre une pénalité sur la distance ℓ_1 entre les coefficients successifs du vecteur β, qui permet à l'estimateur Fused Lasso de répondre à des problèmes où les variables sont ordonnées. Le terme $\mu\sum_{j=2}^{p}|\beta_j - \beta_{j-1}|$ est aussi appelé pénalité *fusion* par Land et Friedman [94].
Alors que la sparsité est toujours assurée par la pénalité Lasso, la pénalité ℓ_1 sur la différence entre les coefficients successifs impose que ces derniers soient égaux (on parle de *sparsité entre coefficients successifs*). Ainsi, lorsque l'objectif est l'estimation du vecteur β^*, la sparsité entre coefficients successifs permet d'interpréter plus facilement les résultats.

Dans le cadre de l'estimation non-paramétrique, la consistance en *sélection* du Fused Lasso et de l'Adaptive Fused Lasso (une version adaptative de celui-ci) a été considérée par Rinaldo [125]. Pour établir cette consistance, l'auteur a introduit une hypothèse restrictive, qui impose aux coefficients de régression non nuls du vecteur β^* d'être regroupés par blocs, et aux coefficients d'un même bloc d'être égaux.

Méthode groupée :
Une méthode prometteuse pour répondre à des problèmes où les variables sont corrélées, et dans le meilleur des cas, ordonnées, est connu sous le nom de *Group Lasso*. Introduite par Yuan et Lin [159], cette méthode considère, en lieu et place des variables individuelles, des groupes de variables. Soit L un entier dans $\{1,\ldots,p\}$ désignant le nombre de groupes considérés. Décrivons alors une partition de $\{1,\ldots,p\}$ par les ensembles $(G_l)_{l\in\{1,\ldots,L\}}$, où G_l est l'ensemble des indices des variables contenues dans le groupe G_l avec $l = 1\ldots,L$. L'estimateur Group Lasso est alors défini par :

$$\hat{\beta}^{G}(\lambda) \in \underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}} \|Y - X\beta\|_n^2 + \lambda\sum_{l=1}^{L}\sqrt{\sum_{j\in G_l}\beta_j^2}. \tag{2.40}$$

Parmi les premiers résultats théoriques, Bach [12] étend la Condition d'Irreprésentabilité (2.25) aux cas des variables groupées lorsque p est supposé fixe, avec $p \leq n$. L'auteur donne dans ce cas la consistance en *sélection de groupes* de la procédure Group Lasso ; i.e.

$$\mathbb{P}(\hat{\Theta}_G = \Theta_G^*) \to 1,$$

où $\Theta_G^* = \{l \in \{1, \ldots, L\} : \exists j \in G_l \text{ avec } \beta_j^* \neq 0\}$ et $\hat{\Theta}_G$ est son équivalent, en considérant $\hat{\beta}_j^G$ au lieu de β_j^*. On remarque que la consistance en *sélection de groupes* établie par Bach [12] n'implique la consistance en *sélection de variables* que lorsque les groupes qui contiennent des variables pertinentes ne sont pas entachées par des variables non pertinentes. Sous des hypothèses similaires, Nardi et Rinaldo [115] considèrent les consistances en *sélection de groupes* et en *estimation* du Group Lasso.

En grande dimension et sous une hypothèse sur la matrice de Gram Ψ^n, Huang et Zhang [81] fournissent des IS en *estimation* pour la norme ℓ_2. Une version synthétique leurs résultats est :

$$\|\hat{\beta}^G - \beta^*\|_2^2 \leq C \left(\frac{s_G + |\Theta_G^*| \log(L)}{n} \right),$$

où $s_G = |\bigcup_{l \in \Theta_G^*} G_l|$, est le nombre total de variables incluses dans les groupes pertinents. Par analogie aux variables pertinentes, on appelle *groupe pertinent*, un groupe contenant au moins une variable pertinente, i.e. un groupe G_l tel que l appartient à Θ_G^*. La borne ci-dessus est obtenue avec grande probabilité. Les auteurs montrent que le Group Lasso améliore l'estimateur Lasso lorsque les groupes de variables sont bien déterminés et que le nombre de variables par groupe est élevé. Toutefois, un compromis doit être réalisé entre s_G et $|\Theta_G^*| \log(L)$. Si les variables pertinentes sont dans un même groupe, alors s_G est égal ou légèrement plus grand que $|\mathcal{A}^*|$, le vrai indice de sparsité. En effet, augmenter la taille des groupes conduit le plus souvent à sélectionner des variables non pertinentes, ce qui augmente s_G. Un nombre élevé de variables par groupe implique que le nombre de groupes L et le nombre de groupes pertinents $|\Theta_G^*|$ sont petits.

Ces résultats complétent les résultats obtenus par Chesneau et Hebiri [44] sur le Group Lasso présentés au Chapitre 4. Des IS pour l'erreur en *prédiction* sont obtenues, et montrent la supériorité dans certains cas du Group Lasso sur le Lasso.

Les bonnes propriétés de l'estimateur Group Lasso incitent à l'élargissement de son champ d'application. L'estimateur Group Lasso a été considéré par McKay Curtis et Ghosal [107] du point de vue de l'inférence bayésienne et par Obozinski, Wainwright, et Jordan [117] dans le cadre de la régression linéaire à réponse multiples. Dans le cadre de la classification, Meier, van de Geer, et Bühlmann [109] exploitent la pénalité groupée combinée avec la perte logistique.

Nonnegative garrote :

Une méthode antérieure au Lasso et très similaire à l'Adaptive Lasso est le *Nonnegative garrote* de Breiman [23, 22]. Définissons la matrice $Z = (Z_1, \ldots, Z_p)$ de taille $n \times p$ telle que $Z_j = X_j \hat{\beta}_j^{init}$, où $\hat{\beta}^{init}$ est un estimateur initial de β^*. Définissons ensuite le vecteur $\hat{d}(\lambda) \in \mathbb{R}_+^p$, solution du problème suivant :

$$\hat{d}(\lambda) \in \begin{cases} \text{argmin}_{d \in \mathbb{R}^p} \|Y - Zd\|_n^2 + \lambda \sum_{j=1}^p d_j \\ \\ \text{s.c. } d_j \geq 0, \ \forall j \in \{1, \ldots, p\}, \end{cases} \tag{2.41}$$

où λ est un paramètre de régularisation positif. Le vecteur $\hat{d}(\lambda) = \hat{d}$ joue le rôle de paramètre d'échelle, de sorte que le Nonnegative garrote est défini par la relation : $\hat{\beta}_j^{NG} = \hat{d}_j \hat{\beta}_j^{init}$. Notons que dans ses travaux, Breiman [23, 22] considère comme estimateur initial, l'EMC $\hat{\beta}^{init} = \hat{\beta}^{MC}$. Dans le cas particulier où le design X est orthogonal ; i.e. $\Psi^n = \frac{X'X}{n} = \mathbf{I}_p$, où $\mathbf{I}_p$ est la matrice identité de taille p, et pour $\hat{\beta}^{init} = \hat{\beta}^{MC}$, il est interéssant de souligner que la solution du problème (2.41) est explicite :

$$\hat{d}_j = \left(1 - \frac{\lambda}{2(\hat{\beta}_j^{init})^2}\right)_+, \quad j = 1, \ldots, p.$$

Cette forme met en avant l'influence de l'estimateur initial sur la solution optimale et plus particulièrement sur sa sparsité.

Comme pour l'Adaptive Lasso, d'autres choix de l'estimateur initial sont possibles. Yuan et Lin [160] étudient les performances du Nonnegative garrote lorsque l'estimateur initial est l'estimateur ridge $\hat{\beta}^R = \hat{\beta}^R(\mu) = (X'X + \mu \mathbf{I}_p)^{-1} X'Y$, où μ est le paramètre de régularisation. Ce choix permet de traiter des problèmes où la matrice Ψ^n est mal déterminée. Yuan et Lin [160] proposent un algorithme de type LARS pour approcher l'estimateur Nonnegative garrote. Ils prouvent également la consistance de cette méthode en *estimation* et en *sélection* quand $p \leq n$ dans des situations où l'estimateur Lasso ne l'est pas. En particulier, aucune condition du type de l'hypothèse 2.3 (CI) n'est requise pour obtenir cette consistance. Seule une borne sur la plus petite valeur propre de Ψ^n est nécessaire, ce qui est une condition non restrictive lorsque $p \leq n$.

2.5 Motivations et plan de la thèse

Les travaux réalisés dans cette thèse visent à apporter des améliorations à l'estimateur Lasso dans trois directions.

1. *Structuration de l'espace des variables.* Nous proposons des méthodes de sélection de variables capables d'incorporer une information a priori sur les variables explicatives. Nous nous intéressons en particulier à deux approches : la première vise à améliorer les performances de l'estimateur Lasso lorsqu'une relation d'ordre est connue entre les variables ; la deuxième consiste à exploiter l'existence d'une structure de groupes entre les variables.

2. *Sparsité générale.* Nous proposons d'étendre la notion de sparsité et d'étudier des problèmes où la sparsité du modèle n'intervient pas de manière habituelle ; i.e le vecteur β^* ne contient pas (ou pas beaucoup) de composantes égales à zéro. La sparsité est observée après une transformation du vecteur β^*. L'intérêt de cette étude est de considérer de nouveaux objectifs exploitant cette sparsité non habituelle et d'observer la modification de l'hypothèse sur la matrice de Gram, dans cette nouvelle configuration.

3. *Objectif de transduction.* Il est intéressant de considérer un objectif différent des ceux présentés dans la Section 2.2.2. Dans le cadre du modèle de régression (2.1), lorsque, en plus des observations $(x_i, y_i)_{i=1\dots,n}$, nous disposons d'observations supplémentaires $\{x_{n+i}\}_{1\leq i\leq m}$ avec $m \geq 1$, il est possible de construire un estimateur capable de fournir une bonne prévision des réponses associées à ces nouvelles observations, en incorporant celles-ci dans sa construction. Ce mode d'échantillonnage renvoie à l'approche transductive décrite par Vapnik [145] dans laquelle la méthode d'estimation fournit une prévision locale et donc plus précise que si nous avions défini une règle globale.

Le reste de manuscrit est organisé comme suit. Dans le chapitre 3, présente l'étude d'un nouvel estimateur : le S-Lasso, qui répond à des problèmes d'estimation lorsque les variables successives sont corrélées. Nous proposons ensuite une extension de ce travail à une large famille d'estimateurs, incluant notamment l'Elastic Net. Nous étudions en particulier un algorithme de type LARS, calculant les solutions de l'estimateur S-Lasso. Enfin, nous comparons l'estimateur S-Lasso aux estimateurs Lasso et de l'Elastic Net.

Le chapitre 4 présente un travail réalisé en collaboration avec Christophe Chesneau. Il est dédié à l'étude théorique d'un estimateur dont les variables sont structurées en groupes. Nous établissons la première IS pour une méthode groupée ainsi qu'une comparaison des performances théoriques avec l'estimateur Lasso.

Le chapitre 5 est écrit en collaboration avec Pierre Alquier. Dans le cadre de problèmes d'inférence lorsqu'une transformation de β^* est sparse, nous présentons une méthode d'es-

timation capable d'exploiter cette sparsité particulière. La méthode proposée permet alors de considérer des objectifs différents de ceux que nous avons considérés jusqu'alors.

En en collaboration avec Pierre Alquier, le chapitre 6 est une étude de prévision dans le cadre transductif, lorsque l'on dispose des observations habituelles $(x_i, y_i)_{i=1\ldots,n}$, et également de nouvelles observations non étiquetées $(x_{n+1} \ldots, x_{n+m})$. Le but est alors de fournir une bonne approximation des réponses associées aux nouveaux individus en exploitant directement les nouvelles observations dans la construction de l'estimateur. Notre approche consiste à étudier une version transductive des estimateurs Lasso et Dantzig Selector. Notre procédure montre de bonnes performances locales autour des observations $x_{n+1} \ldots, x_{n+m}$. De plus, une étude expérimentale est menée, comparant l'estimateur Lasso Transductif à l'estimateur Lasso originel.

Le chapitre 7 se place dans un cadre transductif particulier. Nous considérons une approche séquentielle dans laquelle nous ne disposons que d'une seule nouvelle observation x_{new} par unité de temps. L'objectif est de prédire la réponse de x_{new} par un intervalle de confiance construit directement à partir de cette nouvelle observation. Selon la terminologie de Vovk, Gammerman, et Shafer [150], ces intervalles sont appelés *prédicteurs conformes*. La construction de ces prédicteurs s'appuie sur des estimateurs sparses, comme le Lasso. Dans la méthode proposée, le paramètre de régularisation λ de l'estimateur est déterminé de manière adaptative (il ne dépend que des observations).

Chapter 3

Regularization with the Smooth-Lasso procedure

Abstract: We consider the linear regression problem. We propose the S-Lasso procedure to estimate the unknown regression parameters. This estimator enjoys sparsity of the representation while taking into account correlation between successive covariates (or predictors). The study covers the case when $p \gg n$, i.e., the number of covariates is much larger than the number of observations. From a theoretical point of view, for a fixed p, we establish asymptotic normality and consistency in variable selection results for our procedure. When $p \geq n$, we provide variable selection consistency results and show that the S-Lasso achieved a Sparsity Inequality, i.e., a bound in term of the number of non-zero components of the "true" regression vector. It appears that the S-Lasso has nice variable selection properties compared to its challengers. Furthermore, we provide an estimator of the effective degree of freedom of the S-Lasso estimator. The study is also generalized to a wide family of estimators, which includes the Elastic-Net (Zou and Hastie [167]) as a special case. A simulation study shows that the S-Lasso performs better than the Lasso as far as variable selection is concerned especially when high correlations between successive covariates exist. This procedure also appears to be a good challenger to the Elastic-Net.

3.1 Introduction

We focus on the usual linear regression model:

$$y_i = x_i\beta^* + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{3.1}$$

where the design $x_i = (x_{i,1}, \ldots, x_{i,p}) \in \mathbb{R}^p$ is deterministic, $\beta^* = (\beta_1^*, \ldots, \beta_p^*)' \in \mathbb{R}^p$ is the unknown parameter and $\varepsilon_1, \ldots, \varepsilon_n$, are independent identically distributed (i.i.d.) centered Gaussian random variables with known variance σ^2. We wish to estimate β^* in the sparse case, that is when many of its unknown components equal zero. Thus only a subset of the design covariates $(X_j)_j$ is truly of interest where $X_j = (x_{1,j}, \ldots, x_{n,j})'$, $j = 1, \ldots, p$. Moreover the case $p \gg n$ is not excluded so that we can consider p depending on n. In such a framework, two main issues arise: i) the interpretability of the resulting prediction; ii) the control of the variance in the estimation. Regularization is therefore needed. For this purpose we use selection type procedures of the following form:

$$\tilde{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \|Y - X\beta\|_n^2 + \operatorname{pen}(\beta) \right\}, \tag{3.2}$$

where $X = (x_1', \ldots, x_n')'$, $Y = (y_1, \ldots, y_n)'$ and $\operatorname{pen} : \mathbb{R}^p \to \mathbb{R}$ is a positive convex function called the penalty. For any vector $a = (a_1, \ldots, a_n)'$, we have adopted the notation $\|a\|_n^2 = n^{-1} \sum_{i=1}^n |a_i|^2$ (we denote by $< \cdot, \cdot >_n$ the corresponding inner product in $\mathbb{R}^n$). The choice of the penalty appears to be crucial. Although well-suited for variable selection purpose, Concave-type penalties (Fan and Li [63], Tsybakov and van de Geer [140] and Dalalyan and Tsybakov [46]) are often computationally hard to optimize. Lasso-type procedures (modifications of the l_1 penalized least square (Lasso) estimator introduced by Tibshirani [135]) have been extensively studied during the last few years. Between many others, see Bunea [28], Bunea, Tsybakov, and Wegkamp [33], Zhao and Yu [164] and references inside. Such procedures seem to respond to our objective as they perform both regression parameters estimation and variable selection with low computational cost. We will explore this type of procedures in our study.

In this chapter, we propose a novel modification of the Lasso we call the *Smooth-lasso* (*S-lasso*) estimator. It is defined as the solution of the optimization problem (3.2) when the penalty function is a combination of the Lasso penalty (i.e., $\sum_{j=1}^p |\beta_j|$) and the l_2-fusion penalty (i.e., $\sum_{j=2}^p (\beta_j - \beta_{j-1})^2$). The l_2-fusion penalty was first introduced by Land and Friedman [94]. We add it to the Lasso procedure in order to overcome the variable selection problems observed by the Lasso estimator. Indeed the Lasso estimator has good selection properties but fails in some situations. More precisely, in several works (Bunea [28], Lounici [101], Meinshausen and Bühlmann [112], Wainwright [152], Yuan and Lin [160], Zhao and Yu [164] and Zou [166] among others) conditions for the consistency in variable selection

of the Lasso procedure are given. It was shown that the Lasso is not consistent when high correlations exist between the covariates. We give similar consistency conditions for the S-Lasso procedure and show that it is consistent in variable selection in much more situations than the Lasso estimator. From a practical point of view, problems are also encountered when we solve the Lasso criterion with the Lasso modification of the LARS algorithm (Efron, Hastie, Johnstone, and Tibshirani [61]). Indeed this algorithm tends to select only one representing covariates in each group of correlated covariates. We attempt to respond to this problem in the case where the covariates are ranked so that high correlations can exist between successive covariates. We will see through simulations that such situations support the use of the *S-lasso* estimator. This estimator is inspired by the *Fused-Lasso* (Tibshirani, Saunders, Rosset, Zhu, and Knight [136]). Both S-Lasso and Fused-Lasso combine a l_1-penalty with a fusion term (Land and Friedman [94]). The fusion term is suggested to catch correlations between covariates. More relevant covariates can then be selected due to correlations between them. The main difference between the two procedures is that we use the l_2 distance between the successive coefficients (i.e., the l_2-fusion penalty) whereas the Fused-Lasso uses the l_1 distance (i.e., the l_1-fusion penalty: $\sum_{j=2}^{p} |\beta_j - \beta_{j-1}|$). Hence, compared to the Fused-Lasso, we sacrifice sparsity between successive coefficients in the estimation of β^* in favor of an easier optimization due to the strict convexity of the l_2 distance. However, since sparsity is yet ensured by the Lasso penalty. The l_2-fusion penalty helps us to catch correlations between covariates. Consequently, even if there is no perfect match between successive coefficients our result are still interpretable. Moreover, when successive coefficients are significantly different, a perfect match seems to be not really adapted. In the theoretical point of view, The l_2 distance also helps us to provide theoretical properties for the S-Lasso which in some situations appears to outperforms the Lasso and the Elastic-Net (Zou and Hastie [167]), another Lasso-type procedure. Let us mention that variable selection consistency of the Fused-Lasso and the corresponding Fused adaptive Lasso has also been studied by Rinaldo [125] but in a different context from the one in the present chapter. The result obtained by Rinaldo [125] are established not only under the sparsity assumption, but the model is also supposed to be *blocky*, that is the non-zero coefficients are represented in a block fashion with equal values inside each block.

Many techniques have been proposed to solve the weaknesses of the Lasso. The Fused-Lasso procedure is one of them and we give here some of the most popular methods; the Adaptive Lasso was introduced by Zou [166], which is similar to the Lasso but with adaptive weights used to penalize each regression coefficient separately. This procedure reaches 'Oracles Properties' (i.e. consistency in variable selection and asymptotic normality). Another approach is used in the Relaxed Lasso (Meinshausen [111]) and aims to doubly-control

the Lasso estimate: one parameter to control variable selection and the other to control shrinkage of the selected coefficients. To overcome the problem due to the correlation between covariates, group variable selection has been proposed by Yuan and Lin [159] with the Group-Lasso procedure which selects groups of correlated covariates instead of single covariates at each step. A first step to the variable selection consistency study has been proposed by Bach [12] and Sparsity Inequalities were given by Chesneau and Hebiri [44]. Another choice of penalty has been proposed with the Elastic-Net (Zou and Hastie [167]). It is in the same spirit that we shall treat the S-Lasso from a some theoretical point of view.

The rest of the chapter is organized as follows. In the next section, we present one way to solve the S-Lasso problem with the attractive property of piecewise linearity of its regularization path. Section 3.3 gives theoretical performances of the considered estimator such as consistency in variable selection and asymptotic normality when $p \leq n$ whereas consistency in estimation and variable selection in the high dimensional case are considered in Section 3.4. We also give an estimate of the effective degree of freedom of the S-Lasso estimator in Section 3.5. Then, we provide a way to control the variance of the estimator by scaling in Section 3.6 where a connection with soft-thresholding is also established. A generalization and comparative study to the Elastic-Net is done in Section 3.7. We finally give experimental results in Section 3.8 showing the S-Lasso performances against some popular methods. All proofs are postponed to an Appendix section.

3.2 The S-Lasso procedure

As described above, we define the S-Lasso estimator $\hat{\beta}^{SL}$ as the solution of the optimization problem (3.2) when the penalty function is:

$$\text{pen}(\beta) = \lambda|\beta|_1 + \mu \sum_{j=2}^{p} (\beta_j - \beta_{j-1})^2 , \tag{3.3}$$

where λ and μ are two positive parameters that control the smoothness of our estimator. For any vector $a = (a_1, \ldots, a_p)'$, we have used the notation $|a|_1 = \sum_{j=1}^{p} |a_j|$. Note that when $\mu = 0$, the solution is the Lasso estimator so that it appears as a special case of the S-Lasso estimator. Now we deal with the resolution of the S-Lasso problem (3.2)-(3.3) and its computational cost. From now on, we suppose w.l.o.g. that $X = (x_1, \ldots, x_n)'$ is standardized (that is $n^{-1}\sum_{i=1}^{n} x_{i,j}^2 = 1$ and $n^{-1}\sum_{i=1}^{n} x_{i,j} = 0$) and $Y = (y_1, \ldots, y_n)'$ is centered (that is $n^{-1}\sum_{i=1}^{n} y_i = 0$). The following lemma shows that the S-Lasso criterion can be expressed as a Lasso criterion by augmenting the data artificially.

Lemma 3.1. *Given the data set (X, Y) and (λ, μ). Define the extended dataset $(\widetilde{X}, \widetilde{Y})$ by*

$$\widetilde{X} = \frac{1}{\sqrt{1+\mu}} \begin{pmatrix} X \\ \sqrt{n\mu}\mathbf{J} \end{pmatrix} \quad \textit{and} \quad \widetilde{Y} = \begin{pmatrix} Y \\ \mathbf{0} \end{pmatrix},$$

where $\mathbf{0}$ *is a vector of size* p *containing only zeros and* $\mathbf{J}$ *is the* $p \times p$ *matrix*

$$\mathbf{J} = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 \\ 1 & -1 & \ddots & \ddots & \vdots \\ 0 & 1 & -1 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 1 & -1 \end{pmatrix}. \tag{3.4}$$

Let $r = \lambda/\sqrt{1+\mu}$ *and* $b = \sqrt{1+\mu}\,\beta$*. Then the S-Lasso criterion can be written*

$$\frac{n+p}{n} \left\| \widetilde{Y} - \widetilde{X} b \right\|_{n+p}^2 + r|b|_1.$$

Let $\hat{b}$ *be the minimizer of this Lasso-criterion, then*

$$\hat{\beta}^{SL} = \frac{1}{\sqrt{1+\mu}}\,\hat{b}.$$

This result is a consequence of simple algebra. Lemma 3.1 motivates the following comments on the S-Lasso procedure.

Remark 3.1 (*Regularization paths*)**.** *The S-Lasso modification of the LARS is an iterative algorithm. For a fixed* μ *(appearing in* (3.3)*), it constructs at each step an estimator based on the correlation between covariates and the current residue. Each step corresponds to a value of* λ*. Then for a fixed* μ*, we get the evolution of the S-Lasso estimator coefficients values when* λ *varies. This evolution describes the regularization paths of the S-Lasso estimator which are piecewise linear (Rosset and Zhu [126]). This property implies that the S-Lasso problem can be solved with the same computational cost as the ordinary least square (OLS) estimate using the Lasso modification version of the LARS algorithm.*

Remark 3.2 (*Implementation*)**.** *The number of covariates that the LARS algorithm and its Lasso version can select is limited by the number of rows in the matrix* X*. Applied to the augmented data* $(\widetilde{X}, \widetilde{Y})$ *introduced in Lemma 3.1, the Lasso modification of the LARS algorithm is able to select all the* p *covariates. Then we are no longer limited by the sample size as for the Lasso (Efron, Hastie, Johnstone, and Tibshirani [61]).*

3.3 Theoretical properties of the S-Lasso estimator when $p \leq n$

In this section we introduce the theoretical results according to the S-Lasso with a moderate sample size ($p \leq n$). We first provide rates of convergence of the S-Lasso estimator and show how through a control on the regularization parameters we can establish root-n consistency (asymptotic behavior of $(\hat{\beta}^{SL} - \beta^*)$ at the rate $\sqrt{n}$) and asymptotic normality of $\hat{\beta}^{SL}$, the

S-Lasso estimator. Then we look for variable selection consistency. More precisely, we give conditions under which the S-Lasso estimator succeeds in finding the set of the non-zero regression coefficients. We show that with a suitable choice of the tuning parameter (λ, μ), the S-Lasso is consistent in variable selection. All the results of this section are proved in Appendix A.

3.3.1 Asymptotic Normality

In this section, we allow the tuning parameters (λ, μ) to depend on the sample size n. We emphasize this dependence by adding a subscript n to these parameters. We also fix the number of covariates p. Let us note $\mathbb{I}(\cdot)$ the indicator function and define the sign function such that for any $x \in \mathbb{R}$, $\mathrm{Sgn}(x)$ equals 1, -1 or 0 respectively when x is bigger, smaller or equals 0. Knight and Fu [90] gave the asymptotic distribution of the Lasso estimator. We provide here the asymptotic distribution to the S-Lasso. Let $\Psi^n = n^{-1}X'X$, be Gram matrix, then

Theorem 3.1. *Given the data set (X, Y), assume the Gram matrix verifies*

$$\Psi^n \to \Psi, \quad \text{when } n \to \infty,$$

where Ψ is a positive definite matrix. If there exists a sequence v_n such that $v_n \to 0$ and the regularization parameters verify $\lambda_n v_n^{-1} \to \lambda \geq 0$ and $\mu_n v_n^{-1} \to \mu \geq 0$. Then, if $(\sqrt{n}v_n)^{-1} \to \kappa \geq 0$, we have

$$v_n^{-1}(\hat{\beta}^{SL} - \beta^*) \xrightarrow{\mathcal{D}} \underset{u \in \mathbb{R}^p}{\operatorname{argmin}} V(u), \quad \text{when } n \to \infty,$$

where

$$\begin{aligned} V(u) = -2\kappa u^T W + u^T \Psi u \quad &+ \quad \lambda \sum_{j=1}^{p} \left\{ u_j \,\mathrm{Sgn}(\beta_j^*)\mathbb{I}(\beta_j^* \neq 0) + |u_j|\mathbb{I}(\beta_j^* = 0) \right\} \\ &+ \quad 2\mu \sum_{j=2}^{p} \left\{ (u_j - u_{j-1})(\beta_j^* - \beta_{j-1}^*)\mathbb{I}(\beta_j^* \neq \beta_{j-1}^*) \right\}, \end{aligned}$$

with $W \sim \mathcal{N}(0, \sigma^2\Psi)$.

Remark 3.3. *When $\kappa \neq 0$ is a finite constant: in this case v_n^{-1} is $\mathcal{O}(\sqrt{n})$ so that the estimator $\hat{\beta}^{SL}$ is root-n consistent. Moreover when $\lambda = \mu = 0$, we obtain the following standard regressors asymptotic normality: $\sqrt{n}(\hat{\beta}^{SL} - \beta^*) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2\Psi^{-1})$.*
When $\kappa = 0$: in this case, the rate of convergence is slower than $\sqrt{n}$ so that we no longer have the optimal rate. Moreover the limit is not random anymore.

Note first that the correlation penalty does not alter the asymptotic bias when successive regression coefficients are equal. We also remark that the sequence v_n must be chosen properly as it determines our convergence rate. We would like v_n to be as close as possible to $1/\sqrt{n}$. This sequence is calibrated by the user such that $\lambda_n/v_n \to \lambda$ and $\mu_n/v_n \to \mu$.

3.3.2 Consistency in variable selection

In this section, variable selection consistency of the S-Lasso estimator is considered. For this purpose, we introduce the following sparsity sets: $\mathcal{A}^* = \{j : \beta_j^* \neq 0\}$ and $\hat{\mathcal{A}} = \{j : \hat{\beta}_j^{SL} \neq 0\}$. The set $\mathcal{A}^*$ consists of the non-zero coefficients in the regression vector β^*. The set $\hat{\mathcal{A}}$ consists of the non-zero coefficients in the S-Lasso estimator $\hat{\beta}_j^{SL}$ and is also called the active set of this estimator. Before stating our result, let us introduce some notations. For any vector $a \in \mathbb{R}^p$ and any set of indexes $\mathcal{B} \in \{1, \dots, p\}$, denote by $a_{\mathcal{B}}$ the restriction of the vector a to the indexes in $\mathcal{B}$. In the same way, if we note $|\mathcal{B}|$ the cardinal of the set $\mathcal{B}$, then for any $s \times q$ matrix M, we use the following convention: i) $M_{\mathcal{B},\mathcal{B}}$ is the $|\mathcal{B}| \times |\mathcal{B}|$ matrix consisting of the lines and rows of M whose indexes are in $\mathcal{B}$; ii) $M_{.,\mathcal{B}}$ is the $s \times |\mathcal{B}|$ matrix consisting of the rows of M whose indexes are in $\mathcal{B}$; iii) $M_{\mathcal{B},.}$ is the $|\mathcal{B}| \times q$ matrix consisting of the lines of M whose indexes are in $\mathcal{B}$. Moreover, we define $\widetilde{J}$ the $p \times p$ matrix $\mathbf{J'J}$ where $\mathbf{J}$ was defined in (3.4). Finally we define for $j \in \{1, \dots, p\}$, the quantity $\Omega_j = \Omega_j(\lambda, \mu, \mathcal{A}^*, \beta^*)$ by

$$\Omega_j = \Psi_{j,\mathcal{A}^*}(\Psi_{\mathcal{A}^*,\mathcal{A}^*} + \mu\widetilde{J}_{\mathcal{A}^*,\mathcal{A}^*})^{-1}\left(2^{-1}\operatorname{Sgn}(\beta^*_{\mathcal{A}^*}) + \frac{\mu}{\lambda}\widetilde{J}_{\mathcal{A}^*,\mathcal{A}^*}\beta^*_{\mathcal{A}^*}\right) - \frac{\mu}{\lambda}\widetilde{J}_{j,\mathcal{A}^*}\beta^*_{\mathcal{A}^*}, \tag{3.5}$$

where Ψ is defined as in Theorem 3.1. Now consider the following conditions: *for every* $j \in (\mathcal{A}^*)^c$

$$|\Omega_j(\lambda, \mu, \mathcal{A}^*, \beta^*)| < 1, \tag{3.6}$$

$$|\Omega_j(\lambda, \mu, \mathcal{A}^*, \beta^*)| \leq 1. \tag{3.7}$$

These conditions on the correlation matrix Ψ and the regression vector $\beta^*_{\mathcal{A}^*}$ are the analogues respectively of the sufficient and necessary conditions derived for the Lasso (Zou [166], Zhao and Yu [164] and Yuan and Lin [160]). Now we state the consistency results

Theorem 3.2. *If condition* (3.6) *holds, then for every couple of regularization parameters* (λ_n, μ_n) *such that* $\lambda_n \to 0$, $\lambda_n n^{1/2} \to \infty$ *and* $\mu_n \to 0$, *the S-Lasso estimator* $\hat{\beta}^{SL}$ *as defined in* (3.2)-(3.3) *is consistent in variable selection. That is*

$$\mathbb{P}(\hat{\mathcal{A}} = \mathcal{A}^*) \to 1, \quad \textit{when } n \to \infty.$$

Theorem 3.3. *If there exist sequences* (λ_n, μ_n) *such that* $\hat{\beta}^{SL}$ *converges to* β^* *and* $\hat{\mathcal{A}}$ *converges to* $\mathcal{A}^*$ *in probability, then condition* (3.7) *is satisfied.*

We just have established necessary and sufficient conditions to the selection consistency of the S-Lasso estimator. Due to the assumptions needed in Theorem 3.2 (more precisely $\lambda_n n^{1/2} \to \infty$), root-$n$ consistency and variable selection consistency cannot be treated here simultaneously. We may want to know if the S-Lasso estimator can be consistent in estimation with a slower rate than $n^{1/2}$ and consistent in variable selection in the same time.

Remark 3.4. *Here are special cases of conditions* (3.6)-(3.7).
When $\mu = 0$ and $\mu/\lambda = 0$: these conditions are exactly the sufficient and necessary conditions of the Lasso estimator. In this case, Yuan and Lin [160] showed that the condition (3.6) *becomes necessary and sufficient for the Lasso estimator consistency in variable selection.*
When $\mu = 0$ and $\mu/\lambda = \gamma \neq 0$ (that is, $\mu_n = \mathcal{O}(\lambda_n)$): in this case, condition (3.6) *becomes*

$$\sup_{j \in (\mathcal{A}^*)^c} |\Psi_{j,\mathcal{A}^*} \Psi^{-1}_{\mathcal{A}^*,\mathcal{A}^*} (2^{-1} \operatorname{Sgn}(\beta^*_{\mathcal{A}^*}) + \gamma \widetilde{J}_{\mathcal{A}^*,\mathcal{A}^*} \beta^*_{\mathcal{A}^*}) - \gamma \widetilde{J}_{j,\mathcal{A}^*} \beta^*_{\mathcal{A}^*}| < 1.$$

Here a good calibration of γ leads to consistency in variable selection:

- *if* $(\Psi_{j,\mathcal{A}^*} \Psi^{-1}_{\mathcal{A}^*,\mathcal{A}^*} \widetilde{J}_{\mathcal{A}^*,\mathcal{A}^*} - \widetilde{J}_{j,\mathcal{A}^*}) \beta^*_{\mathcal{A}^*} > 0$, *then γ must be chosen between* $-\dfrac{1 + 2^{-1} \Psi_{j,\mathcal{A}^*} \Psi^{-1}_{\mathcal{A}^*,\mathcal{A}^*} \operatorname{Sgn}(\beta^*_{\mathcal{A}^*})}{(\Psi_{j,\mathcal{A}^*} \Psi^{-1}_{\mathcal{A}^*,\mathcal{A}^*} \widetilde{J}_{\mathcal{A}^*,\mathcal{A}^*} - \widetilde{J}_{j,\mathcal{A}^*}) \beta^*_{\mathcal{A}^*}}$ *and* $\dfrac{1 - 2^{-1} \Psi_{j,\mathcal{A}^*} \Psi^{-1}_{\mathcal{A}^*,\mathcal{A}^*} \operatorname{Sgn}(\beta^*_{\mathcal{A}^*})}{(\Psi_{j,\mathcal{A}^*} \Psi^{-1}_{\mathcal{A}^*,\mathcal{A}^*} \widetilde{J}_{\mathcal{A}^*,\mathcal{A}^*} - \widetilde{J}_{j,\mathcal{A}^*}) \beta^*_{\mathcal{A}^*}}$.
- *if* $(\Psi_{j,\mathcal{A}^*} \Psi^{-1}_{\mathcal{A}^*,\mathcal{A}^*} \widetilde{J}_{\mathcal{A}^*,\mathcal{A}^*} - \widetilde{J}_{j,\mathcal{A}^*}) \beta^*_{\mathcal{A}^*} < 0$, *then γ must be chosen between the same quantities but with inversion in their order.*

When $\mu \neq 0$ and $\mu/\lambda = \gamma \neq 0$: this case is quite similar to the previous. In addition, it allows to have another control on the condition through a calibration with μ, so that condition (3.6) *can be satisfied with a better control.*

We conclude that if we sacrifice the optimal rate of convergence (i.e. root-n consistency), we are able through a proper choice of the tuning parameters (λ_n, μ_n) to get consistency in variable selection. Note that Zou [166] showed that the Lasso estimator cannot be consistent in variable selection even with a slower rate of convergence than $\sqrt{n}$. He then added weights to the Lasso (i.e. the adaptive Lasso estimator) in order to get Oracles Properties (that is both asymptotic normality and variable selection consistency). Note that we can easily adapt techniques used in the adaptive Lasso to provide a weighted S-Lasso estimator which achieved the Oracles Properties.

3.4 Theoretical results when dimension p is larger than sample size n

In this section, we propose to study the performance of the S-Lasso estimator in the high dimensional case. In particular, we provide a non-asymptotic bound on the squared risk.

We also provide bound on the estimation risk under the sup-norm (i.e., the l_∞-norm: $\|\hat{\beta}^{SL} - \beta^*\|_\infty = \sup_j |\hat{\beta}_j^{SL} - \beta_j^*|$). This last result helps us to provide a variable selection consistent estimator obtained through thresholding the S-Lasso estimator. The results of this section are proved in Appendix B.

3.4.1 Sparsity Inequality

Now we establish a Sparsity Inequality (SI) achieved by the S-Lasso estimator, that is a bound on the squared risk that takes into account the sparsity of the regression vector β^*. More precisely, we prove that the rate of convergence is $|\mathcal{A}^*| \log(n)/n$. For this purpose, we need some assumptions on the Gram matrix Ψ^n which is normalized in our setting. Recall that $X_j = (x_{1,j}, \ldots, x_{n,j})'$. Then we define the regularization parameters λ_n and μ_n in the following forms:

$$\lambda_n = \kappa_1 \sigma \sqrt{\frac{\log(p)}{n}}, \quad \text{and} \quad \mu_n = \kappa_2 \sigma^2 \frac{\sqrt{\log(p)}}{n}, \tag{3.8}$$

where $\kappa_1 > 2\sqrt{2}$ and κ_2 is positive constants. Let us define the maximal correlation quantity $\rho_1 = \max_{j \in \mathcal{A}^*} \max_{\substack{k \in \{1,\ldots,p\} \\ k \neq j}} |(\Psi^n)_{j,k}|$. Using these notations, we formulate the following assumptions:

- *Assumption (A1). The true regression vector β^* is such that there exists a finite constant L_1 such that:*

$$\beta_{\mathcal{A}^*}^{*'} \widetilde{J}_{\mathcal{A}^*,\mathcal{A}^*} \beta_{\mathcal{A}^*}^* \leq L_1 \log(p) \, |\mathcal{A}^*|, \tag{3.9}$$

  where $\widetilde{J} = \mathbf{J}'\mathbf{J}$ where $\mathbf{J}$ was defined in (3.4).

- *Assumption (A2). We have:*

$$\rho_1 \leq \frac{1}{16|\mathcal{A}^*|}. \tag{3.10}$$

Note that Assumption (A1) is not restrictive. A sufficient condition is that the larger non-zero component of $\beta_{\mathcal{A}^*}^*$ is bounded by $L_1 \log(p)$ which can be very large. Assumption (A2) is the well-known coherence condition considered by Bunea, Tsybakov, and Wegkamp [32], which has been introduced by Donoho, Elad, and Temlyakov [51]. Most of SIs provided in the literature use such a condition. We refer to Bunea, Tsybakov, and Wegkamp [32] for more details.
Theorem 3.4 below provides an upper bound for the squared error of the estimator $\hat{\beta}^{SL}$ and for its l_1 estimation error which takes into account the sparsity index $|\mathcal{A}^*|$.

Theorem 3.4. *Let us consider the linear regression model* (3.1). *Let $\hat{\beta}^{SL}$ be S-Lasso estimator. Let $\mathcal{A}^*$ be the sparsity set. Suppose that $p \geq n$. If Assumptions (A1)–(A2) hold,*

then with probability greater than $1 - u_{n,p}$*, we have*

$$\|X\hat{\beta}^{SL} - X\beta^*\|_n^2 \leq c_2 \frac{\log(p)|\mathcal{A}^*|}{n}, \tag{3.11}$$

and

$$|\hat{\beta}^{SL} - \beta^*|_1 \leq c_1 \sqrt{\frac{\log(p)}{n}} |\mathcal{A}^*|, \tag{3.12}$$

where $c_2 = (16\kappa_1^2 + L_1\kappa_2)\sigma^2$*,* $c_1 = (16\kappa_1 + L_1\kappa_1^{-1}\kappa_2)\sigma$ *and where* $u_{n,p} = p^{1-\kappa_1^2/8}$ *with* κ_1 *and* κ_2*, the constants appearing in* (3.8)*.*

The proof of Theorem 3.4 is based on the 'argmin' definition of the estimator $\hat{\beta}^{SL}$ and some technical concentration inequalities. Similar bounds were provided for the Lasso estimator by Bunea, Tsybakov, and Wegkamp [33]. Let us mention that the constants c_1 and c_2 are not optimal. We focused our attention on the dependency on n (and then on p and $|\mathcal{A}^*|$). It turns out that our results are near optimal. For instance, for the l_2 risk, the S-Lasso estimator reaches nearly the optimal rate $\frac{|\mathcal{A}^*|}{n} \log(\frac{p}{|\mathcal{A}^*|} + 1)$ up to a logarithmic factor (Bunea, Tsybakov, and Wegkamp [32, Theorem 5.1]).

3.4.2 Sup-norm bound and variable selection

Now we provide a bound on the sup-norm $\|\beta^* - \hat{\beta}^{SL}\|_\infty$. Thanks to this result, one may be able to define a rule in order to get a variable selection consistent estimator when $p \gg n$. That is, we can construct an estimator which succeeds to recover the support of β^* in high dimensional settings.

Small modifications are to be imposed to provide our selection results in this section. Let K_n be the symmetric $p \times p$ matrix defined by $K_n = \Psi^n + \mu_n \widetilde{J}$. Instead of Assumption (A2), we will consider the following

- *Assumption (A3). We assume that*

$$\max_{\substack{j,\, k \in \{1,\ldots,p\} \\ k \neq j}} |(K_n)_{j,k}| \leq \frac{1}{16|\mathcal{A}^*|}.$$

Remark 3.5. *Note that the matrix* $\widetilde{J}$ *is tridiagonal with its off-diagonal terms equal to* -1*. If we do not consider the diagonal terms, we remark that* Ψ^n *and* K_n *differ only in the terms on the second diagonals (i.e.,* $(K_n)_{j-1,j} \neq (\Psi^n)_{j-1,j}$ *for* $j = 2, \ldots, p$ *as soon as* $\mu_n \neq 0$*). Then, as we do not consider the diagonal terms in Assumptions (A2) and (A3), they differ only in the restriction they impose to terms on the second diagonals. Terms in the second diagonals of* Ψ^n *correspond to correlations between successive covariates. Then when high correlations exist between successive covariates, a suitable choice of* μ_n *makes Assumption (A3) satisfied while Assumptions (A2) does not. Hence, Assumption (A3) fits better with setup considered in this chapter.*

In the sequel, a convenient choice of the tuning parameter μ_n is $\mu_n = \kappa_3\sigma/\sqrt{n\log(p)}$, where $\kappa_3 > 0$ is a constant. Moreover, from Assumption (A1), we have $\beta^{*'}_{\mathcal{A}^*}\widetilde{J}_{\mathcal{A}^*,\mathcal{A}^*}\beta^*_{\mathcal{A}^*} \leq L_1\log(p)|\mathcal{A}^*|$. This inequality guarantees the existence of a constant $L_2 > 0$ such that $\|\widetilde{J}\beta^*\|_\infty \leq L_2\log(p)$.

Theorem 3.5. *Let us consider the linear regression model* (3.1). *Let* $\lambda_n = \kappa_1\sigma\sqrt{\log(p)/n}$ *and* $\mu_n = \kappa_3\sigma/\sqrt{n\log(p)}$ *with* $\kappa_1 > 2\sqrt{2}$ *and* $\kappa_3 > 0$. *Suppose that* $p \geq n$. *Under Assumptions (A1) and (A3) and with probability larger than* $1 - p^{1-\frac{\kappa_1^2}{8}}$, *we have*

$$\|\hat{\beta}^{SL} - \beta^*\|_\infty \leq \tilde{c}\sqrt{\frac{\log(p)}{n}},$$

where $\tilde{c}$ *equals to*

$$\frac{1}{1+\frac{B\sigma}{n}}\left(\frac{3}{4} + \frac{1}{\alpha-1} + \frac{4L_1B}{9\alpha^2A^2} + \frac{2L_1B}{3\alpha A^2} + \sqrt{\frac{2L_1B}{3\alpha(\alpha-1)A^2} + \frac{8L_1\,L_2B^2}{9\alpha(\alpha-1)A^4}\lambda_n} + (\frac{4L_2B}{3A^2} + \frac{L_2B}{A^2})\lambda_n\right).$$

Note that the leading term in $\tilde{c}$ is $\frac{3}{4} + \frac{1}{\alpha-1} + \frac{4L_1B}{9\alpha^2A^2} + \frac{2L_1B}{3\alpha A^2} + \sqrt{\frac{2L_1B}{3\alpha(\alpha-1)A^2}}$. One may find back the result obtained for the Lasso by setting L_1 to zero (Lounici [101]). Secondly, the calibration of μ_n aims at making the convergence rate under the sup-norm equal to $\sqrt{\log(p)/n}$. On one hand, the proof of Theorem 3.5 allows us to choose this parameter with a faster convergence to zero without affecting the rate of convergence. On the other hand, a more restrictive Assumption (A1) on $\beta^{*'}_{\mathcal{A}^*}\widetilde{J}_{\mathcal{A}^*,\mathcal{A}^*}\beta^*_{\mathcal{A}^*}$ and $\|\widetilde{J}\beta^*\|_\infty$ can be formulated in order to make μ_n converge slower to zero. If we let $\beta^{*'}_{\mathcal{A}^*}\widetilde{J}_{\mathcal{A}^*,\mathcal{A}^*}\beta^*_{\mathcal{A}^*} \leq L_1\,|\mathcal{A}^*|$ in Assumption (A1), we can set μ_n as $\mathcal{O}(\sqrt{\log(p)/n})$, the slower convergence we can get for μ_n.
Let us now provide a consistent version of the S-Lasso estimator for the selection purpose. Consider $\hat{\beta}^{ThSL} = (\hat{\beta}_1^{ThSL}, \ldots, \hat{\beta}_p^{ThSL})'$, the thresholded S-Lasso estimator defined by $\hat{\beta}_j^{ThSL} = \hat{\beta}_j^{SL}\mathbb{I}(\hat{\beta}_j^{SL} \geq \tilde{c}\sqrt{\log(p)/n})$ where $\tilde{c}$ is given in Theorem 3.5. This estimator consists of the S-Lasso estimator with its small coefficients reduced to zero. We then enforce the selection property of the S-Lasso estimator. Variable selection consistency of this estimator is established under one more restriction:

- *Assumption (A4). The smallest non-zero coefficient of* β^* *is such that there exists a constant* $c_l > 0$ *with*

$$\min_{j\in\mathcal{A}^*}|\beta_j^*| > c_l\sqrt{\frac{\log(p)}{n}}.$$

Assumption (A4) bounds from below the smallest regression coefficient in β^*. This is a common assumption to provide sign consistency in the high dimensional case. This condition appears by Meinshausen and Yu [114], Wainwright [152], Zhang and Huang [163] and Zhao and Yu [164] but with a larger (in term of sample size n dependence) and then more restrictive threshold. We refer to Lounici [101] for a longer discussion. An equivalent lower

bound in the regression coefficients can be found by Bunea [28], Lounici [101]. With this new assumption, we can state the following sign consistency result.

Theorem 3.6. *Let us consider the thresholded S-Lasso estimator $\hat{\beta}^{ThSL}$ as described above. Choose moreover $\lambda_n = \kappa_1 \sigma \sqrt{\log(p)/n}$ and $\mu_n = \kappa_3 \sigma / \sqrt{n \log(p)}$ with the positive constants $\kappa_1 > 2\sqrt{2}$ and κ_3. Under Assumptions (A1), (A3) and (A4), if $c_l > 2\tilde{c}$ with $\tilde{c}$ is given by Theorem 3.5, with probability larger than $1 - p^{1-\frac{\kappa_1^2}{8}}$, we have*

$$\mathrm{Sgn}(\hat{\beta}^{ThSL}) = \mathrm{Sgn}(\beta^*), \tag{3.13}$$

and then as $n \to +\infty$

$$\mathbb{P}(\mathrm{Sgn}(\hat{\beta}^{ThSL}) = \mathrm{Sgn}(\beta^*)) \to 1. \tag{3.14}$$

Remark 3.6. *As observed in Remark 3.5, Assumption (A3) is more easily satisfied when correlation exists between successive covariates. Then in situations where the correlation matrix Ψ^n is tridiagonal with its off-diagonal terms equal to δ with $\delta \in [0, 1]$, the constant κ_3 appearing in the definition of μ_n can be adjusted in order to get Assumption (A3) satisfied.*

3.5 Model Selection

As already mentioned [Remark 3.1 in Section 3.2], each step of the S-Lasso version of the LARS algorithm provides an estimator of β^*. In this section, we are interested in the choice of the best estimator according to its prediction accuracy. Recall that $Y = X\beta^* + \varepsilon$. Let us denote by $\hat{\beta}$ any estimator of β^* which depends on Y and let us set $\hat{y} = X\hat{\beta}$. We aim to minimize the true risk $\mathbb{E}\left\{\|X(\hat{\beta} - \beta^*)\|_n^2\right\}$. First, under the Gaussian assumption, we easily obtain

$$\mathbb{E}\left\{\|\hat{y} - X\beta^*)\|_n^2\right\} = \mathbb{E}\{\|Y - \hat{y}\|_n^2 - \sigma^2 + 2n^{-1} \sum_{i=1}^{n} \mathrm{Cov}(y_i, \hat{y}_i)\},$$

where the expectation is taken over the random variable Y. The last term in this equation is called *optimism* (Efron [58]). Moreover, Tibshirani [135] links this quantity to the *degree of freedom* $\mathrm{df}(\hat{y})$ of the estimator $\hat{y}$, so that the above equality becomes

$$\mathbb{E}\left\{\|\hat{y} - X\beta^*)\|_n^2\right\} = \mathbb{E}\left\{\|Y - \hat{y}\|_n^2 - \sigma^2 + 2n^{-1}\,\mathrm{df}(\hat{y})\sigma^2\right\}. \tag{3.15}$$

This final expression involves the degree of freedom which is unknown. Various methods exist to estimate the degree of freedom as bootstrap (Efron and Tibshirani [60]) or data perturbation methods (Shen and Ye [130]). We give an explicit form to the degree of freedom of the S-Lasso estimator $\hat{\beta}^{SL}$ as in the papers of Efron, Hastie, Johnstone, and Tibshirani [61] and Zou, Hastie, and Tibshirani [169]. This work provides an unbiased estimate of the true risk $\mathbb{E}\left\{\|X(\hat{\beta}^{SL}(\lambda, \mu) - \beta^*)\|_n^2\right\}$ for different pairs of the tuning parameter (λ, μ). Moreover this estimate is obtain with a small computational cost. In the numerical experiments,

we choose the pair (λ, μ) which minimizes this estimate of the true risk over all possible pairs.

Degrees of freedom: the degree of freedom is a quantity of interest in model selection. Before stating our result, let us introduce some useful properties about the regularization paths of the S-Lasso estimator:
Given a response Y, and a regularization parameter $\mu \geq 0$, there is a finite sequence $0 = \lambda^{(K)} < \lambda^{(K-1)} < \ldots < \lambda^{(0)}$ such that $\hat{\beta}^{SL} = \mathbf{0}$ for every $\lambda \geq \lambda^{(0)}$. In this notation, superscripts correspond to the steps of the S-Lasso version of the LARS algorithm.
Given a response Y, and a regularization parameter $\mu \geq 0$, for $\lambda \in (\lambda^{(k+1)}, \lambda^{(k)})$, the same covariates are used to construct the estimator. Let us note $\mathcal{A}_\zeta$ the active set for a fixed couple $\zeta = (\lambda, \mu)$ and $X_{.,\mathcal{A}_\zeta}$ the corresponding design matrix.

In what follows, we will use the subscript ζ to emphasize the fact that the considered quantity depends on ζ.

Theorem 3.7. *For fixed $\mu \geq 0$ and $\lambda > 0$, an unbiased estimate of the effective degree of freedom of the S-Lasso estimate is given by*

$$\widehat{\mathrm{df}}(\hat{y}_\zeta^{SL}) = \mathrm{Tr}\left[X_{.,\mathcal{A}_\zeta}\left(X'_{.,\mathcal{A}_\zeta}X_{.,\mathcal{A}_\zeta} + \mu\widetilde{J}_{\mathcal{A}_\zeta,\mathcal{A}_\zeta}\right)^{-1}X'_{.,\mathcal{A}_\zeta}\right],$$

where $\widetilde{J} = \mathbf{J}'\mathbf{J}$ is defined by

$$\widetilde{J} = \begin{pmatrix} 1 & -1 & 0 & \ldots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \ldots & 0 & -1 & 1 \end{pmatrix}. \tag{3.16}$$

As the estimation given in Theorem 3.7 has an important computational cost, we propose the following estimator of the degree of freedom of the S-Lasso estimator:

$$\widehat{\mathrm{df}}(\hat{y}_\zeta^{SL}) = \frac{|\mathcal{A}_\zeta| - 2}{1 + 2\mu} + \frac{2}{1 + \mu}, \tag{3.17}$$

which is very easy to compute. Let $\mathbf{I}_s$ be the $s \times s$ identity matrix where s is an integer. We found the former approximation of the degree of freedom under the orthogonal covariance matrix assumption (that is $n^{-1}X'X = \mathbf{I}_p$). Moreover we approximate the matrix $(\mathbf{I}_{|\mathcal{A}_\lambda|} + \mu\widetilde{J}_{\mathcal{A}_\lambda,\mathcal{A}_\lambda})$ by the diagonal matrix with $1 + \mu$ in the first and the last terms, and $1 + 2\mu$ in the others.

Remark 3.7 (*Comparison to the Lasso and the Elastic-Net*)**.** *A similar work leads to an estimation of the degree of freedom of the Lasso: $\widehat{\mathrm{df}}(\hat{y}_\zeta^{L}) = |\mathcal{A}_\zeta|$ and to an estimation of the degree of freedom of the Elastic-Net estimator: $\widehat{\mathrm{df}}(\hat{y}_\zeta^{EN}) = |\mathcal{A}_\zeta|/(1 + \mu)$. These*

approximations of the degrees of freedom provide the following comparison for a fixed ζ: $\widehat{\mathrm{df}}(\hat{y}_\zeta^{SL}) \leq \widehat{\mathrm{df}}(\hat{y}_\zeta^{EN}) \leq \widehat{\mathrm{df}}(\hat{y}_\zeta^{L})$. *A conclusion is that the S-Lasso estimator is the one which penalizes the smaller models, and the Lasso estimator the larger. As a consequence, the S-Lasso estimator should select larger models than the Lasso or the Elastic-Net estimator.*

3.6 The Normalized S-Lasso estimator

In this section, we look for a scaled S-Lasso estimator which would have better empirical performance than the original S-Lasso presented above. The idea behind this study is to better control shrinkage. Indeed, using the S-Lasso procedure (3.2)-(3.3) induces double shrinkage: one using the Lasso penalty and the other using the fusion penalty. We want to undo the shrinkage implied by the fusion penalty as shrinkage is already ensured by the Lasso penalty. We then suggest to study the S-Lasso criterion (3.2)-(3.3) without the Lasso penalty (i.e. with only the l_2-fusion penalty) in order to find the constant we have to scale with.

Define

$$\tilde{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|Y - X\beta\|_n^2 + \mu \sum_{j=2}^{p} (\beta_j - \beta_{j-1})^2 .$$

We easily obtain $\tilde{\beta} = ((X'X)/n + \mu\widetilde{J})^{-1}(X'Y)/n := \mathbf{L}^{-1}(X'Y)/n$ where $\widetilde{J}$ is given by (3.16). Moreover as the design matrix X is standardized, the symmetric matrix $\mathbf{L}$ can be written

$$\mathbf{L} = \begin{pmatrix} 1+\mu & \frac{X_1'X_2}{n} - \mu & \frac{X_1'X_3}{n} & \dots & \frac{X_1'X_p}{n} \\ & 1+2\mu & \frac{X_2'X_3}{n} - \mu & \dots & \vdots \\ & & \ddots & \ddots & \frac{X_{p-2}'X_p}{n} \\ & & & 1+2\mu & \frac{X_{p-1}'X_p}{n} - \mu \\ & & & & 1+\mu \end{pmatrix}.$$

In order to get rid of the shrinkage due to the fusion penalty, we force $\mathbf{L}$ to have ones (or close to a diagonal of ones) in its diagonal elements. Then we scale the estimator $\tilde{\beta}$ by a factor c. Here are two choice we will use in the following of this chapter: i) the first is $c = 1 + \mu$ so that the first and the last diagonal elements of $\mathbf{L}^{-1}$ become equal to one; ii) the second is $c = 1 + 2\mu$ which offers the advantage that all the diagonal elements of $\mathbf{L}^{-1}$ become equal to one except the first and the last. This second choice seems to be more appropriate to undo this extra shrinkage and specially in high dimensional problem.

We first give a generalization of Lemma 3.1.

Lemma 3.2. *Given the dataset (X, Y) and (λ_1, μ). Define the augmented dataset $(\widetilde{X}, \widetilde{Y})$*

by

$$\widetilde{X} = \nu_1^{-1} \begin{pmatrix} X \\ \sqrt{n\mu}\mathbf{J} \end{pmatrix} \quad \textit{and} \quad \widetilde{Y} = \begin{pmatrix} Y \\ \mathbf{0} \end{pmatrix},$$

where ν_1 *is a constant which depends only on* μ *and* $\mathbf{J}$ *is given by* (3.4)*. Let* $r = \lambda/\nu_1$ *and* $b = (\nu_2/c)\beta$ *where* ν_2 *is a constant which depends only on* μ*, and* c *is the scaling constant which appears in the previous study. Then the S-Lasso criterion can be written*

$$\frac{n+p}{n} \left\| \widetilde{Y} - \widetilde{X}b \right\|_{n+p}^2 + r|b|_1. \tag{3.18}$$

Let $\hat{b}$ *be the minimizer of this Lasso-criterion, then we define the Scaled Smooth Lasso (SS-Lasso) by*

$$\hat{\beta}^{SSL} = \hat{\beta}^{SSL}(\nu_1, \nu_2, c) = (c/\nu_2)\,\hat{b}.$$

Moreover, let $\widetilde{J} = \mathbf{J}'\mathbf{J}$*. Then we have*

$$\hat{\beta}^{SSL} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{\nu_2}{\nu_1} \beta' \left(\frac{\frac{X'X}{n} + \mu\widetilde{J}}{c} \right) \beta - 2\frac{Y'X}{n}\beta + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{3.19}$$

Equation (3.19) is only a rearrangement of the Lasso criterion (3.18). The SS-Lasso expression (3.19) emphasizes the importance of the scaling constant c. In a way, the SS-Lasso estimator stabilizes the Lasso estimator $\hat{\beta}^L$ (criterion (3.18) based in (X, Y) instead of $(\widetilde{X}, \widetilde{Y})$) as we have

$$\hat{\beta}^L = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \beta' \left(\frac{X'X}{n} \right) \beta - 2\frac{Y'X}{n}\beta + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

The choice of ν_1 and ν_2 should be linked to this scaling constant c in order to get better empirical performances and to have less parameters to calibrate. Let us define some specific cases. i) *Case 1: When* $\nu_1 = \nu_2 = \sqrt{1+\mu}$ *and* $c = 1$: this is the "original" S-Lasso estimator as seen in Section 3.2. ii) *Case 2: When* $\nu_1 = \nu_2 = \sqrt{1+\mu}$ *and* $c = 1+\mu$: we call this scaled S-Lasso estimator Normalized Smooth Lasso (NS-Lasso) and we note it $\hat{\beta}^{NSL}$. In this case, we have $\hat{\beta}^{NSL} = (1+\mu\hat{\beta}^{SL})$. iii) *Case 3: When* $\nu_1 = \nu_2 = \sqrt{1+2\mu}$ *and* $c = 1+2\mu$: we call this scaled version Highly Normalized Smooth Lasso (HS-Lasso) and we note it $\hat{\beta}^{HSL}$.

Others choices are possible for ν_1 and ν_2 in order to better control shrinkage. For instance we can consider a compromise between the NS-Lasso and the HS-Lasso by defining $\nu_1 = 1+\mu$ and $\nu_2 = 1+2\mu$.

Remark 3.8 (*Connection with Soft Thresholding*)**.** *Let us consider the limit case of the NS-Lasso estimator. Note* $\hat{\beta}_\infty^{NSL} = \lim_{\mu\to\infty} \hat{\beta}^{NSL}$*, then using* (3.19)*, we have*

$$\hat{\beta}_\infty^{NSL} = \underset{\beta}{\operatorname{argmin}}\{\beta'\beta - 2Y'X\beta + \lambda|\beta|_1\}.$$

As a consequence, $(\hat{\beta}^{NSL}_{\infty})_j = \left(|Y'X_j| - \frac{\lambda}{2}\right)_+ \mathrm{Sgn}(Y'X_j)$ *which is the Univariate Soft Thresholding (Donoho and Johnstone [54]). Hence, when* $\mu \to \infty$*, the NS-Lasso works as if all the covariates were independent. The Lasso, which corresponds to the NS-Lasso when* $\mu = 0$*, often fails to select covariates when high correlations exist between relevant and irrelevant covariates. It seems that the NS-Lasso is able to avoid such problem by increasing* μ *and working as if all the covariates were independent. Then for a fixed* λ*, the control of the regularization parameter* μ *appears to be crucial. When we vary it, the NS-Lasso bridges the Lasso and the Soft Thresholding.*

3.7 Extension and comparison

All results obtained in the present chapter can be generalized to all penalized least square estimators for which the penalty term can be written as:

$$\mathrm{pen}(\beta) = \lambda|\beta|_1 + \beta' M \beta, \tag{3.20}$$

where M is $p \times p$ matrix. In particular, our study can be extended for instance to the Elastic-Net estimator with the special choice $M = \mathbf{I}_p$. Such an observation underlines the superiority of the S-Lasso estimator on the Elastic-Net in some situations. Indeed, let us consider the variable selection consistency in the high dimensional setting (cf. Section 3.4.2). Regarding the Elastic-Net, Assumption (A3) becomes

- *Assumption (A3-EN). We assume that*

$$\max_{\substack{j,\, k\in\{1,\dots,p\}\\ k\neq j}} |(\Psi^n)_{j,k} + \mu_n \mathbf{I}_p| \leq \frac{1}{16|\mathcal{A}^*|}. \tag{3.21}$$

Since the identity matrix is diagonal and since the maximum in (3.21) is taken over indexes $k \neq j$, condition (3.21) reduces to $\max_{\substack{j,\, k\in\{1,\dots,p\}\\ k\neq j}} |(\Psi^n)_{j,k}| \leq \frac{1}{16|\mathcal{A}^*|}$. This makes Assumption (A3-EN) similar to the assumption needed to get the variable selection consistency of the Lasso estimator (Bunea [28]). Hence, we get no gain to use the Elastic-Net in a variable selection consistency point of view in our framework. This ables us to think that the S-Lasso outperforms the Elastic-Net at least on examples as the one in Remark 3.6. Recently, Jia and Yu [85] studied the variable selection consistency of the Elastic-Net under an assumption called *Elastic Irrepresentable Condition*:

- *(EIC). There exists a positive constant* θ *such that for any* $j \in (\mathcal{A}^*)^c$

$$|\Psi_{j,\mathcal{A}^*}(\Psi_{\mathcal{A}^*,\mathcal{A}^*} + \mu \mathbf{I}_{\mathcal{A}^*})^{-1}\left(2^{-1}\,\mathrm{Sgn}(\beta^*_{\mathcal{A}^*}) + \frac{\mu}{\lambda}\beta^*_{\mathcal{A}^*}\right)| \leq 1 - \theta.$$

This condition can be seen as a generalization of the *Irrepresentable Condition* involved in the Lasso variable selection consistency.
Let us discuss how the two assumptions can be compared in the case $p \gg n$. First, note that Assumption (A3-EN), as well as EIC suggests low correlations between covariates. Moreover Assumption (A1), (A4) and (A3-EN) seem more restrictive than EIC as all the correlations are constrained in (3.21). However, EIC is harder to interpret in term of the coefficients of the regression vector β^*. It also depends on the sign of β^*. The main difference is that the selection consistency result in the present chapter holds uniformly on the solutions of the Elastic Net criterion while the result from Jia and Yu [85] hinges upon the existence of a consistent solution for variable selection. Obviously, this is more restrictive as we are certain to provide the sign-consistent solution under the EIC. Finally, we have also provided results on the sup-norm and sparsity inequalities on the squared risk of our estimators. Such results are new for estimators defined with the penalty (3.20), including the S-Lasso and the Elastic-Net.

3.8 Experimental results

In the present section we illustrate the good prediction and selection properties of the NS-Lasso and the HS-Lasso estimators. For this purpose, we compare it to the Lasso and the Elastic-Net. It appears that S-Lasso is a good challenger to the Elastic-Net even when large correlations between covariates exist (Zou and Hastie [167]). We further show that in most cases, our procedure outperforms the Elastic-Net and the Lasso when we consider the ratio between the relevant selected covariates and irrelevant selected covariates.

Simulations:
Data. Four simulations are generated according to the linear regression model

$$y = x\beta^* + \sigma\varepsilon, \quad \varepsilon \sim \mathcal{N}(0,1),\ x = (X_1, \ldots, X_p) \in \mathbb{R}^p.$$

The first and the second examples were introduced in the original Lasso paper of Tibshirani [135]. The third simulation creates a grouped covariates situation. It was introduced by Zou and Hastie [167] and aims to point the efficiency of the Elastic-Net compared to the Lasso. The last simulation introduces large correlation between successive covariates.

(a) In this example, we simulate 20 observations with 8 covariates. The true regression vector is $\beta^* = (3, 1.5, 0, 0, 2, 0, 0, 0)'$ so that only three covariates are truly relevant. Let $\sigma = 3$ and the correlation between X_j and X_k such that $\text{Cov}(X_j, X_k) = 2^{-|j-k|}$.

(b) The second example is the same as the first one, except that we generate 50 observations and that $\beta_j^* = 0.85$ for every $j \in \{1, \dots, 8\}$ so that all the covariates are relevant.

(c) In the third example, we simulate 50 data with 40 covariates. The true regression vector is such that $\beta_j^* = 3$ for $j = 1, \dots, 15$ and $\beta_j^* = 0$ for $j = 16, \dots, 40$. Let $\sigma = 15$ and the covariates generated as follows:

$$\begin{aligned} X_j &= Z_1 + \varepsilon_j, \quad Z_1 \sim \mathcal{N}(0,1), \qquad j = 1, \dots, 5, \\ X_j &= Z_2 + \varepsilon_j, \quad Z_2 \sim \mathcal{N}(0,1), \qquad j = 6, \dots, 10, \\ X_j &= Z_3 + \varepsilon_j, \quad Z_3 \sim \mathcal{N}(0,1), \qquad j = 11, \dots, 15, \end{aligned}$$

where ε_j, $j = 1, \dots, 15$, are i.i.d. $\mathcal{N}(0, 0.01)$ variables. Moreover for $j = 16, \dots, 40$, the X_j's are i.i.d $\mathcal{N}(0,1)$ variables.

(d) In the last example, we generate 50 data with 30 covariates. The true regression vector is such that

$$\begin{aligned} \beta_j = &\quad 3 - 0.1j \qquad && j = 1, \dots, 10, \\ \beta_j = &\quad -5 + 0.3j \qquad && j = 20, \dots, 25, \\ \beta_j = &\quad 0 \qquad && \text{for the others } j. \end{aligned}$$

The noise is such that $\sigma = 9$ and the correlations are such that $\text{Cov}(X_j, X_k) = \exp\left(-\frac{|j-k|}{2}\right)$ for $(j,k) \in \{11, \dots, 25\}^2$ and the others covariates are i.i.d. $\mathcal{N}(0,1)$, also independent from $X_{11}, \dots, X_{25}$. In this model there are big correlation between relevant covariates and even between relevant and irrelevant covariates.

Validation. The selection of the tuning parameters λ and μ is based on the minimization of a BIC-type criterion (Schwartz [127]). For a given $\hat{\beta}$ the associated BIC error is defined as:

$$\text{BIC}(\hat{\beta}) = \|Y - X\hat{\beta}\|_n^2 + \frac{\log(n)\sigma^2}{n} \widehat{\text{df}}(\hat{\beta}),$$

where $\widehat{\text{df}}(\hat{\beta})$ is given by (3.17) if we consider the S-Lasso and denotes its analogous quantities if we consider the Lasso or the Elastic-Net. Such a criterion provides an accurate estimator which enjoys good variable selection properties (Shao [129] and Yang [156]). In simulation studies, for each replication, we also provide the Mean Square Error (MSE) of the selected estimator on a new and independent dataset with the same size as training set (that is n). This gives an information on the robustness of the procedures.

Interpretations. All the results exposed here are based on 200 replications. Figure 3.1 and Figure 3.2 give respectively the BIC error and the test error of the considered proce-

Method	Example (a)	Example (b)	Example (c)	Example (d)
Lasso	3.8 [±0.1]	6.5 [±0.1]	6 [±0.1]	18.4 [±0.2]
E-Net	4.9 [±0.1]	6.9 [±0.1]	15.9 [±0.1]	20.5 [±0.2]
NS-Lasso	3.9 [±0.1]	6.5 [±0.1]	15.3 [±0.2]	18.9 [±0.2]
HS-Lasso	3.5 [±0.1]	5.9 [±0.1]	15 [±0.1]	18.1 [±0.2]

Table 3.1: Mean of the number of non-zero coefficients [and its standard error] selected respectively by the Lasso, the Elastic-Net (E-Net), the Normalized Smooth Lasso (NS-Lasso) and the Highly Smooth Lasso (HS-Lasso) procedures.

Method	Example (a)	Example (c)	Example (d)
Lasso	2.3 [±0.1]	2.9 [±0.1]	4.7 [±0.2]
E-Net	1.7 [±0.1]	13.1 [±0.3]	3.4 [±0.2]
NS-Lasso	2.5 [±0.1]	13.5 [±0.3]	6.8 [±0.3]
HS-Lasso	1.79 [±0.1]	11.4 [±0.3]	6.4 [±0.3]

Table 3.2: Mean of the ratio between the number of relevant covariates and the number of noise covariates (SNR) [and its standard error] that each of the Lasso, the Elastic-Net, the NS-Lasso and the HS-Lasso procedures selected.

dures in each example. According to the selection part, Figure 3.3 shows the frequencies of selection of each covariate for all the procedures, and Table 3.1 shows the mean of the number of non-zeros coefficients that each procedure selected. Finally for each procedure, Table 3.2 gives the ratio between the number of relevant covariates and the number of noise covariates that the procedures selected. Let us call SNR this ratio. Then we can express this ratio as

$$\text{SNR} = \frac{\sum_{j \in \hat{\mathcal{A}}} \mathbb{I}(j \in \mathcal{A}^*)}{\sum_{j \in \hat{\mathcal{A}}} \mathbb{I}(j \notin \mathcal{A}^*)}.$$

This is a good indication of the selection power of the procedures.

As the Lasso is a special case of the S-Lasso and the Elastic-Net, the Lasso BIC error (Figure 3.1) is always larger than the BIC error for the other methods. These two seem to have equivalent BIC errors. When considering the test error (Figure 3.2), it seems again that all the procedures are similar in all of the examples. They manage to produce good prediction independently of the sparsity of the model.

The more attractive aspect concerns variable selection. For this purpose we treat each example separately.

Example (a): the Elastic-Net selects a model which is too large (Table 3.1). This is reflected by the worst SNR (Table 3.2). As a consequence, we can observe in Figure 3.3 that it also includes the second covariate more often than the other procedures. This is due to the "grouping effect" as the first covariate is relevant. For similar reasons, the S-Lasso often selects the second covariate. However, this covariate is less selected than by the Elastic-Net

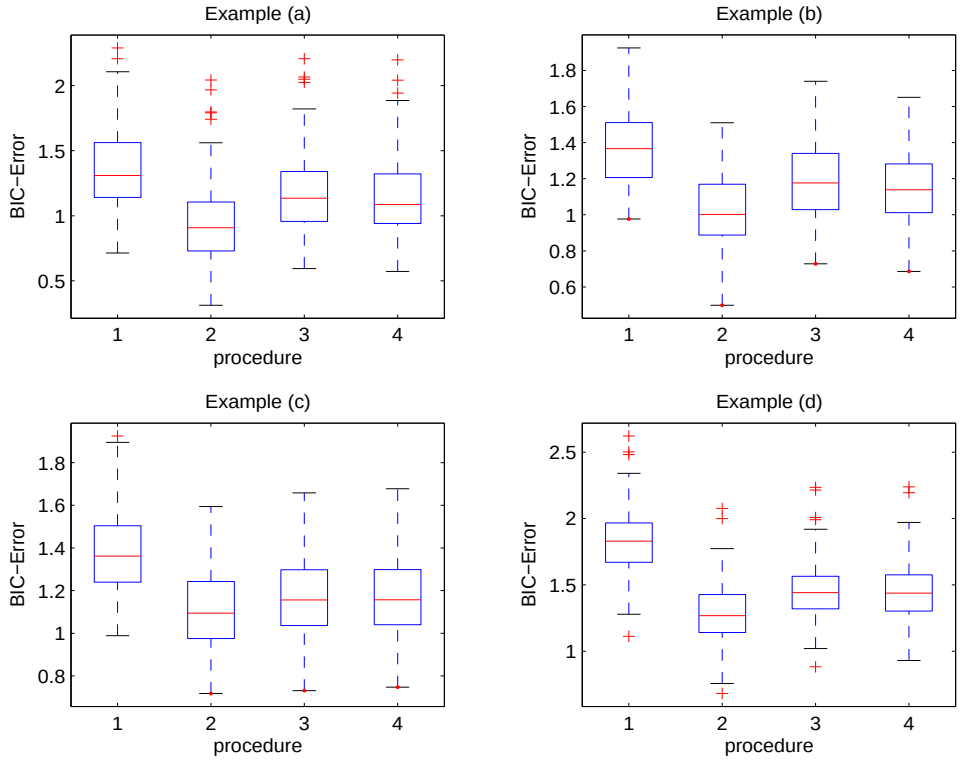

Figure 3.1: BIC error in each example. For each plot, we construct the boxplot for the procedure 1 = Lasso; 2 = Elastic-Net; 3 = NS-Lasso; 4 = HS-Lasso

as the S-Lasso seems to be a little bit disturbed by the third covariate which is irrelevant. This aspect of the S-Lasso procedure is also present in the selection of the covariate 5 as its neighbor covariates 4 and 6 are irrelevant. We can also observe that the S-Lasso procedure is the one which selects less often irrelevant covariates when these covariates are far away from relevant ones (in term of indices distance). Finally, even if the Lasso procedure selects less often the relevant covariates than the Elastic-Net and the S-Lasso procedures, it also has as good SNR. The Lasso presents good selection performances in this example.

Example (b): we can see in Figure 3.3 how the S-Lasso and Elastic-Net selection depends on how the covariates are ranked. They both select more covariates in the middle (that is covariates 2 to 7) than the ones in the borders (covariates 1 and 8) than the Lasso. We also remark that this aspect is more emphasized for the S-Lasso than for the Elastic-Net.

Example (c): the Lasso procedure performs poorly. It selects more noise covariates and less relevant ones than the other procedures (Figure 3.3). It also has the worst SNR (Table 3.2). In this example, Figure 3.3 also shows that the Elastic-Net selects more often relevant covariates than the S-Lasso procedures but it also selects more noise covariates than the NS-lasso procedure. Then even if the Elastic-Net has very good performance in variable

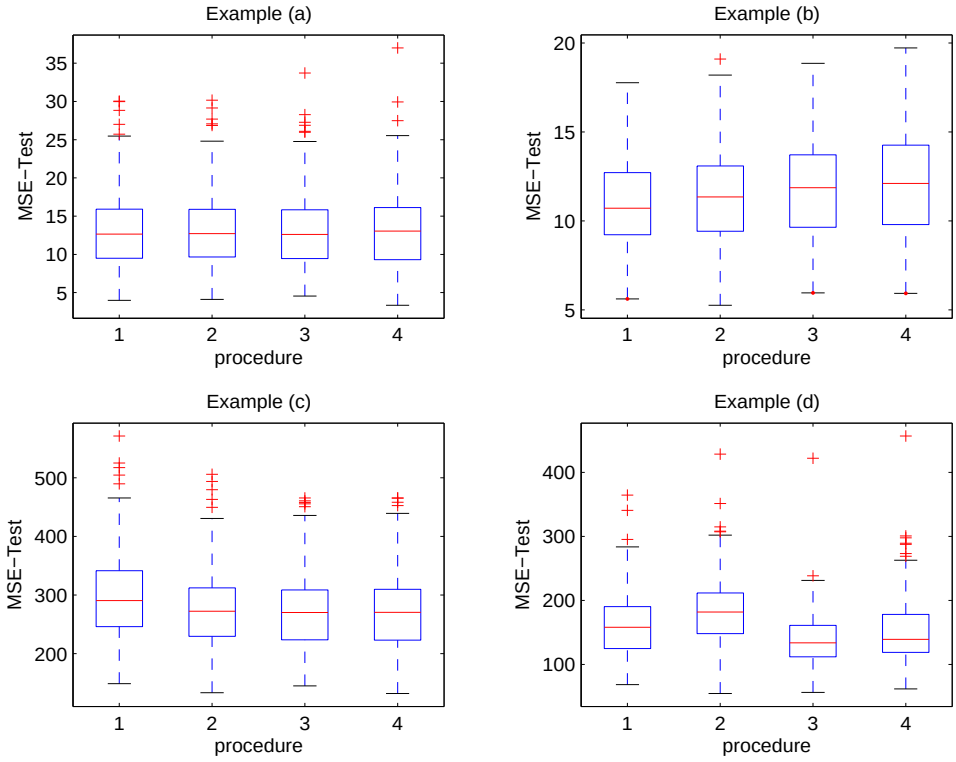

Figure 3.2: Test Error in each example. For each plot, we construct the boxplot for the procedure 1 = Lasso; 2 = Elastic-Net; 3 = NS-Lasso; 4 = HS-Lasso

selection, the NS-Lasso procedure has similar performances with a close SNR (Table 3.2). The NS-Lasso appears to have very good performance in this example. However, it selects again less often relevant covariates at the border than the Elastic-Net.

Example (d): we decompose the study into two parts. First, the independent part which considers covariates $X_1, \ldots, X_{10}$ and $X_{26}, \ldots, X_{30}$. The second part considers the other covariates which are dependent. Regarding the independent covariates, Figure 3.3 shows that all the procedures perform roughly in the same way, though the S-Lasso procedure enjoys a slightly better selection (in both relevant and noise group of covariates). For the dependent and relevant covariates, the Lasso performs worst than the other procedures. It selects clearly less often these relevant covariates. As in example (c), the reason is that the Lasso modification of the LARS algorithm tends to select only one representative of a group of highly correlated covariates. The high value of the SNR for the Lasso (when compared to the Elastic-Net) is explained by its good performance when it treat noise covariates. In this example the Elastic-Net correctly selects relevant covariates but it is also the procedure which selects the more noise covariates and has the worst SNR. We also note that both the NS-Lasso and HS-Lasso outperform the Lasso and Elastic-Net. This gain is emphasized

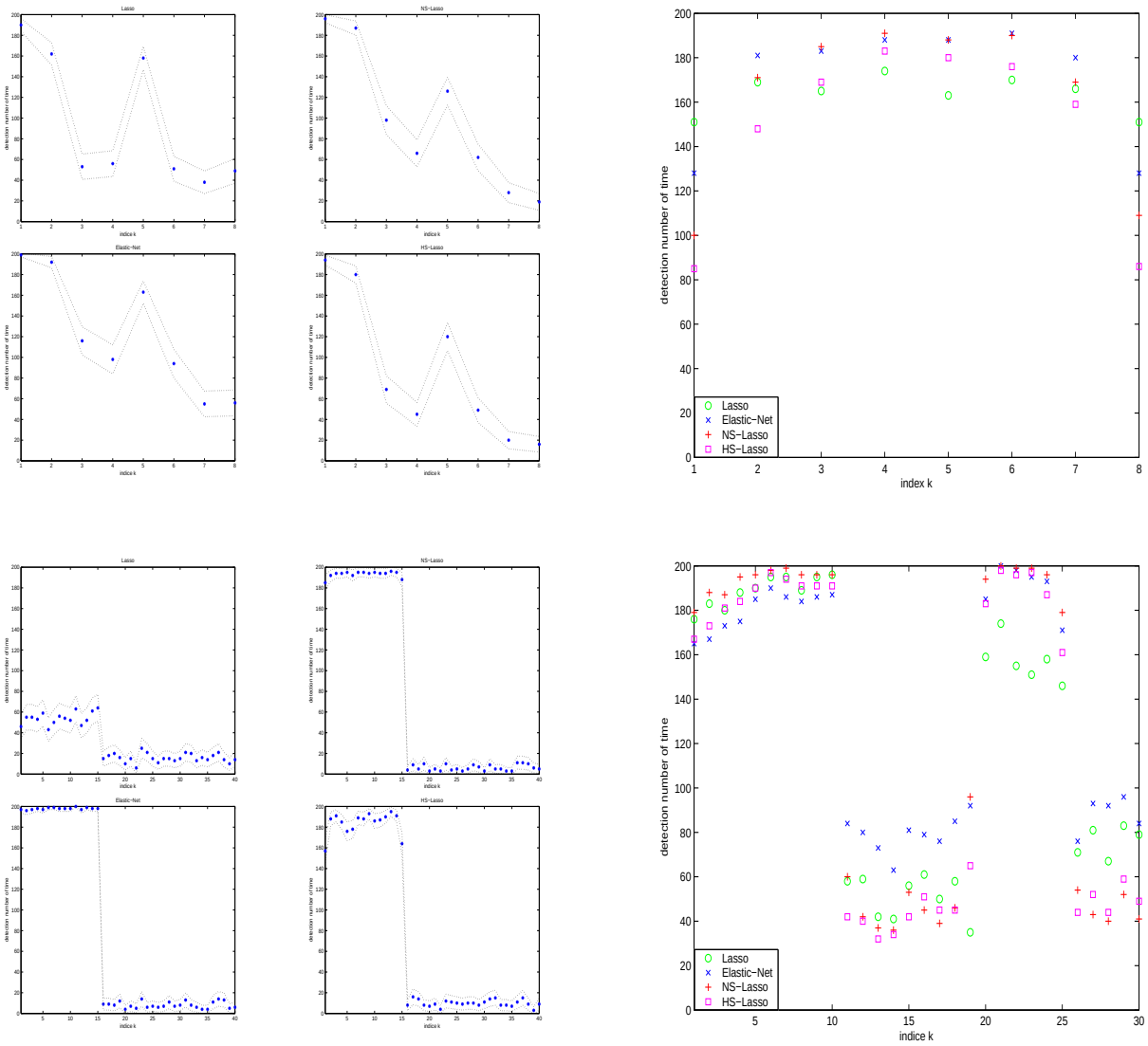

Figure 3.3: Number of covariates detections for each procedure in all the examples (Top-Left: Example (a); Top-Right: Example (b); Bottom-Left: Example (c); Bottom-Right: Example (d))

especially in the center of the groups. Observe that for the covariates X_{20}, X_{21}, X_{25} and X_{26} (that is the borders), the NS-Lasso and HS-Lasso have slightly worst performance than in the center of the groups. This is again due to the attraction we imposed by the fusion penalty (3.3) in the S-Lasso criterion.

Conclusion of the experiments. The S-Lasso procedure seems to respond to our expectations. Indeed, when successive correlations exist, it tends to select the whole group of these relevant covariates and not only one representing the group as done by the Lasso procedure. It also appears that the S-Lasso procedure has very good selection properties according to both relevant and noise covariates. However it has slightly worst performance in the borders than in the centers of groups of covariates (due to attractions of irrelevant covariates). It almost always has a better SNR than the Elastic-Net, so we can take it as a good challenger for this procedure.

3.9 Conclusion

In this chapter, we introduced a new procedure called the Smooth-Lasso which takes into account correlation between successive covariates. We established several theoretical results. The main conclusions are that when $p \leq n$, the S-Lasso is consistent in variable selection and asymptotically normal with a rate lower than $\sqrt{n}$. In the high dimensional setting, we provided a condition related to the coherence mutual condition, under which the thresholded version of the Smooth-Lasso is consistent in variable selection. This condition is fulfilled when correlations between successive covariates exist. Moreover, simulation studies showed that normalized versions of the Smooth-Lasso have nice properties of variable selection which are emphasized when high correlations exist between successive covariates. It appears that the Smooth-Lasso almost always outperforms the Lasso and is a good challenger of the Elastic-Net.

3.10 Proofs in the case $p \leq n$

Since the matrix $\Psi^n + \mu_n \tilde{J}$ plays a crucial role in the proves, we use to shorten the notation $K_n = \Psi^n + \mu_n \tilde{J}$ and when $p \leq n$ we define $K = \Psi + \mu \tilde{J}$, its limit.
In this appendix we prove the results when $p \leq n$.

Proof of Theorem 3.1. Let $\mathcal{R}_n$ be

$$\begin{aligned}\mathcal{R}_n(u) = \|Y - X(\beta^* + v_n u)\|_n^2 &+ \lambda_n \sum_{j=1}^{p} |\beta_j^* + v_n u_j| \\ &+ \mu_n \sum_{j=2}^{p} \left(\beta_j^* - \beta_{j-1}^* + v_n(u_j - u_{j-1})\right)^2,\end{aligned}$$

for $u = (u_1, \ldots, u_p)' \in \mathbb{R}^p$ and let $\hat{u} = \operatorname{argmin}_u \mathcal{R}_n(u)$. Note that $\hat{u} = v_n^{-1}(\hat{\beta}^{SL} - \beta^*)$. Let $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$, we then have

$$\begin{aligned}\mathcal{R}_n(u) - \mathcal{R}_n(0) &=: \overline{V}_n(u) \\ &= v_n^2\, u' \left(\frac{X'X}{n}\right) u - 2\frac{v_n}{\sqrt{n}}\frac{\varepsilon' X}{\sqrt{n}} u + v_n \lambda_n \sum_{j=1}^{p} v_n^{-1} \left(|\beta_j^* + v_n u_j| - |\beta_j^*|\right) \\ &\quad + v_n \mu_n \sum_{j=2}^{p} v_n^{-1} \left\{\left(\beta_j^* - \beta_{j-1}^* + v_n(u_j - u_{j-1})\right)^2 - \left(\beta_j^* - \beta_{j-1}^*\right)^2\right\} \\ &= v_n^2 \left[u' \left(\frac{X'X}{n}\right) u - \frac{2}{v_n\sqrt{n}}\frac{\varepsilon' X}{\sqrt{n}} u + \frac{\lambda_n}{v_n} \sum_{j=1}^{p} v_n^{-1} \left(|\beta_j^* + v_n u_j| - |\beta_j^*|\right) \right. \\ &\quad \left. + \frac{\mu_n}{v_n} \sum_{j=2}^{p} v_n^{-1} \left\{\left(\beta_j^* - \beta_{j-1}^* + v_n(u_j - u_{j-1})\right)^2 - \left(\beta_j^* - \beta_{j-1}^*\right)^2\right\} \right] \\ &= v_n^2 V_n(u).\end{aligned}$$

Note that $\hat{u} = \operatorname{argmin}_u \mathcal{R}_n(u) = \operatorname{argmin}_u V_n(u)$, we then have to consider the limit distribution of $V_n(u)$. First, we have $\frac{X'X}{n} \to \Psi$. Moreover, as $1/(v_n\sqrt{n}) \to \kappa$ and as given X, the random variable $\frac{X'\varepsilon}{\sqrt{n}} \xrightarrow{\mathcal{D}} W$, with $W \sim \mathcal{N}(0, \sigma^2\Psi)$, the Slutsky theorem implies that

$$\frac{2}{v_n\sqrt{n}}\frac{\varepsilon' X}{\sqrt{n}} u \xrightarrow{\mathcal{D}} 2\kappa W' u.$$

Now we treat the last two terms. If $\beta_j^* \neq 0$,

$$v_n^{-1} \left(|\beta_j^* + v_n u_j| - |\beta_j^*|\right) \to u_j \operatorname{Sgn}(\beta_j^*),$$

and is equal to $|u_j|$ otherwise. Then, as

$$\frac{\lambda_n}{v_n} \sum_{j=1}^{p} v_n^{-1} \left(|\beta_j^* + v_n u_j| - |\beta_j^*|\right) \to \lambda \sum_{j=1}^{p} \left\{u_j \operatorname{Sgn}(\beta_j^*) \mathbb{I}(\beta_j^* \neq 0) + |u_j| \mathbb{I}(\beta_j^* = 0)\right\},$$

For the remaining term, we show that if $\beta_j \neq \beta_{j-1}$,

$$v_n^{-1} \left\{\left(\beta_j^* - \beta_{j-1}^* + v_n(u_j - u_{j-1})\right)^2 - \left(\beta_j^* - \beta_{j-1}^*\right)^2\right\} \to 2(u_j - u_{j-1})(\beta_j^* - \beta_{j-1}^*),$$

and is equal to $\frac{(u_j-u_{j-1})^2}{n}$ otherwise. But μ_n converge to 0, implies that

$$\frac{\mu_n}{v_n}\sum_{j=2}^{p} v_n^{-1}\left\{\left(\beta_j^*-\beta_{j-1}^*+v_n(u_j-u_{j-1})\right)^2-\left(\beta_j^*-\beta_{j-1}^*\right)^2\right\}\rightarrow$$

$$2\mu\sum_{j=2}^{p}\left\{(u_j-u_{j-1})(\beta_j^*-\beta_{j-1}^*)\mathbb{I}(\beta_j^*\neq\beta_{j-1}^*)\right\}.$$

Therefore we have $V_n(u)\rightarrow V(u)$ in distribution, for every $u\in\mathbb{R}^p$. And since Ψ is a positive defined matrix, $V(u)$ has a unique minimizer. Moreover as $V_n(u)$ is convex, standard M-estimation results of van der Vaart [144] lead to: $\hat{u}_n\rightarrow \operatorname{argmin}_u V(u)$. □

Proof of Theorem 3.2. We begin by giving two results which we will use in our proof. The first one concerns the optimality conditions of the S-Lasso estimator. Recall that by definition

$$\hat{\beta}^{SL}=\underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}}\,\|Y-X\beta\|_n^2+\lambda_n|\beta|_1+\mu_n\beta'\widetilde{J}\beta.$$

Note $f(a)|_{a=a_0}$ the evaluation of the function f at the point a_0. As the above problem is a non-differentiable convex problem, classical tools lead to the following optimality conditions for the S-Lasso estimator:

Lemma 3.3. *The vector $\hat{\beta}^{SL}=(\hat{\beta}_1^{SL},\ldots,\hat{\beta}_p^{SL})'$ is the S-Lasso estimate as defined in* (3.2)-(3.3) *if and only if*

$$\left.\frac{\|Y-X\beta\|_n^2+\mu_n\beta'\widetilde{J}\beta}{d\beta_j}\right|_{\beta_j=\hat{\beta}_j^{SL}} = -\lambda_n\operatorname{Sgn}(\hat{\beta}_j^{SL}) \qquad \text{for } j:\ \hat{\beta}_j^{SL}\neq 0, \tag{3.22}$$

$$\left|\left.\frac{\|Y-X\beta\|_n^2+\mu_n\beta'\widetilde{J}\beta}{d\beta_j}\right|_{\beta_j=\hat{\beta}_j^{SL}}\right| \leq \lambda_n \qquad \text{for } j:\ \hat{\beta}_j^{SL}=0. \tag{3.23}$$

Recall that $\mathcal{A}^*=\{j:\beta_j^*\neq 0\}$, the second result states that if we restrict ourselves to the covariates which we are after (i.e. indexes in $\mathcal{A}^*$), we get a consistent estimate as soon as the regularization parameters λ_n and μ_n are properly chosen.

Lemma 3.4. *Let $\tilde{\beta}_{\mathcal{A}^*}$ a minimizer of*

$$\|Y-X_{\mathcal{A}^*}\beta_{\mathcal{A}^*}\|_n^2+\lambda_n\sum_{j\in\mathcal{A}^*}|\beta_j|+\mu_n\beta'_{\mathcal{A}^*}\widetilde{J}_{\mathcal{A}^*,\mathcal{A}^*}\beta_{\mathcal{A}^*}.$$

If $\lambda_n\rightarrow 0$ and $\mu_n\rightarrow 0$, then $\tilde{\beta}_{\mathcal{A}^*}$ converges to $\beta^*_{\mathcal{A}^*}$ in probability.

This lemma can be see as a special and restricted case of Theorem 3.1. We now prove Theorem 3.2. Let $\tilde{\beta}_{\mathcal{A}^*}$ as in Lemma 3.4. We define an estimator $\tilde{\beta}$ by extending $\tilde{\beta}_{\mathcal{A}^*}$ by zeros on $(\mathcal{A}^*)^c$. Hence, consistency of $\tilde{\beta}$ is ensure as a simple consequence of Lemma 3.4. Now we need to prove that with probability tending to one, this estimator is optimal for the problem (3.2)-(3.3). That is the optimal conditions (3.22)-(3.23) are fulfilled with probability tending to one.

From now on, we write $\mathcal{A}$ for $\mathcal{A}^*$. By definition of $\tilde{\beta}_{\mathcal{A}}$, the optimality condition (3.22) is satisfied. We now must check the optimality condition (3.23). Combining the fact that $Y = X\beta^* + \varepsilon$ and the convergence of the matrix $X'X/n$ and the vector $\varepsilon' X/\sqrt{n}$, we have

$$n^{-1}(X'Y - X'X_{\mathcal{A}}\tilde{\beta}_{\mathcal{A}}) = \Psi_{.,\mathcal{A}}(\beta^*_{\mathcal{A}} - \tilde{\beta}_{\mathcal{A}}) + \mathcal{O}_p(n^{-1/2}). \tag{3.24}$$

Moreover, the optimality condition (3.22) for the estimator $\tilde{\beta}$ can be written as

$$n^{-1}(X'_{.,\mathcal{A}}Y - X'_{.,\mathcal{A}}X_{.,\mathcal{A}}\tilde{\beta}_{\mathcal{A}}) = \frac{\lambda_n}{2}\operatorname{Sgn}(\tilde{\beta}_{\mathcal{A}}) - \mu_n\widetilde{J}_{\mathcal{A},\mathcal{A}}(\beta^*_{\mathcal{A}} - \tilde{\beta}_{\mathcal{A}}) + \mu_n\widetilde{J}_{\mathcal{A},\mathcal{A}}\beta^*_{\mathcal{A}}. \tag{3.25}$$

Combining (3.24) and (3.25), we easily obtain

$$(\beta^*_{\mathcal{A}} - \tilde{\beta}_{\mathcal{A}}) = (\Psi_{\mathcal{A},\mathcal{A}} + \mu_n\widetilde{J}_{\mathcal{A},\mathcal{A}})^{-1}\left(\frac{\lambda_n}{2}\operatorname{Sgn}(\tilde{\beta}_{\mathcal{A}}) + \mu_n\widetilde{J}_{\mathcal{A},\mathcal{A}}\beta^*_{\mathcal{A}}\right) + \mathcal{O}_p(n^{-1/2}).$$

Since $\tilde{\beta}$ is consistent in estimation and $\lambda_n n^{1/2} \to \infty$, for each $j \in \mathcal{A}^c$, the left hand side in the optimality condition (3.23)

$$\frac{1}{\lambda_n n}(X'_j Y - X'_j X_{.,\mathcal{A}}\tilde{\beta}_{\mathcal{A}}) - \frac{\mu_n}{\lambda_n}\widetilde{J}_{j,\mathcal{A}}\tilde{\beta}_{\mathcal{A}} =: L_j^{(n)},$$

converges to

$$\Psi_{j,\mathcal{A}}(K_{\mathcal{A},\mathcal{A}})^{-1}\left(2^{-1}\operatorname{Sgn}(\beta^*_{\mathcal{A}}) + \frac{\mu}{\lambda}\widetilde{J}_{\mathcal{A},\mathcal{A}}\beta^*_{\mathcal{A}}\right) - \frac{\mu}{\lambda}\widetilde{J}_{j,\mathcal{A}}\beta^*_{\mathcal{A}} =: L_j.$$

By condition (3.6), this quantity is strictly smaller than one. Then

$$\lim_{n\to\infty}\mathbb{P}\left(\forall j \in \mathcal{A}^c,\, |L_j^{(n)}| \leq 1\right) \geq \prod_{j\in\mathcal{A}^c}\mathbb{P}\left(|L_j| \leq 1\right) = 1,$$

which ends the proof. □

Proof of Theorem 3.3. We prove the theorem by contradiction by assuming that there exists a $j \in (\mathcal{A}^*)^c$ such that there exists a $i \in \mathcal{A}^*$ and

$$|\Omega_j(\lambda, \mu, \mathcal{A}^*, \beta^*)| > 1,$$

where the Ω_j are given by (3.5). Since $\hat{\mathcal{A}} = \mathcal{A}^*$ with probability tending to one, optimality condition (3.22) implies

$$\hat{\beta}^{SL}_{\mathcal{A}} = ((K_n)_{\mathcal{A},\mathcal{A}})^{-1} \left(\frac{X'_{.,\mathcal{A}} Y}{n} - \frac{\lambda_n}{2} \operatorname{Sgn}(\hat{\beta}^{SL}_{\mathcal{A}}) \right). \tag{3.26}$$

Using this expression of $\hat{\beta}^{SL}_{\mathcal{A}}$ and $Y = X_{.,\mathcal{A}}\beta^*_{\mathcal{A}} + \varepsilon$, then for every $j \in \mathcal{A}^c$,

$$\begin{aligned}
\frac{X'_j Y}{n} - \frac{X'_j X_{.,\mathcal{A}} \hat{\beta}^{SL}_{\mathcal{A}}}{n} &= \frac{X'_j Y}{n} - \frac{X'_j X_{.,\mathcal{A}}}{n} ((K_n)_{\mathcal{A},\mathcal{A}})^{-1} \frac{X'_{.,\mathcal{A}} Y}{n} \\
&\quad + \frac{\lambda_n}{2} \frac{X'_j X_{.,\mathcal{A}}}{n} ((K_n)_{\mathcal{A},\mathcal{A}})^{-1} \operatorname{Sgn}(\hat{\beta}^{SL}_{\mathcal{A}}) \\
&= \frac{X'_j Y}{n} - \frac{X'_j X_{.,\mathcal{A}}}{n} ((K_n)_{\mathcal{A},\mathcal{A}})^{-1} \frac{X'_{.,\mathcal{A}} \varepsilon}{n} - \frac{X'_j X_{.,\mathcal{A}}}{n} \beta^*_{\mathcal{A}} \\
&\quad + \frac{X'_j X_{.,\mathcal{A}}}{n} ((K_n)_{\mathcal{A},\mathcal{A}})^{-1} \left(\frac{\lambda_n}{2} \operatorname{Sgn}(\hat{\beta}^{SL}_{\mathcal{A}}) + \mu_n \widetilde{J}_{\mathcal{A},\mathcal{A}} \beta^*_{\mathcal{A}} \right).
\end{aligned}$$

Therefore,

$$n^{-1}(X'_j Y - X'_j X_{.,\mathcal{A}} \hat{\beta}^{SL}_{\mathcal{A}}) - \mu_n \widetilde{J}_{j,\mathcal{A}} \beta^{SL}_{\mathcal{A}} = A_n + B_n,$$

with

$$\begin{cases}
A_n = \frac{X'_j Y}{n} - \frac{X'_j X_{.,\mathcal{A}}}{n} ((K_n)_{\mathcal{A},\mathcal{A}})^{-1} \frac{X'_{.,\mathcal{A}} \varepsilon}{n} - \frac{X'_j X_{.,\mathcal{A}}}{n} \beta^*_{\mathcal{A}} \\
B_n = \frac{X'_j X_{.,\mathcal{A}}}{n} ((K_n)_{\mathcal{A},\mathcal{A}})^{-1} \left(\frac{\lambda_n}{2} \operatorname{Sgn}(\hat{\beta}^{SL}_{\mathcal{A}}) + \mu_n \widetilde{J}_{\mathcal{A},\mathcal{A}} \beta^*_{\mathcal{A}} \right) - \mu_n \widetilde{J}_{j,\mathcal{A}} \hat{\beta}^{SL}_{\mathcal{A}}.
\end{cases}$$

We treat this two terms separately. First as $\hat{\beta}^{SL}_{\mathcal{A}}$ converges in probability to $\beta^*_{\mathcal{A}}$ and empirical covariance matrices convergence, the sequence B_n/λ_n converges to

$$B = \Psi_{j,\mathcal{A}} (K_{\mathcal{A},\mathcal{A}})^{-1} (2^{-1} \lambda \operatorname{Sgn}(\beta^*_{\mathcal{A}}) + \mu \lambda^{-1} \widetilde{J}_{\mathcal{A},\mathcal{A}} \beta^*_{\mathcal{A}}) - \mu \lambda^{-1} \widetilde{J}_{j,\mathcal{A}} \beta^*_{\mathcal{A}}.$$

By assumption $|B| > 1$. This implies that $\mathbb{P}\left(B_n/\lambda_n \geq (1 + |B|)/2\right)$ converges to one.

With regard to the other term, since $Y = X\beta^* + \varepsilon$ we have

$$\begin{aligned}
A_n &= \frac{X'_j \varepsilon}{n} - \frac{X'_j X_{.,\mathcal{A}}}{n} ((K_n)_{\mathcal{A},\mathcal{A}})^{-1} \frac{X'_{.,\mathcal{A}} \varepsilon}{n} \\
&= n^{-1} \sum_{k=1}^{n} \varepsilon_k (x_{k,j} - \Psi_{j,\mathcal{A}} (K_{\mathcal{A},\mathcal{A}})^{-1} x'_{k,\mathcal{A}}) + o_p(n^{-1/2}) \\
&= n^{-1} \sum_{k=1}^{n} c_n + o_p(n^{-1/2}) = C_n + o_p(n^{-1/2}),
\end{aligned}$$

where c_n are i.i.d. random variables with mean 0 and variance:

$$\begin{aligned}
s^2 = \mathbb{V}ar(c_k) &= \mathbb{E}(c_k^2) = \mathbb{E}[\mathbb{E}(c_k^2|X)] \\
&= \mathbb{E}\left[\mathbb{E}(\varepsilon_k^2|X)(x_{k,j} - \Psi_{j,\mathcal{A}} (K_{\mathcal{A},\mathcal{A}})^{-1} x'_{k,\mathcal{A}})^2 \right] \\
&= \sigma^2 \mathbb{E}\left[\Psi_{j,j} + \Psi_{j,\mathcal{A}} (K_{\mathcal{A},\mathcal{A}})^{-1} \Psi_{\mathcal{A},\mathcal{A}} (K_{\mathcal{A},\mathcal{A}})^{-1} \Psi_{\mathcal{A},j} \right. \\
&\quad \left. - 2\Psi_{j,\mathcal{A}} (K_{\mathcal{A},\mathcal{A}})^{-1} \Psi_{\mathcal{A},j} \right].
\end{aligned}$$

Thus, by the central limit theorem, $n^{1/2}C_n$ is asymptotically normal with mean 0 and covariance matrix s^2/n, which is finite. Thus $\mathbb{P}(n^{1/2}A_n > 0)$ converges to 1/2.

Finally, $\mathbb{P}((A_n + B_n)/\lambda_n > (1 + |B|)/2)$ is asymptotically bounded below by 1/2. Thus $|(A_n + B_n)/\lambda_n|$ is asymptotically bigger than 1 with a positive probability, that is to say the optimality condition (3.23) is not satisfied. Then $\hat{\beta}^{SL}$ is not optimal. We get a contradiction, which concludes the proof. □

3.11 Proofs in the high-dimensional case

Proof of Theorem 3.4. Using the definition of the penalized estimator (3.2)–(3.3), for any $\beta \in \mathbb{R}^p$, we have

$$\begin{aligned} &\|X\hat{\beta}^{SL} - X\beta^*\|_n^2 - \frac{2}{n}\sum_{i=1}^n \varepsilon_i x_i \hat{\beta}^{SL} + \lambda_n |\hat{\beta}^{SL}|_1 + \mu_n (\hat{\beta}^{SL})' \widetilde{J} \hat{\beta}^{SL} \\ \leq \;& \|X\beta - X\beta^*\|_n^2 - \frac{2}{n}\sum_{i=1}^n \varepsilon_i x_i \beta + \lambda_n |\beta|_1 + \mu_n \beta' \widetilde{J} \beta. \end{aligned}$$

Therefore, if we chose $\beta = \beta^*$, we obtain the following inequalities:

$$\begin{aligned} \|X\hat{\beta}^{SL} - X\beta^*\|_n^2 &\leq \lambda_n \sum_{j=1}^p \left(|\beta_j^*| - |\hat{\beta}_j^{SL}| \right) + \frac{2}{n}\sum_{i=1}^n \varepsilon_i x_i (\hat{\beta}^{SL} - \beta^*) \\ &\quad + \mu_n (\beta^{*\prime} \widetilde{J} \beta^* - (\hat{\beta}^{SL})' \widetilde{J} \hat{\beta}^{SL}) \\ &\leq \lambda_n \sum_{j=1}^p \left(|\beta_j^*| - |\hat{\beta}_j^{SL}| \right) + \frac{2}{n}\sum_{i=1}^n \varepsilon_i x_i (\hat{\beta}^{SL} - \beta^*) \\ &\quad + \mu_n \beta^{*\prime} \widetilde{J} \beta^*, \end{aligned} \tag{3.27}$$

as $\beta' \widetilde{J} \beta \geq 0$ for any $\beta \in \mathbb{R}^p$. In order to control (3.27), we use in a first time Assumption (A1) so that $\mu_n \beta^{*\prime} \widetilde{J} \beta^* \leq L_1 \kappa_2 \sigma^2 \dfrac{\log(p)|\mathcal{A}^*|}{n}$. Second we bound the residual term in the same way as in the paper of Bunea, Tsybakov, and Wegkamp [33]. Then, we only present here the main lines. Recall that $\mathcal{A} = \mathcal{A}^* = \{j : \beta_j^* \neq 0\}$. Then, on the event $\Lambda_{n,p} = \{\max_{j=1,\ldots,p} 4|V_j| \leq \lambda_n\}$ with $V_j = n^{-1}\sum_{i=1}^n x_{i,j}\varepsilon_i$, we have

$$\|X\hat{\beta}^{SL} - X\beta^*\|_n^2 + 2^{-1}\lambda_n \sum_{j=1}^p \left| \hat{\beta}_j^{SL} - \beta_j^* \right| \leq \lambda_n \sum_{j \in \mathcal{A}} \left| \hat{\beta}_j^{SL} - \beta_j^* \right| + L_1 \kappa_2 \sigma^2 \frac{\log(p)|\mathcal{A}|}{n}. \tag{3.28}$$

This inequality is obtained thanks to the fact that $|\beta_j^* - \hat{\beta}_j^{SL}| + |\beta_j^*| - |\hat{\beta}_j^{SL}| = 0$ for any $j \notin \mathcal{A}$ and to the triangular inequality. The rest of the proof consists in bounding this term $\lambda_n \sum_{j \in \mathcal{A}} \left| \hat{\beta}_j^{SL} - \beta_j^* \right|$. Using similar arguments as in the paper of Bunea, Tsybakov,

and Wegkamp [33], we can write

$$\begin{aligned}\sum_{j\in\mathcal{A}}(\hat{\beta}_j^{SL}-\beta_j^*)^2 &\leq \|X\hat{\beta}^{SL}-X\beta^*\|_n^2+2\rho_1\sum_{k\in\mathcal{A}}|\hat{\beta}_k^{SL}-\beta_k^*|\sum_{j=1}^{p}|\hat{\beta}_j^{SL}-\beta_j^*| \\ &\quad -\rho_1\left(\sum_{j\in\mathcal{A}}|\hat{\beta}_j^{SL}-\beta_j^*|\right)^2. \end{aligned} \tag{3.29}$$

But $\left(\sum_{j\in\mathcal{A}}\left|\hat{\beta}_j^{SL}-\beta_j^*\right|\right)^2 \leq |\mathcal{A}|\sum_{j\in\mathcal{A}}(\hat{\beta}_j^{SL}-\beta_j^*)^2$, then

$$\begin{aligned}&\left(\sum_{j\in\mathcal{A}}\left|\hat{\beta}_j^{SL}-\beta_j^*\right|\right)^2 \\ &\leq |\mathcal{A}|\quad\left\{\|X\hat{\beta}^{SL}-X\beta^*\|_n^2+2\,\rho_1\sum_{k\in\mathcal{A}}|\hat{\beta}_k^{SL}-\beta_k^*|\sum_{j=1}^{p}|\hat{\beta}_j^{SL}-\beta_j^*|\right. \\ &\qquad\left.-\,\rho_1\left(\sum_{j\in\mathcal{A}}|\hat{\beta}_j^{SL}-\beta_j^*|\right)^2\right\}. \end{aligned} \tag{3.30}$$

A simple optimization implies

$$\sum_{j\in\mathcal{A}}\left|\hat{\beta}_j^{SL}-\beta_j^*\right| \leq \frac{2\rho_1|\mathcal{A}|\sum_{j=1}^{p}|\hat{\beta}_j^{SL}-\beta_j^*|}{1+\rho_1|\mathcal{A}|}+\frac{\sqrt{|\mathcal{A}|}\,\|X\hat{\beta}^{SL}-X\beta^*\|_n^2}{1+\rho_1|\mathcal{A}|}. \tag{3.31}$$

Now, use Assumption (A2) to bound the left hand side of the inequality (3.31) and combine this to (3.28) to get

$$\|X\hat{\beta}^{SL}-X\beta^*\|_n^2+\lambda_n\sum_{j=1}^{p}\left|\hat{\beta}_j^{SL}-\beta_j^*\right| \leq 16\lambda_n^2|\mathcal{A}|+L_1\kappa_2\sigma^2\frac{\log(p)|\mathcal{A}|}{n}. \tag{3.32}$$

This proves (3.11). Finally (3.12) follows directly by dividing by λ_n both sides of this last inequality. A concentration inequality to bound $\mathbb{P}\left(\max_{j=1,\ldots,p}4|V_j|\leq\lambda_n\right)$ allows us to conclude the proof. □

Lemma 3.5. *Let $\Lambda_{n,p}$ be the random event defined by $\Lambda_{n,p}=\{\max_{j=1,\ldots,p}4|V_j|\leq\lambda_n\}$ where $V_j=n^{-1}\sum_{i=1}^{n}x_{i,j}\varepsilon_i$. Let us choose a $\kappa_1>2\sqrt{2}$ and $\lambda_n=\kappa_1\sigma\sqrt{n^{-1}\log(p)}$. Then*

$$\mathbb{P}\left(\max_{j=1,\ldots,p}4|V_j|\leq\lambda_n\right)\geq 1-p^{1-\frac{\kappa_1^2}{8}}.$$

Proof. Since $V_j\sim\mathcal{N}(0,n^{-1}\sigma^2)$, an elementary Gaussian inequality gives

$$\begin{aligned}\mathbb{P}\left(\max_{j=1,\ldots,p}\lambda_n^{-1}|V_j|\geq 4^{-1}\right) &\leq p\max_{j=1,\ldots,p}\mathbb{P}\left(\lambda_n^{-1}|V_j|\geq 4^{-1}\right) \\ &\leq p\exp\left(-\kappa_1^2\log(p)/8\right) \\ &= p^{1-\kappa_1^2/8}. \end{aligned}$$

This ends the proof. □

Proof of Theorem 3.5. Through this proof, for any $a \in \mathbb{R}^p$, let us denote by $a_{\mathcal{A}}$, the p-dimensional vector such that $(a_{\mathcal{A}})_j = a_j$ if $j \in \mathcal{A}$ and zero otherwise. Moreover, we recall that $K_n = \Psi^n + \mu_n \widetilde{J}$. Now, note that we can write the KKT conditions (3.22)-(3.23) as

$$\|K_n(\hat{\beta}^{SL} - \beta^*) - \frac{X'\varepsilon}{n} + \mu_n \widetilde{J}\beta^*\|_\infty \leq \frac{\lambda_n}{2}. \tag{3.33}$$

Recall that $\Lambda_{n,p} = \{\max_{j=1,\dots,p} 2|V_j| \leq \lambda_n\}$ with $V_j = \frac{X_j'\varepsilon}{n}$, then applying (3.33) and Assumption (A4), we have on $\Lambda_{n,p}$ and for any $j \in \{1, \dots, p\}$

$$\begin{aligned}
|(K_n)_{j,j}(\hat{\beta}_j^{SL} - \beta_j^*)| &= |\{K_n(\hat{\beta}^{SL} - \beta^*)\}_j - \sum_{\substack{k=1 \\ k\neq j}}^{p} (K_n)_{j,k}(\hat{\beta}_k^{SL} - \beta_k^*) + \mu_n(\widetilde{J}\beta^*)_j| \\
&\leq \frac{\lambda_n}{2} + |\frac{X_j'\varepsilon}{n}| + \sum_{\substack{k=1 \\ k\neq j}}^{p} |(K_n)_{j,k}(\hat{\beta}_k^{SL} - \beta_k^*) + \mu_n(\widetilde{J}\beta^*)_j| \\
&\leq \frac{3\lambda_n}{4} + \frac{1}{3\alpha|\mathcal{A}|}|\hat{\beta}^{SL} - \beta^*|_1 + \mu_n|(\widetilde{J}\beta^*)_j|.
\end{aligned}$$

Then

$$\|K_n(\hat{\beta}^{SL} - \beta^*)\|_\infty \leq \frac{3\lambda_n}{4} + \frac{1}{3\alpha|\mathcal{A}|}|\hat{\beta}^{SL} - \beta^*|_1 + \mu_n\|\widetilde{J}\beta^*\|_\infty. \tag{3.34}$$

Let us now bound $|\hat{\beta}^{SL} - \beta^*|_1$. Thanks to (3.27), we can write

$$\begin{aligned}
\lambda_n|\hat{\beta}^{SL}|_1 &\leq \lambda_n|\beta^*|_1 + \frac{2}{n}\sum_{i=1}^{p} \varepsilon_i x_i(\hat{\beta}^{SL} - \beta^*) + \mu_n \beta^{*\prime}\widetilde{J}\beta^* \\
\overset{\text{on } \Lambda_{n,p}}{\Longleftrightarrow} \lambda_n|\hat{\beta}^{SL}|_1 &\leq \lambda_n|\beta^*|_1 + \frac{\lambda_n}{2}|\hat{\beta}^{SL} - \beta^*|_1 + \mu_n \beta^{*\prime}\widetilde{J}\beta^*.
\end{aligned}$$

Dividing by λ_n, and adding $2^{-1}|\hat{\beta}^{SL} - \beta^*|_1 - |\hat{\beta}^{SL}|_1$, we get on the event $\Lambda_{n,p}$

$$\begin{aligned}
2^{-1}|\hat{\beta}^{SL} - \beta^*|_1 &\leq (|\hat{\beta}^{SL} - \beta^*|_1 + |\beta^*|_1 - |\hat{\beta}^{SL}|_1) + \frac{\mu_n}{\lambda_n}\beta^{*\prime}\widetilde{J}\beta^* \\
\iff \quad |\hat{\beta}^{SL} - \beta^*|_1 &\leq 2|\hat{\beta}_{\mathcal{A}}^{SL} - \beta_{\mathcal{A}}^*|_1 + 2\frac{\mu_n}{\lambda_n}\beta_{\mathcal{A}}^{*\prime}\widetilde{J}\beta_{\mathcal{A}^*}^* \qquad (3.35) \\
\iff \quad |\hat{\beta}^{SL} - \beta^*|_1 &\leq 2\sqrt{|\mathcal{A}|}\,\|\hat{\beta}_{\mathcal{A}}^{SL} - \beta_{\mathcal{A}}^*\|_2 + 2\frac{\mu_n}{\lambda_n}\beta_{\mathcal{A}}^{*\prime}\widetilde{J}\beta_{\mathcal{A}^*}^*, \qquad (3.36)
\end{aligned}$$

where we used the Cauchy Schwarz inequality in the last line. Combine (3.34) and (3.36), we easily get

$$\begin{aligned}
\|\hat{\beta}^{SL} - \beta^*\|_\infty \leq \frac{1}{1+\mu_n} \quad &\Big(\tfrac{3\lambda_n}{4} + \tfrac{2}{3\alpha|\mathcal{A}|}\sqrt{|\mathcal{A}|}\,\|\hat{\beta}_{\mathcal{A}}^{SL} - \beta_{\mathcal{A}}^*\|_2 \\
&+\mu_n\|\widetilde{J}\beta^*\|_\infty + \tfrac{2\mu_n}{3\alpha\lambda_n|\mathcal{A}|}\beta_{\mathcal{A}}^{*\prime}\widetilde{J}\beta_{\mathcal{A}}^*\Big). \qquad (3.37)
\end{aligned}$$

The final step consists in bounding $\|\hat{\beta}_{\mathcal{A}}^{SL} - \beta_{\mathcal{A}}^*\|_2$. First, using the KKT condition (3.33), we remark that $\|K_n(\hat{\beta}^{SL} - \beta^*)\|_\infty \leq 3\lambda_n/4 + \mu_n\|\widetilde{J}\beta^*\|_\infty$ on $\Lambda_{n,p}$. This and equation (3.36)

lied to

$$(\hat{\beta}^{SL}-\beta^*)'K_n(\hat{\beta}^{SL}-\beta^*) \leq \|K_n(\hat{\beta}^{SL}-\beta^*)\|_\infty |\hat{\beta}^{SL}-\beta^*|_1 \\ \leq (\frac{3\lambda_n}{4}+\mu_n\|\widetilde{J}\beta^*\|_\infty)(2\sqrt{|\mathcal{A}|}\|\hat{\beta}^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}\|_2+2\frac{\mu_n}{\lambda_n}\beta^{*\prime}_{\mathcal{A}}\widetilde{J}\beta^*_{\mathcal{A}}). \tag{3.38}$$

On the other hand, using Assumption (A4), and similar arguments as in the paper of Lounici [101],

$$\begin{aligned}\frac{(\hat{\beta}^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}})'K_n(\hat{\beta}^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}})}{\|\hat{\beta}^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}\|_2^2} &= \frac{(\hat{\beta}^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}})'\,\mathrm{diag}(K_n)(\hat{\beta}^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}})}{\|\hat{\beta}^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}\|_2^2} \\ &\quad+\frac{(\hat{\beta}^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}})'(K_n-\mathrm{diag}(K_n))(\hat{\beta}^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}})}{\|\hat{\beta}^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}\|_2^2} \\ &\geq 1-\frac{1}{3\alpha|\mathcal{A}|}\sum_{j,k=1}^{p}\frac{|(\hat{\beta}^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}})_j|\,|(\hat{\beta}^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}})_k|}{\|\hat{\beta}_{\mathcal{A}}-\beta^*_{\mathcal{A}}\|_2^2} \\ &\geq 1-\frac{1}{3\alpha|\mathcal{A}|}\frac{\|\hat{\beta}^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}\|_1^2}{\|\hat{\beta}^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}\|_2^2},\end{aligned}$$

where we used in the second inequality the fact that $\mathrm{diag}(K_n)$ has larger diagonal elements than 1 since the diagonal elements in Ψ^n and $\widetilde{J}$ are respectively equal to 1 and larger than 0. Now, twice using Assumption (A4), one deduces

$$\begin{aligned}\frac{(\hat{\beta}^{SL}-\beta^*)'K_n(\hat{\beta}^{SL}-\beta^*)}{\|\hat{\beta}^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}\|_2^2} &\geq \frac{(\hat{\beta}^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}})'K_n(\hat{\beta}^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}})}{\|\hat{\beta}^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}\|_2^2}+\frac{(\hat{\beta}^{SL}_{\mathcal{A}^c}-\beta^*_{\mathcal{A}^c})'K_n(\hat{\beta}^{SL}_{\mathcal{A}^c}-\beta^*_{\mathcal{A}^c})}{\|\hat{\beta}^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}\|_2^2} \\ &\geq 1-\frac{1}{3\alpha|\mathcal{A}|}\frac{|\hat{\beta}^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}|_1^2}{\|\hat{\beta}^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}\|_2^2}-\frac{|\hat{\beta}^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}|_1|\hat{\beta}^{SL}_{\mathcal{A}^c}-\beta^*_{\mathcal{A}^c}|_1}{3\alpha|\mathcal{A}|\,\|\hat{\beta}^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}\|_2^2} \\ &\geq 1-\frac{|\hat{\beta}^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}|_1^2}{\alpha|\mathcal{A}|\,\|\hat{\beta}^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}\|_2^2}-\frac{2\mu_n\,\beta^{*\prime}_{\mathcal{A}}\widetilde{J}\beta^*_{\mathcal{A}}}{3\alpha\lambda_n|\mathcal{A}|}\frac{|\hat{\beta}^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}|_1}{\|\hat{\beta}^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}\|_2^2} \\ &\geq (1-\frac{1}{\alpha})-\frac{2\mu_n\,\beta^{*\prime}_{\mathcal{A}}\widetilde{J}\beta^*_{\mathcal{A}}}{3\alpha\lambda_n|\mathcal{A}|}\frac{|\hat{\beta}^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}|_1}{\|\hat{\beta}^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}\|_2^2}.\end{aligned}$$

where we used the fact that (3.35) implies $|\hat{\beta}^{SL}_{\mathcal{A}^c}-\beta^*_{\mathcal{A}^c}|_1 \leq 2|\hat{\beta}^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}|_1+2\frac{\mu_n}{\lambda_n}\beta^{*\prime}_{\mathcal{A}}\widetilde{J}\beta^*_{\mathcal{A}}$ in the third line. The last inequalities can be summed-up by

$$(\hat{\beta}^{SL}-\beta^*)'K_n(\hat{\beta}^{SL}-\beta^*) \geq (1-\frac{1}{\alpha})\|\hat{\beta}^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}\|_2^2-\frac{2\mu_n\,\beta^{*\prime}_{\mathcal{A}}\widetilde{J}\beta^*_{\mathcal{A}}}{3\alpha\lambda_n|\mathcal{A}|}|\hat{\beta}^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}|_1. \tag{3.39}$$

Let us consider (3.38) and (3.39). An optimization work over $\|\hat{\beta}^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}\|_2$ provides us the following bound:

$$\begin{aligned}\|\hat{\beta}^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}\|_2 \leq\ & (\tfrac{\alpha}{\alpha-1})\left[(\tfrac{3\lambda_n}{2}+2\mu_n\|\widetilde{J}\beta^*\|_\infty)\sqrt{|\mathcal{A}|}+\tfrac{2\mu_n}{3\alpha\lambda_n\sqrt{|\mathcal{A}|}}\beta^{*\prime}_{\mathcal{A}}\widetilde{J}\beta^*_{\mathcal{A}}\right] \\ &+\sqrt{\tfrac{\alpha}{\alpha-1}\left(\tfrac{3\lambda_n}{2}+2\mu_n\|\widetilde{J}\beta^*\|_\infty\right)\tfrac{\mu_n}{\lambda_n}\beta^{*\prime}_{\mathcal{A}}\widetilde{J}\beta^*_{\mathcal{A}}}.\end{aligned} \tag{3.40}$$

Thanks to Assumption (A1), $\beta^{*'}_{\mathcal{A}}\widetilde{J}\beta^{*}_{\mathcal{A}} \leq L_1 \log(p)|\mathcal{A}|$ and $\|\widetilde{J}\beta^*\|_\infty \leq L_2 \log(p)$. Moreover the tuning parameters λ_n and μ_n are chosen in the form $\lambda_n = \kappa_1\sigma\sqrt{\log(p)/n}$ and $\mu_n = \kappa_3\sigma/n$. Then we conclude from (3.37) and (3.40)

$$\|\hat{\beta}^{SL} - \beta^*\|_\infty \leq \frac{1}{1+\frac{\kappa_3\sigma}{n}} \quad \left(\frac{3}{4} + \frac{1}{\alpha-1} + \frac{4L_1\kappa_3}{9\alpha^2\kappa_1^2} + \frac{2L_1\kappa_3}{3\alpha\kappa_1^2} + \sqrt{\frac{2L_1\kappa_3}{3\alpha(\alpha-1)\kappa_1^2} + \frac{8L_1\,L_2\kappa_3^2}{9\alpha(\alpha-1)\kappa_1^4}\lambda_n} \right.$$
$$\left. +(\frac{4L_2\kappa_3}{3\kappa_1^2} + \frac{L_2\kappa_3}{\kappa_1^2})\lambda_n\right)\lambda_n.$$

This ends the proof. □

Proof of Theorem 3.7. The proof of this theorem is essentially an adaptation of the one concerning the Lasso by Zou, Hastie, and Tibshirani [169]. We do not give the whole proof but only mention the important steps and let the reader refer to Zou, Hastie, and Tibshirani [169] for more details. The main points in the proof are Stein's lemma and these few facts:

- For every couple (λ, μ), the S-Lasso estimator is a continuous function of Y.
- For every couple $(\lambda, \mu) = \zeta$, the active set $\mathcal{A}_\zeta$ and the sign vector of $\hat{\beta}^{SL}_\zeta$ which we denote by Sgn_ζ are piecewise constant with respect to Y, out of a set with Lebesgue measure equal to 0.

The detailed proof uses these points and the explicit form of the estimator $\hat{\beta}^{SL}$ given by (3.26). This proof is the same as the one by Zou, Hastie, and Tibshirani [169] so that we omit it here. □

Chapter 4

A Sparsity Inequality for the Grouped Variables Lasso

Abstract: We consider the linear regression model with Gaussian error. We estimate the unknown parameters by a procedure inspired from the Group Lasso estimator introduced by Yuan and Lin [159]. We show that this estimator satisfies a sparsity inequality, i.e., a bound in terms of the number of non-zero components of the oracle regression vector. We prove that this bound is better, in some cases, than the one achieved by the Lasso and the Dantzig selector.

4.1 Introduction

We consider the linear regression model

$$y_i = x_i\beta^* + \varepsilon_i, \qquad i = 1, \dots, n, \tag{4.1}$$

where the design $x_i = (x_{i,1}, \dots, x_{i,p}) \in \mathbb{R}^p$ is deterministic, $\beta^* = (\beta_1^*, \dots, \beta_p^*)' \in \mathbb{R}^p$ is the unknown parameter vector of interest and $\varepsilon_1, \dots, \varepsilon_n$, are i.i.d. centered Gaussian random variables with variance σ^2. We wish to estimate β^* in the sparse case, i.e., when many of its components are equal to zero. If we define the covariates $\xi_j = (x_{1,j}, \dots, x_{n,j})'$, $j = 1, \dots, p$, the sparsity of the model means that only a small subset of $(\xi_j)_j$ is relevant for explaining the response y_i, $i = 1, \dots, n$. We are mainly interested in the case where the number of the covariates p is much larger than the sample size n. In such a situation, the classical methods of estimation such as ordinary least squares are inconsistent. In the last decade, a wide variety of procedures has been developed for estimation and variable selection under sparsity assumption. Most popular procedures are of the form:

$$\tilde{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \|Y - X\beta\|_n^2 + \operatorname{pen}(\beta) \right\}, \tag{4.2}$$

where $X = (x_1', \dots, x_n')'$, $Y = (y_1, \dots, y_n)'$, $\operatorname{pen} : \mathbb{R}^p \to \mathbb{R}$ is a penalty function, and, for any vector $a = (a_1, \dots, a_n)'$, $\|a\|_n^2 = n^{-1} \sum_{i=1}^n a_i^2$ (we denote by $\langle \cdot, \cdot \rangle_n$ the corresponding inner product in $\mathbb{R}^n$). When X is standardized, the Lasso procedure introduced by Tibshirani [135] is defined by (4.2) with $\operatorname{pen}(\beta) = 2\lambda_{n,p} \sum_{i=1}^n |\beta_i|$, where $\lambda_{n,p}$ is a tuning parameter. This estimator is attractive as it performs both regression parameters estimation and variable selection. In the literature, the theoretical and empirical properties of the Lasso procedure have been extensively studied. See, for instance, Efron, Hastie, Johnstone, and Tibshirani [61], Meinshausen and Bühlmann [112], Fan and Li [63], Knight and Fu [90], Zou [166] and Zhao and Yu [164], among others. Recent extensions of the Lasso and their performances can be found in the papers of Fan and Li [63], Meinshausen [111], Zou [166], Zou and Hastie [167] and Tibshirani, Saunders, Rosset, Zhu, and Knight [136].

In this paper, we study a "grouped" version of the Lasso procedure. It is defined with a penalty of the form $\operatorname{pen}(\beta) = 2\lambda_{n,p} \sum_{l=1}^L \sqrt{\sum_{j \in G_l} \|\xi_j\|_n^2 {\beta_j}^2}$, where $L \in \{1, \dots, p\}$ is the number of groups, $(G_l)_{l \in \{1, \dots, L\}}$ is a sequence of groups forming a partition of $\{1, \dots, p\}$ and $\lambda_{n,p}$ is a tuning parameter. It can be viewed as a slight modification of the Group Lasso procedure developed by Yuan and Lin [159]. For the sake of clarity, we call our modified Group Lasso: Grouped Variables Lasso. We measure its performance by considering a statistical approach derived from confidence balls. We aim to find the smallest bound $\varphi_{n,p}$ such that

$$\mathbb{P}\left(\|X\widehat{\beta} - X\beta^*\|_n^2 \le C\, \varphi_{n,p} \right) \ge 1 - u_{n,p}, \tag{4.3}$$

where $\widehat{\beta}$ is the Grouped Variables Lasso estimator, $u_{n,p}$ is a positive sequence of the form $n^{-\alpha}p^{-\gamma}$ with $\alpha > 0$, $\gamma > 0$ and C is a positive constant which does not depend on n and p. Under some assumptions on $(G_l)_{l\in\{1,\ldots,L\}}$ and X, we obtain a rate $\varphi_{n,p}$ depending only on n, on p and on an index of sparsity of the model. From this point of view, the inequality (4.3) is a *Sparsity Inequality* (SI) for the Grouped Variable Lasso estimator. Such SIs have already been investigated for other estimators (Bunea, Tsybakov, and Wegkamp [33], Dalalyan and Tsybakov [46], Koltchinskii [93], van de Geer [143] and Candès and Tao [35]). As a benchmark, we use the SIs provided for the Lasso estimator introduced by Bunea, Tsybakov, and Wegkamp [30] and for the Dantzig selector introduced by Bickel, Ritov, and Tsybakov [17]. If we compare the corresponding $\varphi_{n,p}$, we remark that the one achieved by the Grouped Variables Lasso is smaller than the one achieved by the Lasso and the Dantzig selector. This illustrates the fact that, in some situations, the Grouped Variables Lasso exploits the sparsity of the model more efficiently than the Lasso and the Dantzig selector.

The rest of the paper is organized as follows. The Grouped Variables Lasso estimator is described in Section 2. Section 3 presents the assumptions made on the model. The theoretical performance of the considered estimator is investigated in Section 4. The proofs are postponed to Section 5.

4.2 The Grouped Variables Lasso (GVL) estimator

In this study, for any real number a, $[a]$ denotes the integer part of a. We define the Grouped Variables Lasso (GVL) estimator by

$$\widehat{\beta} = \underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \|Y - X\beta\|_n^2 + 2\lambda_{n,p} \sum_{l=1}^{L} \sqrt{\sum_{j\in G_l} \|\xi_j\|_n^2 {\beta_j}^2} \right\}, \tag{4.4}$$

where

$$\lambda_{n,p} = \kappa\sigma\sqrt{n^{-1}\log(np)}, \tag{4.5}$$

$\kappa > 2$, $L \in \{1,\ldots,p\}$ and $(G_l)_{l\in\{1,\ldots,L\}}$ is a partition of $\{1,\ldots,p\}$. We suppose that L and $(G_l)_{l\in\{1,\ldots,L\}}$ satisfy

$$\max_{l\in\{1,\ldots,L\}} \operatorname{Card}(G_l) \leq \min\left([\log(np)], p\right).$$

Additional assumptions on X, L and $(G_l)_{l\in\{1,\ldots,L\}}$ will be specified in Section 4.3.

Naturally, we obtain the Lasso estimator when $\widehat{\beta}$ is defined with, for any $l \in \{1,\ldots,p\}$, $\operatorname{Card}(G_l) = 1$. Moreover, the GVL estimator is a slight modification of the Group Lasso estimator developed by Yuan and Lin [159]. The main difference is the maximal size of groups $\min(p, [\log(np)])$ which, as we will see in Section 4.4, offers significant advantages from the theoretical point of view. We also do not assume that groups are such that, for

any $l \in \{1, \ldots, L\}$, $X'_{G_l} X_{G_l} = I_{\mathrm{Card}(G_l)}$, where X_{G_l} is the restriction of X on the block G_l and $I_{\mathrm{Card}(G_l)}$ is the $\mathrm{Card}(G_l) \times \mathrm{Card}(G_l)$ identity matrix. In practice, let us just mention that we can use the Group Lasso modification of the LARS algorithm (see Efron, Hastie, Johnstone, and Tibshirani [61] and Yuan and Lin [159]) to compute an approximation of the solution of the Grouped Variable Lasso estimator. This can be done by adding the above mentioned restriction on the groups sizes. We also refer to Kim, Kim, and Kim [89] and Meier, van de Geer, and Bühlmann [109] for recent developments concerning the Group Lasso method.

For any real number a, we set $(a)_+ = \max(a, 0)$. If $X'X = I_n$, $L = n/[\log n]$ is supposed to be an integer and, for any $l \in \{1, \ldots, L\}$, $G_l = \big\{k \in \{1, \ldots, n\} : (l-1)[\log n] + 1 \leq k \leq l[\log n]\big\}$, each component of the GVL estimator $\widehat{\beta}$ in the group G_l can be expressed in the following form $\widehat{\beta}_i = \left(1 - \left(\kappa\sigma\sqrt{2n^{-1}\log n}\right) / \sqrt{\sum_{j \in G_l} y_j{}^2}\right)_+ y_i$. In this case, $\widehat{\beta}$ can be viewed as a slight modification of the blockwise Stein estimator. This construction enjoys powerful theoretical properties in various statistical approaches (oracle inequalities, (near) minimax optimality,...). See, for instance, Cai [34] and Cavalier and Tsybakov [40].

4.3 Assumptions

Recall that $X = (x_{i,j})_{i,j}$ is the $n \times p$ design matrix and, for any $j \in \{1, \ldots, p\}$, $\xi_j = (x_{1,j}, \ldots, x_{n,j})'$. Let $\rho_p = (\rho_p(j,k))_{j,k}$ be the correlation matrix defined by

$$\rho_p(j,k) = \frac{\langle \xi_j, \xi_k \rangle_n}{\|\xi_j\|_n \|\xi_k\|_n}, \qquad (j,k) \in \{1, \ldots, p\}^2.$$

We now present three assumptions we need to establish a SI for the GVL estimator. They are related to the correlation matrix ρ_p:

– *Assumption (A1). For any $l \in \{1, \ldots, L\}$, consider the set $\mathcal{S}_2^l = \{a = (a_j)_{j \in G_l} \in \mathbb{R}^{\mathrm{Card}(G_l)};\ \sum_{j \in G_l} a_j^2 \leq 1\}$. There exists a constant $C_* \geq 1$ independent of n and of p such that*

$$\max_{l=1,\ldots,L} \sup_{a \in \mathcal{S}_2^l} \left(\sum_{j \in G_l} \sum_{k \in G_l} a_j a_k \rho_p(j,k) \right) \leq C_*.$$

The second assumption must be satisfied for a subset $\mathcal{B} \subseteq \{1, \ldots, L\}$ to be specified later.

– *Assumption (A2)($\mathcal{B}$). The correlation matrix ρ_p satisfies*

$$\max_{l \in \mathcal{B}} \max_{m=1,\ldots,L} \sqrt{\sum_{j \in G_l} \sum_{\substack{k \in G_m \\ k \neq j}} \rho_p^2(j,k)} \leq (32)^{-1} \mathrm{Card}(\mathcal{B})^{-1}. \tag{4.6}$$

Remark 4.1. *The condition in Assumption (A1) is equivalent to requiring that the largest eigenvalues of the diagonal blocks of the matrix ρ_p (i.e., eigenvalues of the correlation matrices restricted to covariates in the same group) are bounded by C_*.*

Lemma 4.1 below provides an example of a standard family of matrices satisfying Assumption (A1).

Lemma 4.1. *Let $X = (x_{i,j})_{i,j}$ be a $n \times p$ matrix and, for any $j \in \{1, \ldots, p\}$, $\xi_j = (x_{1,j}, \ldots, x_{n,j})'$. Suppose that, for any $(j,k) \in \{1, .., p\}^2$, we have*

$$\langle \xi_j, \xi_k \rangle_n = r_n z_j z_k b_{|j-k|},$$

where $r = (r_n)_n$ is a sequence of real numbers, $z = (z_u)_u$ denotes a positive sequence and $b = (b_u)_u$ denotes a sequence in $l_1(\mathbb{N})$ with $b_0 > 0$. Then X satisfies Assumption (A1) with $C_* = 1 + 2b_0^{-1}\|b\|_{l_1}$, where $\|b\|_{l_1} = \sum_{j=1}^{p} |b_j|$.

Here are some comments on Assumption (A2)($\mathcal{B}$). In our study, Assumption (A2)($\mathcal{B}$) only needs to be satisfied for a particular set $\mathcal{B} = \Theta_G \subseteq \{1, \ldots, L\}$ (to be defined in Subsection 4.4.1). This set characterizes the sparsity of the model. Note also that Assumption (A2)($\mathcal{B}$) can be viewed as an extension of the mutual coherence condition considered by Bunea, Tsybakov, and Wegkamp [32] ,i.e., $\sup_{j \in \mathcal{B}} \sup_{k \neq j} |\rho_p(j,k)| \leq (45)^{-1} \operatorname{Card}(\mathcal{B})^{-1}$, where $\mathcal{B} = \Theta_L = \{j \in \{1, \ldots, p\};\ \beta_j \neq 0\}$. This mutual coherence condition has been introduced by Donoho, Elad, and Temlyakov [51]. When we treat the case $p \geq n$, such coherence condition is standard as almost all SIs provided in the literature need a similar condition.

Remark 4.2. *For any two sets $\mathcal{B}_1$ and $\mathcal{B}_2$ such that $\mathcal{B}_1 \subseteq \mathcal{B}_2 \subseteq \{1, \ldots, L\}$, Assumption (A2)($\mathcal{B}_2$) implies Assumption (A2)($\mathcal{B}_1$).*

Remark 4.3. *If $\mathcal{B} = \{1, \ldots, L\}$ then Assumption (A2)($\mathcal{B}$) implies Assumption (A1) with $C_* = 1 + (32)^{-1}$. This is a consequence of the Hölder inequality.*

Remark 4.4. *By constructing groups with large sizes, one can increase the value of the right hand side expression in* (4.6)*. However, such a choice for groups tends to increase the value of the left hand side expression in* (4.6) *as the double sum will be calculated over a larger number of terms. In practice, one can take into account these two facts to construct adapted groups. Nevertheless, the choice of the groups should be essentially dicted by the nature of the applications.*

An example of a $n \times p$ matrix $X = (x_{i,j})_{i,j}$ satisfying Assumptions (A1) and (A2)($\mathcal{B}$) for any $\mathcal{B} \subseteq \{1, \ldots, L\}$, is the one characterized by the equality $\langle \xi_j, \xi_k \rangle_n = n^{\nu} p^{-\alpha|j-k|}$, with

$\nu \in \mathbb{R}$, $\alpha \geq 3$ and $p \geq 32$. Here is a concise proof: Thanks to Lemma 4.1, Assumption (A1) is satisfied for any constant $C_* \geq 1 + 2p/(p-1)$ (for instance, $C_* = 4$). Moreover $\max_{l=1,\ldots,L} \max_{m=1,\ldots,L} \sqrt{\sum_{j \in G_l} \sum_{\substack{k \in G_m \\ k \neq j}} p^{-2\alpha|j-k|}} \leq p^{-\alpha} \max_{l=1,\ldots,L} \operatorname{Card}(G_l) \leq p^{-\alpha+1} \leq (32)^{-1} L^{-1} \leq (32)^{-1} \operatorname{Card}(\mathcal{B})^{-1}$ and Assumption (A2)($\mathcal{B}$) is satisfied.

When $p \leq n$, Assumption (A2)($\mathcal{B}$) can replaced by the following:

– *Assumption (A3). Consider the $p \times p$ Gram matrix Ψ_n defined by $\Psi_n = \left(\langle \xi_j, \xi_k \rangle_n\right)_{j,k}$. For any $p \geq 2$, there exists a constant $c_p > 0$ such that the matrix $\Psi_n - c_p \operatorname{diag}(\Psi_n)$ is positive semi-definite.*

Assumption (A3) is the same as by Bunea, Tsybakov, and Wegkamp [32, Assumption (A3)]. Further details can be found by Bunea, Tsybakov, and Wegkamp [32, Remarks 4-5]. Assumption (A3) is, for instance, always fulfilled for positive matrices Ψ_n. It is important to notice that this assumption can be helpful when the "group mutual coherence" assumption is not satisfied; Assumptions (A2)($\mathcal{B}$) and (A3) can recover different types of design matrices.

4.4 Theoretical properties

In this section, we investigate some theoretical properties of the GVL estimator. Notice that all the results include the case $p \geq n$.

4.4.1 Main results

Here we provide SIs achieved by the GVL estimator. These SIs take advantage of the group structure of the estimator. The key is the introduction of a *group sparsity set* Θ_G defined by:

$$\Theta_G = \left\{ l \in \{1, \ldots, L\} : \text{ there exists an integer } j_0 \in G_l \text{ such that } \beta^*_{j_0} \neq 0 \right\}, \tag{4.7}$$

where G_l is defined in Section 2. Such a set contains group indexes and characterizes the sparsity of the model. Indeed, the "sparser" the model is, the smaller the sparsity index $\operatorname{Card}(\Theta_G)$ is. Proposition 4.1 below provides an upper bound for the squared error of the GVL estimator. This bound brings into play the sparsity index inferred by the group sparsity set Θ_G.

Proposition 4.1. *We consider the linear regression model (4.1). Let $\Lambda_{n,p}$ be the random event defined by*

$$\Lambda_{n,p} = \left\{ \max_{l=1,\ldots,L} \sqrt{\sum_{j \in G_l} \|\xi_j\|_n^{-2} V_j^2} \leq 2^{-1} \lambda_{n,p} \right\}, \tag{4.8}$$

where $V_j = n^{-1}\sum_{i=1}^{n} x_{i,j}\varepsilon_i$ *and* $\lambda_{n,p}$ *is defined by (4.5). Let* $\widehat{\beta}$ *be the GVL estimator defined by (4.4) and* Θ_G *be the group sparsity set defined by (4.7). Suppose that* X *satisfies Assumption* $(A2)(\Theta_G)$*. Then, on* $\Lambda_{n,p}$*, we have*

$$\|X\widehat{\beta} - X\beta^*\|_n^2 \leq C\, n^{-1}\log(np)\,\mathrm{Card}(\Theta_G), \tag{4.9}$$

where $C = 16\kappa^2\sigma^2$*.*

The proof of Proposition 4.1 is based on the 'argmin' definition of the estimator $\widehat{\beta}$ and some technical inequalities.

Remark 4.5. *We can improve the convergence rate in* (4.9) *by making the sparsity index* $\mathrm{Card}(\Theta_G)$ *as small as possible, in other words, by choosing groups with as large as possible sizes. The best choice consists then in considering each group with the size* $\min(p, [\log(np)])$*. This means that we construct* L *groups with* $L = p/\min(p, [\log(np)])$*, supposed to be an integer.*

Theorem 4.1 below states that, under some assumptions on X, the SI of equation (4.9) is true with high probability.

Theorem 4.1. *We consider the linear regression model (4.1). Let* $\widehat{\beta}$ *be the GVL estimator defined by (4.4) and* Θ_G *be the group sparsity set defined by (4.7). Suppose that* X *satisfies Assumptions (A1) and* $(A2)(\Theta_G)$*. Then we have*

$$\mathbb{P}\left(\|X\widehat{\beta} - X\beta^*\|_n^2 \leq C\, n^{-1}\log(np)\,\mathrm{Card}(\Theta_G)\right) \geq 1 - u_{n,p}, \tag{4.10}$$

where $C = 16\kappa^2\sigma^2$ *and* $u_{n,p} = L(np)^{-(2^{-1}\kappa-1)^2/(2C_*)}$*, with* C_* *is the constant appearing in Assumption (A1).*

The proof of Theorem 4.1 uses Proposition 4.1 and a concentration inequality of the form $\mathbb{P}(\Lambda_{n,p}^c) \leq u_{n,p}$, where $\Lambda_{n,p}^c$ denotes the complementary of the set (4.8).

Corollary 4.1 below states that, when $p \leq n$, Theorem 4.1 holds with Assumption (A3) instead of Assumption $(A2)(\mathcal{B})$.

Corollary 4.1. *We consider the linear regression model (4.1). Let* Θ_G *be the group sparsity set defined by (4.7). Suppose that* X *satisfies Assumptions (A1) and (A3). Then the GVL estimator (4.4) satisfies the inequality (4.10) with* $C = 16{c_p}^{-1}\kappa^2\sigma^2$*, where* c_p *is the constant appearing in Assumption (A3).*

The proof of Corollary 4.1 is similar to the proof of Proposition 4.1.

4.4.2 Comparison with the Lasso and the Dantzig selector

A result similar to Theorem 4.1 has been proved for the Lasso estimator by Bunea, Tsybakov, and Wegkamp [30], and for the Dantzig selector by Candès and Tao [35]. Moreover Bickel, Ritov, and Tsybakov [17] stated that the squared error of the Lasso and the Dantzig selector are equivalent up to a constant factor. In these works, the authors provided similar SIs. The main difference lies in the sparsity index $\mathrm{Card}(\Theta_G)$. For both the Lasso estimator and the Dantzig selector, it is replaced by $\mathrm{Card}(\Theta^*)$, where $\Theta^* = \{j \in \{1,\dots,p\};\, \beta_j^* \neq 0\}$. Since

$$\mathrm{Card}(\Theta_G) \leq \mathrm{Card}(\Theta^*),$$

Theorem 4.1 states that, with high probability, the GVL estimator can have a smaller squared error than the Lasso estimator. This illustrates the fact that, in some cases, the GVL estimator exploits better the sparsity of the model than the Lasso estimator and the Dantzig selector. Moreover, $\mathrm{Card}(\Theta_G)$ can be asymptotically significantly smaller than $\mathrm{Card}(\Theta^*)$. For example, if $p = n$, $L = n/[\log n]$ is supposed to be an integer, for any $l \in \{1,\dots,L\}$, $G_l = \left\{k \in \{1,\dots,n\} : (l-1)[\log n] + 1 \leq k \leq l[\log n]\right\}$ and the unknown parameter vector $\beta^* = (\beta_1^*,\dots,\beta_n^*)'$ is defined by $\beta^* = (\underbrace{1,\dots,1}_{\log n}, \underbrace{0,\dots,0}_{n-\log n})$, then $\mathrm{Card}(\Theta_G) = 1$ whereas $\mathrm{Card}(\Theta^*) = \log n$.

4.5 Proofs

Proof of Lemma 4.1. For the sake of simplicity and loss of generality, we work on the set G_1 of the form $G_1 = \{1,\dots,m\}$, $m \in \{1,\dots,\min([\log(np)], p)\}$. Let us notice that, for any $u \in G_1$, we have $\|\xi_u\|_n = z_u\sqrt{r_n b_0}$. Therefore, for any $(j,k) \in \{1,\dots,p\}^2$, we have $\rho_p(j,k) = b_0^{-1} b_{|j-k|}$. Hence

$$\begin{aligned}
&\sum_{j\in G_1}\sum_{k\in G_1} a_j a_k \rho_p(j,k) \\
&\quad= \quad b_0^{-1}\sum_{j=1}^{m}\sum_{k=1}^{m} a_j a_k b_{|j-k|} = \sum_{j=1}^{m} a_j^2 + 2b_0^{-1}\sum_{j=2}^{m}\sum_{k=1}^{j-1} a_j a_k b_{j-k} \\
&\quad\leq \quad \sum_{j=1}^{m} a_j^2 + b_0^{-1}\sum_{j=2}^{m}\sum_{u=1}^{j-1}(a_j^2 + a_{j-u}^2) b_u.
\end{aligned}$$

For any $a \in \mathcal{S}_2^l$, we have $\sum_{j=1}^m a_j^2 \leq 1$. Therefore

$$\sum_{j=2}^{m}\sum_{u=1}^{j-1} a_j^2 b_u = \sum_{j=2}^{m} a_j^2 \sum_{u=1}^{j-1} b_u \leq \|b\|_{l_1}$$

and

$$\sum_{j=2}^{m}\sum_{u=1}^{j-1} a_{j-u}^2 b_u = \sum_{u=1}^{m-1} b_u \sum_{j=u+1}^{m} a_{j-u}^2 \leq \|b\|_{l_1}.$$

Hence

$$\sup_{a\in\mathcal{S}_2^l}\left(\sum_{j\in G_1}\sum_{k\in G_1}a_j a_k\rho_p(j,k)\right)\leq(1+2b_0^{-1}\|b\|_{l_1})=C_*.$$

This inequality can easily be extended to any set G_l. Thus, the matrix X satisfies Assumption (A1) with $C_*=1+2b_0^{-1}\|b\|_{l_1}$. □

Proof of Proposition 4.1. For the sake of simplicity, for any $j\in\{1,\ldots,p\}$, we set $w_{n,j}=\lambda_{n,p}\|\xi_j\|_n$. By definition of the penalized estimator (4.4), for any $\beta\in\mathbb{R}^p$, we have

$$\begin{aligned}&\|X\widehat{\beta}-X\beta^*\|_n^2+2\sum_{l=1}^{L}\sqrt{\sum_{j\in G_l}w_{n,j}^2\widehat{\beta}_j^2}-\frac{2}{n}\sum_{i=1}^{n}\varepsilon_i x_i\widehat{\beta}\\ \leq\;&\|X\beta-X\beta^*\|_n^2+2\sum_{l=1}^{L}\sqrt{\sum_{j\in G_l}w_{n,j}^2\beta_j^2}-\frac{2}{n}\sum_{i=1}^{n}\varepsilon_i x_i\beta.\end{aligned}$$

Therefore, taking $\beta=\beta^*$, we obtain the following inequality:

$$\begin{aligned}\|X\widehat{\beta}-X\beta^*\|_n^2\;\leq\;&2\sum_{l=1}^{L}\left[\sqrt{\sum_{j\in G_l}w_{n,j}^2(\beta_j^*)^2}-\sqrt{\sum_{j\in G_l}w_{n,j}^2\widehat{\beta}_j^2}\right]\\&+\frac{2}{n}\sum_{i=1}^{n}\varepsilon_i x_i\left(\widehat{\beta}-\beta^*\right).\end{aligned}\tag{4.11}$$

Recall that $V_j=n^{-1}\sum_{i=1}^n x_{i,j}\varepsilon_i$ and using the Hölder inequality, we have on the event $\Lambda_{n,p}$

$$\begin{aligned}\frac{2}{n}\sum_{i=1}^{n}\varepsilon_i x_i\left(\widehat{\beta}-\beta^*\right)&=2\sum_{l=1}^{L}\sum_{j\in G_l}V_j\left(\widehat{\beta}_j-\beta_j^*\right)\\&\leq 2\sum_{l=1}^{L}\sqrt{\sum_{j\in G_l}w_{n,j}^{-2}V_j^2}\sqrt{\sum_{j\in G_l}w_{n,j}^2\left(\widehat{\beta}_j-\beta_j^*\right)^2}\\&\leq\sum_{l=1}^{L}\sqrt{\sum_{j\in G_l}w_{n,j}^2\left(\widehat{\beta}_j-\beta_j^*\right)^2}.\end{aligned}\tag{4.12}$$

It follows from (4.11), (4.12) and the definition of the group sparsity set Θ_G (see (4.7)) that

$$\begin{aligned}&\|X\widehat{\beta}-X\beta^*\|_n^2+\sum_{l=1}^{L}\sqrt{\sum_{j\in G_l}w_{n,j}^2\left(\widehat{\beta}_j-\beta_j^*\right)^2}\\&\leq\;2\sum_{l=1}^{L}\sqrt{\sum_{j\in G_l}w_{n,j}^2\left(\widehat{\beta}_j-\beta_j^*\right)^2}+2\sum_{l=1}^{L}\left[\sqrt{\sum_{j\in G_l}w_{n,j}^2(\beta_j^*)^2}-\sqrt{\sum_{j\in G_l}w_{n,j}^2\widehat{\beta}_j^2}\right]\\&=\;2\sum_{l\in\Theta_G}\sqrt{\sum_{j\in G_l}w_{n,j}^2\left(\widehat{\beta}_j-\beta_j^*\right)^2}+2\sum_{l\in\Theta_G}\left[\sqrt{\sum_{j\in G_l}w_{n,j}^2(\beta_j^*)^2}-\sqrt{\sum_{j\in G_l}w_{n,j}^2\widehat{\beta}_j^2}\right].\end{aligned}$$

Therefore using the Minkowski inequality, we have

$$\|X\widehat{\beta} - X\beta^*\|_n^2 + \sum_{l=1}^{L} \sqrt{\sum_{j\in G_l} w_{n,j}^2 \left(\widehat{\beta}_j - \beta_j^*\right)^2} \leq 4 \sum_{l\in\Theta_G} \sqrt{\sum_{j\in G_l} w_{n,j}^2 \left(\widehat{\beta}_j - \beta_j^*\right)^2}$$
$$\leq 4\sqrt{\mathrm{Card}(\Theta_G)} \sqrt{\sum_{l\in\Theta_G} \sum_{j\in G_l} w_{n,j}^2 \left(\widehat{\beta}_j - \beta_j^*\right)^2}. \tag{4.13}$$

Now, let us bound the term $\sum_{l\in\Theta_G} \sum_{j\in G_l} w_{n,j}^2 \left(\widehat{\beta}_j - \beta_j^*\right)^2$. By a simple decomposition, we have

$$\|X\widehat{\beta} - X\beta^*\|_n^2 = \sum_{l\in\Theta_G} \sum_{j\in G_l} \|\xi_j\|_n^2 \left(\widehat{\beta}_j - \beta_j^*\right)^2 + n^{-1} \sum_{i=1}^{n} \left(\sum_{l\notin\Theta_G} \sum_{j\in G_l} x_{i,j}(\widehat{\beta}_j - \beta_j^*) \right)^2$$
$$+ R(\Theta_G), \tag{4.14}$$

where

$$\begin{aligned} R(\Theta_G) &= 2 \sum_{l\in\Theta_G} \sum_{m\notin\Theta_G} \sum_{j\in G_l} \sum_{k\in G_m} \langle \xi_j, \xi_k \rangle_n (\widehat{\beta}_j - \beta_j^*)(\widehat{\beta}_k - \beta_k^*) \\ &+ \sum_{l\in\Theta_G} \sum_{\substack{m\in\Theta_G \\ m\neq l}} \sum_{j\in G_l} \sum_{k\in G_m} \langle \xi_j, \xi_k \rangle_n (\widehat{\beta}_j - \beta_j^*)(\widehat{\beta}_k - \beta_k^*) \\ &+ \sum_{l\in\Theta_G} \sum_{j\in G_l} \sum_{\substack{k\in G_l \\ k\neq j}} \langle \xi_j, \xi_k \rangle_n (\widehat{\beta}_j - \beta_j^*)(\widehat{\beta}_k - \beta_k^*). \end{aligned}$$

Note that $R(\Theta_G)$ is such that

$$|R(\Theta_G)| \leq 2 \sum_{l\in\Theta_G} \sum_{m=1}^{L} \sum_{j\in G_l} \sum_{\substack{k\in G_m \\ k\neq j}} |\langle \xi_j, \xi_k \rangle_n| \, |\widehat{\beta}_j - \beta_j^*| \, |\widehat{\beta}_k - \beta_k^*|.$$

Moreover, since $n^{-1} \sum_{i=1}^{n} \left(\sum_{l\notin\Theta_G} \sum_{j\in G_l} x_{i,j}(\widehat{\beta}_j - \beta_j^*) \right)^2 \geq 0$, the equality (4.14) implies that:

$$\begin{aligned} &\sum_{l\in\Theta_G} \sum_{j\in G_l} w_{n,j}^2 \left(\widehat{\beta}_j - \beta_j^*\right)^2 \\ &\leq \left(\kappa\sigma n^{-1}\sqrt{\log(np)}\right)^2 n\Big(\|X\widehat{\beta} - X\beta^*\|_n^2 - R(\Theta_G)\Big) \\ &\leq \left(\kappa\sigma n^{-1}\sqrt{\log(np)}\right)^2 n\Big(\|X\widehat{\beta} - X\beta^*\|_n^2 + |R(\Theta_G)|\Big) \\ &\leq \left(\kappa\sigma n^{-1}\sqrt{\log(np)}\right)^2 n\Big(\|X\widehat{\beta} - X\beta^*\|_n^2 \\ &\quad +2 \sum_{l\in\Theta_G} \sum_{m=1}^{L} \sum_{j\in G_l} \sum_{\substack{k\in G_m \\ k\neq j}} |\langle \xi_j, \xi_k \rangle_n| \, |\widehat{\beta}_j - \beta_j^*| \, |\widehat{\beta}_k - \beta_k^*|\Big). \end{aligned} \tag{4.15}$$

Let us set $\Pi_{j,k} = w_{n,j}^{-1} w_{n,k}^{-1} \langle \xi_j, \xi_k \rangle_n$. The Cauchy-Schwarz inequality yields

$$
\begin{aligned}
&\sum_{l\in\Theta_G} \sum_{m=1}^{L} \sum_{j\in G_l} \sum_{\substack{k\in G_m \\ k\neq j}} |\langle \xi_j, \xi_k \rangle_n| \, |\widehat{\beta}_j - \beta_j^*| \, |\widehat{\beta}_k - \beta_k^*| \\
&= \sum_{l\in\Theta_G} \sum_{m=1}^{L} \sum_{j\in G_l} \sum_{\substack{k\in G_m \\ k\neq j}} |\Pi_{j,k}| \, w_{n,j} w_{n,k} \, |\widehat{\beta}_j - \beta_j^*| \, |\widehat{\beta}_k - \beta_k^*| \\
&\leq \sum_{l\in\Theta_G} \sum_{m=1}^{L} \sqrt{\sum_{j\in G_l} \sum_{\substack{k\in G_m \\ k\neq j}} \Pi_{j,k}^2} \sqrt{\sum_{j\in G_l} \sum_{k\in G_m} w_{n,j}^2 w_{n,k}^2 \left(\widehat{\beta}_j - \beta_j^*\right)^2 \left(\widehat{\beta}_k - \beta_k^*\right)^2} \\
&\leq \sup_{l\in\Theta_G} \sup_{m=1,\ldots,L} \sqrt{\sum_{j\in G_l} \sum_{\substack{k\in G_m \\ k\neq j}} \Pi_{j,k}^2} \left(\sum_{l=1}^{L} \sqrt{\sum_{j\in G_l} w_{n,j}^2 \left(\widehat{\beta}_j - \beta_j^*\right)^2} \right)^2 \\
&= B(\Theta_G).
\end{aligned}
$$

Combining (4.13), (4.15), the previous inequality and using an elementary inequality of convexity, we obtain

$$
\begin{aligned}
&\|X\widehat{\beta} - X\beta^*\|_n^2 + \sum_{l=1}^{L} \sqrt{\sum_{j\in G_l} w_{n,j}^2 \left(\widehat{\beta}_j - \beta_j^*\right)^2} \\
&\quad \leq 4n^{1/2} \left(\kappa\sigma n^{-1} \sqrt{\log(np)}\right) \sqrt{\mathrm{Card}(\Theta_G)} \sqrt{\|X\widehat{\beta} - X\beta^*\|_n^2 + 2B(\Theta_G)} \\
&\quad \leq 4n^{1/2} \left(\kappa\sigma n^{-1} \sqrt{\log(np)}\right) \sqrt{\mathrm{Card}(\Theta_G)} \sqrt{\|X\widehat{\beta} - X\beta^*\|_n^2} \\
&\quad + 4\sqrt{2} n^{1/2} \left(\kappa\sigma n^{-1} \sqrt{\log(np)}\right) \sqrt{\mathrm{Card}(\Theta_G) B(\Theta_G)}. \qquad (4.16)
\end{aligned}
$$

An application of Assumption (A2)($\mathcal{B}$), with $\mathcal{B} = \Theta_G$ yields

$$
4\sqrt{2} n^{1/2} \left(\kappa\sigma n^{-1} \sqrt{\log(np)}\right) \sqrt{\mathrm{Card}(\Theta_G) B(\Theta_G)} \leq \sum_{l=1}^{L} \sqrt{\sum_{j\in G_l} w_{n,j}^2 \left(\widehat{\beta}_j - \beta_j^*\right)^2}. \qquad (4.17)
$$

It follows from (4.16) and (4.17) that

$$
\|X\widehat{\beta} - X\beta^*\|_n^2 \leq 4n^{1/2} \left(\kappa\sigma n^{-1} \sqrt{\log(np)}\right) \sqrt{\mathrm{Card}(\Theta_G)} \|X\widehat{\beta} - X\beta^*\|_n.
$$

Therefore,

$$
\|X\widehat{\beta} - X\beta^*\|_n^2 \leq 16n \left(\kappa\sigma n^{-1} \sqrt{\log(np)}\right)^2 \mathrm{Card}(\Theta_G) = C\, n^{-1} \log(np)\, \mathrm{Card}(\Theta_G),
$$

where $C = 16\kappa^2\sigma^2$. This ends the proof of Proposition 4.1. □

Proof of Theorem 4.1. Thanks to Proposition 4.1, it is enough to prove that

$$\mathbb{P}\left(\max_{l=1,\dots,L}\sqrt{\sum_{j\in G_l}\|\xi_j\|_n^{-2}V_j^2}\geq 2^{-1}\lambda_{n,p}\right)\leq L(np)^{-(2^{-1}\kappa-1)^2/(2C_*)}.$$

For the sake of simplicity, we set $v_{n,j}=\sqrt{\sum_{i=1}^n x_{i,j}^2}=n^{1/2}\|\xi_j\|_n$. We have

$$\begin{aligned}\mathbb{P}\left(\max_{l=1,\dots,L}\sqrt{\sum_{j\in G_l}\|\xi_j\|_n^{-2}V_j^2}\geq 2^{-1}\lambda_{n,p}\right)&\leq\sum_{l=1}^{L}\mathbb{P}\left(\sqrt{\sum_{j\in G_l}\|\xi_j\|_n^{-2}V_j^2}\geq 2^{-1}\lambda_{n,p}\right)\\&\leq\ L\max_{l=1,\dots,L}\mathbb{P}\left(\sqrt{\sum_{j\in G_l}v_{n,j}^{-2}V_j^2}\geq 2^{-1}\kappa\sigma n^{-1}\sqrt{\log(np)}\right).\end{aligned}\tag{4.18}$$

In order to bound this last term, we introduce the Borell inequality. For further details about this inequality, see, for instance, Adler [1].

Lemma 4.2 (The Borell inequality)**.** *Let $\mathcal{D}$ be a subset of $\mathbb{R}$ and $(\eta_t)_{t\in\mathcal{D}}$ be a centered Gaussian process. Suppose that*

$$\mathbb{E}\left(\sup_{t\in\mathcal{D}}\eta_t\right)\leq N\quad\text{and}\quad\sup_{t\in\mathcal{D}}Var(\eta_t)\leq Q.$$

Then, for any $x>0$, we have

$$\mathbb{P}\left(\sup_{t\in\mathcal{D}}\eta_t\geq x+N\right)\leq\exp(-x^2/(2Q)).$$

Let us consider the set $\mathcal{S}_2^l$ defined by $\mathcal{S}_2^l=\{a=(a_j)_{j\in G_l}\in\mathbb{R}^{\mathrm{Card}(G_l)};\ \sum_{j\in G_l}a_j^2\leq 1\}$, and the centered Gaussian process $\mathcal{Z}(a)$ defined by

$$\mathcal{Z}(a)=\sum_{j\in G_l}a_jV_jv_{n,j}^{-1}.$$

By an argument of duality, we have

$$\sup_{a\in\mathcal{S}_2^l}\mathcal{Z}(a)=\sup_{a\in\mathcal{S}_2^l}\sum_{j\in G_l}a_jv_{n,j}^{-1}V_j=\sqrt{\sum_{j\in G_l}v_{n,j}^{-2}V_j^2}.$$

In order to use Lemma 4.2, let us investigate the upper bounds for $\mathbb{E}(\sup_{a\in\mathcal{S}_2^l}\mathcal{Z}(a))$ and $\sup_{a\in\mathcal{S}_2^l}Var(\mathcal{Z}(a))$, in turn.

The upper bound for $\mathbb{E}(\sup_{a\in\mathcal{S}_2^l}\mathcal{Z}(a))$. Since $V_j\sim\mathcal{N}\left(0,\sigma^2n^{-1}\|\xi_j\|_n^2\right)$, the Cauchy-Schwarz inequality and the fact that $\max_{l\in\{1,\dots,L\}}\mathrm{Card}(G_l)\leq\min\left([\log(np)],p\right)$ imply

$$\begin{aligned}\mathbb{E}(\sup_{a\in\mathcal{S}_2^l}\mathcal{Z}(a))&=\mathbb{E}\left(\sqrt{\sum_{j\in G_l}v_{n,j}^{-2}V_j^2}\right)\leq\sqrt{\sum_{j\in G_l}v_{n,j}^{-2}\mathbb{E}(V_j^2)}\\&=\sqrt{\sum_{j\in G_l}v_{n,j}^{-2}(\sigma^2n^{-1}\|\xi_j\|_n^2)}=\sigma n^{-1}\sqrt{\mathrm{Card}(G_l)}\\&\leq\sigma n^{-1}\sqrt{\log(np)}.\end{aligned}$$

So, we set $N = \sigma n^{-1}\sqrt{\log(np)}$.

The upper bound for $\sup_{a\in\mathcal{S}_2^l} Var(\mathcal{Z}(a))$. We have

$$Var(\mathcal{Z}(a)) = \sum_{j\in G_l}\sum_{k\in G_l} a_j a_k v_{n,j}^{-1} v_{n,k}^{-1}\mathbb{E}(V_j V_k),$$

with $\mathbb{E}(V_j V_k) = n^{-2}\sum_{u=1}^n\sum_{v=1}^n x_{u,j}x_{v,k}\mathbb{E}(\epsilon_u\epsilon_v) = \sigma^2 n^{-1}\langle \xi_j, \xi_k\rangle_n$. This with Assumption (A1) imply

$$\sup_{a\in\mathcal{S}_2^l} Var(\mathcal{Z}(a)) = \sigma^2 n^{-2}\sup_{a\in\mathcal{S}_2^l}\left(\sum_{j\in G_l}\sum_{k\in G_l} a_j a_k \rho_p(j,k)\right) \leq C_*\sigma^2 n^{-2}.$$

So, we set $Q = C_*\sigma^2 n^{-2}$.

Combining the obtained values of N and Q with Lemma 4.2, for any $l \in \{1,\ldots,L\}$, we have

$$\begin{aligned}
&\mathbb{P}\left(\sqrt{\sum_{j\in G_l} v_{n,j}^{-2} V_j^2} \geq 2^{-1}\kappa\sigma n^{-1}\sqrt{\log(np)}\right)\\
&= \mathbb{P}\left(\sup_{t\in\mathcal{D}}\eta_t \geq (2^{-1}\kappa - 1)\sigma n^{-1}\sqrt{\log(np)} + N\right)\\
&\leq \exp\left(-(2^{-1}\kappa-1)^2\sigma^2 n^{-2}\log(np)/(2Q)\right) = (np)^{-(2^{-1}\kappa-1)^2/(2C_*)}. \qquad (4.19)
\end{aligned}$$

Putting (4.18) and (4.19) together, we obtain

$$\mathbb{P}\left(\max_{l=1,\ldots,L}\sqrt{\sum_{j\in G_l}\|\xi_j\|_n^{-2} V_j^2} \geq 2^{-1}\lambda_{n,p}\right) \leq L(np)^{-(2^{-1}\kappa-1)^2/(2C_*)} = u_{n,p}.$$

This ends the proof of Theorem 4.1. □

Chapter 5

Generalization of ℓ_1 constraints for high dimensional regression problems

Abstract: We consider the linear regression problem where the number p of covariates is possibly larger than the number n of observations. In the paper, we propose to approximate the unknown regression parameters under sparsity assumptions with a class of estimators that are motivated by geometrical considerations. Popular estimators based on the control of the ℓ_1 norm of the regression coefficients (such as the LASSO and the Dantzig selector for example) can be seen as special cases of our estimator for which we derive Sparsity Inequalities, i.e., bounds involving the sparsity of the parameter we try to estimate. In such a generalized setup, we show that it is possible to consider variations of the loss function to be minimized. In particular, under a suitable setting, we derive a new estimator that is a transductive version of the LASSO, and we analyze its performance with milder assumptions than in the well-known results about the "usual" LASSO.

5.1 Introduction

In many modern applications, one has to deal with very large datasets. Regression problems may involve a large number of covariates, possibly larger than the sample size. In this situation, a major issue lies in dimension reduction which can be performed through the selection of a small amount of relevant covariates. For this purpose, numerous regression methods have been proposed in the literature, ranging from the classical information criteria such as C_p, AIC and BIC to the more recent regularization-based techniques such as the l_1 penalized least square estimator, known as the LASSO Tibshirani [135], and the Dantzig selector Candès and Tao [35] among many others. Regularized regression methods have recently witnessed several developments due to the attractive feature of computational feasibility, even for high dimensional data when the number of covariates p is large. In the present paper, we focus on the LASSO and the Dantzig selector with what they have in common: both obey to a geometric constraint which we now introduce. Consider the linear regression model $Y = X\beta^* + \varepsilon$, where Y is a vector in $\mathbb{R}^n$, $\beta^* \in \mathbb{R}^p$ is the parameter vector, X is an $n \times p$ real-valued matrix with possibly much fewer rows than columns, $n \ll p$, and ε is a random noise vector in $\mathbb{R}^n$. The analysis of regularized regression methods for high dimensional data usually involves a sparsity assumption on β^* through the *sparsity index* $\|\beta^*\|_0 = \sum_{j=1,\dots,p} \mathbb{I}(\beta_j^* \neq 0)$ where $\mathbb{I}(\cdot)$ is the indicator function. For any $q \geq 1$, $d \geq 0$ and $a \in \mathbb{R}^d$, denote by $\|a\|_q^q = \sum_{i=1}^d |a_i|^q$ and $\|a\|_\infty = \max_{1 \leq i \leq d} |a_i|$, the ℓ_q and the ℓ_∞ norms respectively. When the design matrix X is normalized, the LASSO and the Dantzig selector minimize respectively $\|X\beta\|_2^2$ and $\|\beta\|_1$ under the constraint $\|X'(Y - X\beta)\|_\infty \leq s$ where s is a positive tuning parameter (e.g. Osborne, Presnell, and Turlach [118], Alquier [4] for the dual form of the LASSO). This geometric constraint is central in the approach developed in the present paper and we shall use it in a general perspective. In the sequel, we consider three specific problems in the high-dimensional setting (i.e., $p \geq n$):

Goal 1 - *Prediction:* The reconstruction of the signal $X\beta^*$ with the best possible accuracy is first considered. The quality of the reconstruction with an estimator $\hat{\beta}$ is often measured with the squared error $\|X\hat{\beta} - X\beta^*\|_2^2$. In the standard form, results are stated as follows: under assumptions on the matrix X and with high probability, the prediction error is bounded by $C \log(p) \|\beta^*\|_0$ where C is a positive constant. Such results for the prediction issue have been obtained in Bickel, Ritov, and Tsybakov [17], Bunea [28], Bunea, Tsybakov, and Wegkamp [32] and Bunea, Tsybakov, and Wegkamp [33] for the LASSO and by Bickel, Ritov, and Tsybakov [17] for the Dantzig selector. We also refer to Koltchinskii [91], Koltchinskii [92], Meier, van de Geer, and Bühlmann [110], van de Geer [143], Dalalyan and Tsybakov [46] and Chesneau and Hebiri [44] for related works with different estimators

(non-quadratic loss, penalties slightly different from ℓ_1 and/or random design). The results obtained in the works above-mentioned are optimal up to a logarithmic factor as it has been proved in Bunea, Tsybakov, and Wegkamp [32].

Goal 2 - *Estimation:* Another wishful thinking is that the estimator $\hat{\beta}$ is close to β^* in terms of the ℓ_q distance for $q \geq 1$. The estimation bound is of the form $C \|\beta^*\|_0 (\log(p)/n)^{q/2}$ where C is a positive constant. Such results are stated for the LASSO in Bunea, Tsybakov, and Wegkamp [32, 33] when $q = 1$, for the Dantzig selector in Candès and Tao [35] when $q = 2$ and have been generalized in Bickel, Ritov, and Tsybakov [17] with $1 \leq q \leq 2$ for both the LASSO end the Dantzig selector.

Goal 3 - *Selection:* Since we consider variable selection methods, the identification of the true support $\{j : \beta_j^* \neq 0\}$ of the vector β^* is to be considered. One expects that the estimator $\hat{\beta}$ and the true vector β^* share the same support at least when n grows to infinity. This is known as the variable selection consistency problem and it has been considered for the LASSO estimator in several works Bunea [28], Meinshausen and Bühlmann [112], Meinshausen and Yu [114], Wainwright [152] and Zhao and Yu [164]. Recently, Lounici [101] provided the variable selection consistency of the Dantzig selector. Other popular selection procedures, based on the LASSO estimator, such as the Adaptive LASSO Zou [166], the SCAD Fan and Li [63], the S-LASSO Hebiri [78] and the Group-LASSO Bach [12], have also been studied under this angle.

In the present paper, we address these three goals under a different sparsity assumption than the usual one. Namely, we relate the notion of sparsity to the sparsity index of the vector $P\beta^*$, for some matrix $P \in \mathbb{R}^{p \times p}$. Therefore, sparsity implies here that many components $(P\beta^*)_j$ are equal to 0. Naturally, when P equals I_p, the $p \times p$ identity matrix, we recover the standard assumption on the sparsity index of β^*. We consider a general family of estimators which are defined as solutions of different optimization functions but with the same set of constraint $\|X'(Y - X\beta)\|_\infty \leq s$ (when X is normalized). This family includes the LASSO and the Dantzig selector as special cases. We respond to the three goals described above but with some modifications. Concerning **Goal 1**, instead of the prediction of $X\beta^*$, we aim at recovering $Z\beta^*$ for some matrix $Z \in \mathbb{R}^{m \times p}$ with $m \in \mathbb{N}$. This matrix can be taken equal to X. However, different choices of matrices Z can be considered in such a way to cover other fields such as the transductive setting (Section 5.4.2). As far as estimation (**Goal 2**) and selection (**Goal 3**) are concerned, the whole study takes into consideration the sparse vector $P\beta^*$ instead of β^*. By exploiting the sparsity of $P\beta^*$, we provide similar results

to those presented in the conventional case. However, we need assumptions which are less restrictive in some situations. We also show, in the high-dimensional case $p \gg n$, that it is possible to derive consistent results in situations where β^* is not sparse (in the usual sense).

The paper is organized as follows. In the next section, we specify the setting and the estimators considered in this paper. A short description of the results stated in the sequel are also provided. In Section 5.3, we state our main results. More precisely, we present in Section 5.3.1, different assumptions used through the paper and compare them to the assumptions used in previous works. Using techniques from Bunea, Tsybakov, and Wegkamp [33], we study the performance of the estimators in the different contexts (**Goal 1** to **Goal 3**). Applications of these results are then considered in Section 5.4: the transductive LASSO and the correlation selector Alquier [4]. Finally Section 5.6 is dedicated to the proofs.

5.2 Model and estimator

In this section, we present the general setting. We first introduce the model and the estimators which are considered in the sequel. We also briefly present the results obtained in this paper with some technical arguments which should help the reader to better understand the progression of the paper.

Model. We focus on the usual linear regression model:

$$y_i = x_i\beta^* + \varepsilon_i, \qquad i = 1, \dots, n, \tag{5.1}$$

where the design $x_i = (x_{i,1}, \dots, x_{i,p}) \in \mathbb{R}^p$ is deterministic, $\beta^* = (\beta_1^*, \dots, \beta_p^*)' \in \mathbb{R}^p$ is the unknown parameter vector of interest and $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. centered Gaussian random variables with known variance σ^2. Let X denote the matrix where the i-th line is x_i and the j-th column is X_j with $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, p\}$. Then:

$$X = (x_1', \dots, x_n')' = (X_1, \dots, X_p).$$

For the sake of simplicity, it is often assumed that the observations are normalized in such a way that $X_j'X_j/n = 1$. In this paper, we will not make such an assumption, but we will discuss the consequences of such a normalization in the various results. For this purpose, let us introduce to following notation. For any $j \in \{1, \dots, p\}$,

$$\xi_j = \frac{X_j'X_j}{n} = \frac{1}{n}\sum_{i=1}^{n} x_{i,j}^2 \quad \text{and} \qquad \Xi = \begin{pmatrix} \xi_1^{1/2} & & 0 \\ & \ddots & \\ 0 & & \xi_p^{1/2} \end{pmatrix}.$$

Let us also put $Y = (y_1, \ldots, y_n)'$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$, so we have the following matricial form for Model (5.1): $Y = X\beta^* + \varepsilon$. As mentioned in Section 5.1, we assume that $P\beta^*$ is sparse and base our approach in exploiting this sparsity of the model through the *sparsity index* $\|P\beta^*\|_0 = \sum_{j=1\ldots,p} \mathbb{I}((P\beta^*)_j \neq 0)$.

Estimator. Considering the case where $p \gg n$, dimension reduction is fundamental. It aims at producing consistent estimators while being easy to interpret. The estimators defined in this paper share the same constraint $\|\Xi^{-1}X'(Y - X\beta)\|_\infty \leq s$ where $s > 0$ is a tuning parameter to be specified later. Let us call this constraint the *Dantzig Constraint* ($DC(s)$), as it appears in the definition of the Dantzig selector Candès and Tao [35] when $\xi_j = 1$. This constraint consists in a threshold on the correlation between a covariate X_j, $j \in \{1, \ldots, p\}$ and the residual $Y - X\beta$. Let us define the set:

$$DC(s) = \{\beta \in \mathbb{R}^p : \left\|\Xi^{-1}X'(Y - X\beta)\right\|_\infty \leq s\}, \tag{5.2}$$

where $s > 0$ is a tuning parameter depending on n and p to be specified later. This set is also interpreted as a confidence region for β^* Alquier [4] and produce a geometrical motivation for the study of the following estimators:

$$\textbf{Program I:} \qquad \hat{\beta} = \underset{\beta \in DC(s)}{\operatorname{argmin}} \|Z\beta\|_2^2, \tag{5.3}$$

$$\textbf{Program II:} \qquad \tilde{\beta} = \underset{\beta \in DC(s)}{\operatorname{argmin}} \|\Xi P\beta\|_1, \tag{5.4}$$

where $Z \in \mathbb{R}^{m \times p}$ with $m \in \mathbb{N}$ and $DC(s)$ is the Dantzig Constraint given by (5.2). For unicity reasons, Let us assume from now on the following condition:

– *Kernel condition. The matrix Z is such that* $\ker Z = \ker X$.

The connection between the estimators $\hat{\beta}$ and $\tilde{\beta}$ defined respectively in (5.3) and (5.4) is made in the following way. The *Kernel condition* implies[1] that there exists an invertible matrix $P \in \mathbb{R}^{p \times p}$ such that

$$(X'X)P = (Z'Z). \tag{5.5}$$

When this matrix P coincides with the matrix P used in the definition of $\tilde{\beta}$ in (5.4), **Program I** and **Program II** will produce estimators that have, in theory and in practice, about the same performances. In this paper, both of the solutions $\hat{\beta}$ and $\tilde{\beta}$ are used to predict $Z\beta^*$ whenever $P\beta^*$ is sparse.

Sketch of main results and technical tools. Most of the results which are stated

[1] See the section dedicated to proofs, more precisely Section 5.6.1 page 126 for a proof.

here rely on the exploitation the sparsity index $\|P\beta^*\|_0$, under assumptions on a symmetric matrix W, defined[2] such that

$$(Z'Z)W(X'X) = (X'X). \tag{5.6}$$

In the sequel, in the high dimensional case (when $p \gg n$) and under the sparsity assumption $\|P\beta^*\|_0 \ll p$, we respond mainly to three objectives described in Section 5.1. Here is a sketch of the main results:

Goal 1 - *Prediction:* To ensure to reconstruction of $Z\beta^*$, we prove that, with high probability, $\|Z\beta - Z\beta^*\|_2^2 \leq C \log(p) \|P\beta^*\|_0$ where C is a positive constant and β is either $\hat{\beta}$ or $\tilde{\beta}$ defined in (5.3) and (5.4) respectively.

Goal 2 - *Estimation:* We hope that the solution β (where β is either $\hat{\beta}$ or $\tilde{\beta}$ defined in (5.3) and (5.4) respectively) is such that $P\beta$ is close to $P\beta^*$. We state that with high probability $\|P\beta - P\beta^*\|_1 \leq C\sqrt{\log(p)/n} \|P\beta^*\|_0$ where C is a positive constant.

Goal 3 - *Selection:* Variable selection consistency seems less interesting in our study as soon as $P \neq I_p$. However, we set that with high probability $\|P\beta - P\beta^*\|_\infty \leq C\sqrt{\log(p)/n}$ where C is a positive constant. One then can easily provide variable selection consistency results using such an inequality.

The results stated in the present paper can be interpret as *Sparsity Inequalities* (SIs), bounds which depend on the oracle vector β^* through the sparsity index $\|P\beta^*\|_0$. Furthermore, let us mention that one technical argument to provide our result is based on a dual form of **Program I**: given X, Z and β^*, the relation (5.5) permits us to introduce Γ, defined as

$$\Gamma = \{\gamma \in \mathbb{R}^p : \ X'X\gamma = Z'Z\beta^*\}.$$

This set consists of all vectors γ which belong to the space in which the transformation of β^* is sparse. We can define the sparsest vector in Γ by

$$\gamma^* \in \underset{\gamma \in \Gamma}{\operatorname{argmax}} \operatorname{Card}\{j \in \{1,...,p\}, \quad \gamma_j = 0\} = \underset{\gamma \in \Gamma}{\operatorname{argmax}} \|\gamma\|_0 . \tag{5.7}$$

As the matrix P is invertible, denote $\beta^{**} \in \mathbb{R}^p$ the vector such that $\beta^{**} = P^{-1}\gamma^*$ and consequently, we have $\|\gamma^*\|_0 = \|P\beta^{**}\|_0$, the sparsity index. Because of the Kernel condition, we have[3] $Z\beta^* = Z\beta^{**}$. Then estimating β^{**} instead of β^* does not affect the prediction objective **Goal 1**. From now on, for the sake of simplicity, let P_j denote the j-th row of P, so we can write, for any $j \in \{1,...,p\}$: $\gamma_j^* = P_j\beta^{**}$. One of the main points in the proofs is to link the study of **Program I**, with the study of

$$\textbf{Program I-Dual} \qquad \hat{\gamma} = \underset{\gamma \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \|Y - X\gamma\|_2^2 + 2s \|\Xi\gamma\|_1 + \gamma' M \gamma \right\}, \tag{5.8}$$

[2]Here again, the existence of such a W is given in Section 5.6.1.

[3]See Section 5.6.1.

for some matrix $M \in \mathbb{R}^{p\times p}$ related to Z. An explicit form of M will be provided in Section 5.3.2. Here again, note the form of the program when $\xi_j = 1$ for $j = 1, \ldots, p$:

$$\underset{\gamma\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \|Y - X\gamma\|_2^2 + 2s\,\|\gamma\|_1 + \gamma' M\gamma \right\}.$$

In this paper we present two applications based on the sparsity induced by the transformation $P\beta$. We consider in Section 5.4.1 the Correlation Selector introduced by Alquier [4]. We also consider the Transductive LASSO. In such a case, one choice for Z may be the unlabeled dataset (More details are given in Section 5.4.2).

5.3 Main results

In this section we state all the theoretical results according to the solutions of **Program I** and **Program II**. We start with presenting different assumptions used through the paper in Section 5.3.1. This is the occasion to compare our hypotheses with the ones already used in the previous works mentioned above. Then, we study the performance of the estimators $\hat{\beta}$ and $\tilde{\beta}$ defined by **Programs I** and **II** respectively. We also relate the solutions of **Programs I** to those of **Program I-Dual** thanks to a link between Z and M (Section 5.3.2).

5.3.1 Assumptions

We present here the assumptions we need to state the Sparsity Inequalities provided in Sections 5.3.3 and 5.3.4. Note that they essentially involve the matrix Ω defined as follows:

$$\Omega = \frac{1}{n}(X'X)W(X'X). \tag{5.9}$$

We denote by $\Omega_{j,k}$, the (j,k) coefficient of Ω. We just remind that the definition of Ω involves the matrix W, given by (5.6). Using this notation we introduce the assumptions with more precision. The first assumption is used to respond to the prediction **Goal 1** and estimation **Goal 2** objectives.

– *Assumption (A1). There is a constant $c > 0$ such that, for any $\alpha \in \mathbb{R}^p$ such that*

$$\sum_{j:\gamma_j^*=0} \xi_j^{\frac{1}{2}} |\alpha_j| \leq 3 \sum_{j:\gamma_j^*\neq 0} \xi_j^{\frac{1}{2}} |\alpha_j|,$$

where γ^* is given by (5.7)*, we have*

$$\sum_{j:\gamma_j^*\neq 0} \alpha_j^2 \leq c\alpha'\Omega\alpha. \tag{5.10}$$

When we deal with variable selection **Goal 3**, we replace Assumption (A1) by the following:

– *Assumption (A2). Let us assume that , for any* $j \in \{1, ..., p\}$, $\xi_j = 1$ *and that*

$$\rho = \sup_{j \in \{1,...,p\}} \sup_{k \neq j} |\Omega_{j,k}| \leq \frac{\inf_{\gamma_j^* \neq 0} \Omega_{j,j}}{14 \|\gamma^*\|_0}. \tag{5.11}$$

We now give some comments about the assumptions. First, note that both of these assumptions are modifications of the well-known *mutual coherence* condition introduced by Donoho, Elad, and Temlyakov [51] - but the mutual coherence condition is about the Gram matrix $n^{-1}X'X$ while the assumptions presented here involve the matrix Ω. Assumption (A2) is closer to the mutual coherence condition than Assumption (A1). It is also more restrictive. For a selection purpose **Goal 3**, the mutual coherence assumption is used in Lounici [101]. Moreover Assumption (A1) can be seen as a modification of more general assumptions that can be found in Bickel, Ritov, and Tsybakov [17], Bunea [28], Bunea, Tsybakov, and Wegkamp [32] and Bunea, Tsybakov, and Wegkamp [33]. For example, using a slight modification of a proof given in Bickel, Ritov, and Tsybakov [17], we can prove the following result.

Lemma 5.1. *Assumption (A2)* $\Rightarrow$ *Assumption (A1) with constant* $c = \frac{2}{\inf_{\gamma_j^* \neq 0} \Omega_{j,j}}$.

For the sake of completeness, the proof is given in Section 5.6 in its full length. It is known that in the high dimensional case ($p \gg n$), such assumptions are hard to relax when the considered estimators are solution of a convex minimization problem as in (5.3) and (5.4); see Bickel, Ritov, and Tsybakov [17] and Bunea, Tsybakov, and Wegkamp [33, Remarks 4 and 5] for other comments on that topic. In the case where $n \geq p$, if $Z'Z$ and $X'X$ are invertible matrices, then we can find a constant c such that, for any $\alpha \in \mathbb{R}^p$, $\alpha'\alpha \leq c\alpha'\Omega\alpha$. Of course, this implies that Assumption (A1) is satisfied with this specific choice of the constant c.

5.3.2 Dual form of Program I

Let us put (remember that W is given by (5.6)):

$$M = (X'X)W(X'X) - (X'X). \tag{5.12}$$

Theorem 5.1. *All the solutions* $\hat{\beta}$ *of* **Program I** *are given by* $(Z'Z)\hat{\beta} = (X'X)\hat{\gamma}$ *where* $\hat{\gamma}$ *is any solution of* **Program I-Dual**, *with* M *given by* (5.12). *Moreover, when* $\hat{\gamma}$ *is unique, all the solutions of* **Program I** *give the same value to* $X\hat{\beta}$, *and also to* $Z\hat{\beta}$.

The proof is given in Section 5.6. Note that, taking $Z = X$ gives $M = 0$ and allows the choice $P = I_p$. So, $\hat{\gamma}$ is a solution the the LASSO program and we can take $\hat{\beta} = \hat{\gamma}$. Theorem 5.1 can be seen as a generalization of the dual form of the LASSO given by Alquier [4] and Osborne, Presnell, and Turlach [118].

5.3.3 Sparse inequalities and sup-norm bound for Program I

In this section, we provide sparse inequalities (SIs) and a sup-norm bound for **Program I**. First Theorem 5.2 provides bounds on the squared error (corresponding to **Goal 1**) and to the distance between the estimated and true parameters (corresponding to **Goal 2**). The main key is to use the sparsity index $\|\gamma^*\|_0 = \|P\beta^{**}\|_0$ where γ^* is given by (5.7).

Theorem 5.2. *Let us consider the linear regression model* (5.1). *Let $\hat{\gamma}$ be any solution of the the quadratic* **Program I-Dual**. *Let $\hat{\beta} = P^{-1}\hat{\gamma}$, so by Theorem 5.1, $\hat{\beta}$ is a solution of* **Program I**. *Let us choose $\kappa > 2\sqrt{2}$ and $s = \kappa\sigma\sqrt{n\log(p)}$. Under Assumption (A1), with probability larger than $1 - p^{1-\frac{\kappa^2}{8}}$, we have*

$$\left\|Z(\hat{\beta} - \beta^{**})\right\|_2^2 = \left\|Z(\hat{\beta} - \beta^*)\right\|_2^2 \leq 16c\kappa^2\sigma^2\log(p)\sum_{P_j\beta^{**}\neq 0}\xi_j, \tag{5.13}$$

and

$$\|\Xi\,(\hat{\gamma} - \gamma^*)\|_1 = \left\|\Xi P\left(\hat{\beta} - \beta^*\right)\right\|_1 \leq 16c\kappa\sigma\sqrt{\frac{\log(p)}{n}}\sum_{P_j\beta^{**}\neq 0}\xi_j. \tag{5.14}$$

The proof of this result can be found in Section 5.6.

Corollary 5.1. *Under the conditions of Theorem 5.2, if we moreover assume that the matrix X is normalized in order to have $\xi_j = 1$ for any j, then we have, with probability larger than $1 - p^{1-\frac{\kappa^2}{8}}$,*

$$\left\|Z(\hat{\beta} - \beta^*)\right\|_2^2 \leq 16c\kappa^2\sigma^2\log(p)\,\|P\beta^{**}\|_0\,, \tag{5.15}$$

and

$$\left\|P\left(\hat{\beta} - \beta^{**}\right)\right\|_1 \leq 16c\kappa\sigma\sqrt{\frac{\log(p)}{n}}\,\|P\beta^{**}\|_0\,. \tag{5.16}$$

Theorem 5.2 and its corollary state that with high probability, we can consistently perform prediction **Goal 1** and estimation **Goal 2**, exploiting the sparsity of the projected β^*, that is $\gamma^* = P\beta^{**}$. Note that the obtained rates are near optimal up to a logarithmic factor. Indeed, in our setting, it is proved in Bunea, Tsybakov, and Wegkamp [32, Theorem 5.1] that the optimal rate for the l_2 risk (5.15) is $\log\left(\frac{p}{\|P\beta^{**}\|_0} + 1\right)\|P\beta^{**}\|_0$.

We provide now a bound on the sup-norm $\|\gamma^* - \hat{\gamma}\|_\infty$. As described in Remark 5.1, such a result would help us to easily get an estimator of γ^* which is consistent in variable selection **Goal 3**. That is, an estimator which succeed to recover the true support of γ^*, the sparse projection of β^*.

Theorem 5.3. *Let us consider the linear regression model* (5.1). *Let $\hat{\gamma}$ be any solution of the quadratic* **Program I-Dual**. *Let us choose $\kappa > 2\sqrt{2}$ and $s = \kappa\sigma\sqrt{n\log(p)}$. Under*

Assumption (A2), with probability larger than $1-p^{1-\frac{\kappa^2}{8}}$*, we have simultaneously Inequalities* (5.15), (5.16) *and*

$$\|\hat{\gamma}-\gamma^*\|_\infty \leq \frac{3\kappa\sigma}{\inf_{1\leq j\leq p}\Omega_{j,j}}\sqrt{\frac{\log(p)}{n}}. \tag{5.17}$$

The proof of this result can be found in Section 5.6. Note also that these results generalize the results obtained by Bunea [28], Bunea, Tsybakov, and Wegkamp [33] and Lounici [101] as the LASSO can be seen as special cases of our estimator.

5.3.4 Sparsity Inequalities and sup-norm bound for Program II

For readability, let us recall **Program II**:

$$\begin{cases} \operatorname{argmin}_{\beta\in\mathbb{R}^p} \|\Xi P\beta\|_1 \\ \text{s.t. } \left\|\Xi^{-1}X'(Y-X\beta)\right\|_\infty \leq s. \end{cases}$$

Here again we want to estimate $Z\beta^* = Z\beta^{**}$ when $P\beta^{**}$ is assumed to be sparse (as in Theorem 5.2). In such a context, analog results to those obtained in Section 5.3.3 can be obtained. First we state:

Theorem 5.4. *Let us assume that Assumption (A1) is satisfied. Let* $\tilde{\beta}$ *be a solution of* **Program II**. *Let us choose* $\kappa > 2\sqrt{2}$ *and* $s = \kappa\sigma\sqrt{n\log(p)}$. *Then with probability larger than* $1-p^{1-\frac{\kappa^2}{8}}$*, we have*

$$\left\|Z(\tilde{\beta}-\beta^*)\right\|_2^2 \leq 9c\kappa^2\sigma^2\log(p)\sum_{P_j\beta^{**}\neq 0}\xi_j,$$

and

$$\left\|\Xi P\left(\tilde{\beta}-\beta^{**}\right)\right\|_1 \leq 6c\kappa\sigma\sqrt{\frac{\log(p)}{n}}\sum_{P_j\beta^{**}\neq 0}\xi_j.$$

The proof is given in Section 5.6. In the same way as for the solution of **Program I**, we can provide an analog to Theorem 5.3.

Theorem 5.5. *With the notations of the previous theorem, under Assumption (A2), if we moreover assume that the matrix* X *is normalized in order to have* $\xi_j = 1$ *for any* j*, with probability greater than* $1-p^{1-\frac{\kappa^2}{8}}$*, we have simultaneously*

$$\left\|Z(\tilde{\beta}-\beta^*)\right\|_2^2 \leq 9c\kappa^2\sigma^2\log(p)\left\|P\beta^{**}\right\|_0,$$

$$\left\|P\left(\tilde{\beta}-\beta^{**}\right)\right\|_1 \leq 6c\kappa\sigma\sqrt{\frac{\log(p)}{n}}\left\|P\beta^{**}\right\|_0,$$

and

$$\left\|P\left(\tilde{\beta}-\beta^{**}\right)\right\|_\infty \leq \frac{2\kappa\sigma}{\inf_{1\leq j\leq p}\Omega_{j,j}}\sqrt{\frac{\log(p)}{n}}.$$

Note that these results generalize the results obtained by Bickel, Ritov, and Tsybakov [17], Lounici [101] as the Dantzig selector can be seen as special cases of our estimator.

Remark 5.1. *Thanks to Theorems 5.3 and 5.5, we can easily construct a sign-consistent estimator (an estimator $\overline{\gamma}$ of the vector γ^* given by* (5.7) *such that it shares asymptotically and in probability, not only the same support (sparsity set) but also the same sign of its components with γ^*). This estimator $\overline{\gamma}$ is defined as a thresholded version of $\hat{\gamma}$ where $\hat{\gamma}$ is either solution of* **Program I** *or is equal to $P\tilde{\beta}$ with $\tilde{\beta}$ solution of* **Program II**. *The threshold used is respectively equal to the bound in the sup-norm result appearing in Theorem 5.3 and 5.5. Some more technical tools to establish the result are needed and we refer to Bunea [28] and Lounici [101] for more details.*

5.4 Applications

We present now two applications of the estimators considered in the previous sections.

5.4.1 The Correlation Selector

In Alquier [4], an estimator is introduced for the case where most of the X_j's have a null correlation with Y while we think that all together, these covariates can provide a good prevision for Y: namely, we assume that the $(X'X)\beta^*$ is sparse.

Here (in this subsection only), let us assume that $(X'X)$ is invertible - this implies that $p \leq n$. Let us also assume that X is normalized, so $\xi_j = 1$ for any j. Then, if we take $Z = (X'X)$ then we can take $P = (X'X)$ too and $\Omega = I_p/n$, this means that Assumptions (A1) and (A2) are satisfied in any case. So, **Program I** involves the minimization of $\|(X'X)\beta\|_2^2$ while **Program II** involves the minimization of $\|(X'X)\beta\|_1$. Actually, it is proved by Alquier [4] that the estimator defined by

$$\hat{\beta}_{CS} = \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \|(X'X)\beta\|_q^q \qquad \text{s.t. } \left\|X'\left(Y - X\beta\right)\right\|_\infty \leq s$$

for any $q \geq 1$ does not depend on q. Theorem 5.5 gives, with probability larger than $1 - n^{1-\frac{\kappa^2}{8}}$,

$$\left\|\frac{X'X}{n}(\hat{\beta}_{CS} - \beta^*)\right\|_2^2 \leq 9\kappa^2\sigma^2\frac{\log(p)}{n}\left\|(X'X)\beta^*\right\|_0, \tag{5.18}$$

$$\left\|\frac{X'X}{n}\left(\hat{\beta}_{CS} - \beta^*\right)\right\|_1 \leq 6\kappa\sigma\sqrt{\frac{\log(p)}{n}}\left\|(X'X)\beta^*\right\|_0,$$

and

$$\left\|\frac{X'X}{n}\left(\hat{\beta}_{CS} - \beta^*\right)\right\|_\infty \leq 2\kappa\sigma\sqrt{\frac{\log(p)}{n}}.$$

Note that Inequality (5.18) was already proved by Alquier [4], but that the proof by Alquier [4] could not be extended to the cases of the ℓ_1-norm and ℓ_∞-norm. However,

the proof by Alquier [4] allows to extend Inequality (5.18) to the case where $p \geq n$ without hypotheses; here, $\Omega = I/n$ would not be possible in this case and so we would have additional hypotheses. Finally, notice that Inequality (5.18) does not involve a natural norm. However, adding one more hypotheses, we obtain the following result.

Corollary 5.2. *Let us assume that there is a $\zeta > 0$ such that $\zeta(X'X)/n - I_p$ is definite positive. Then we have, with probability larger than $1 - n^{1-\frac{\kappa^2}{8}}$,*

$$\left\| X(\hat{\beta}_{CS} - \beta^*) \right\|_2^2 \leq 9\zeta\kappa^2\sigma^2 \log(p) \left\| (X'X)\beta^* \right\|_0 .$$

5.4.2 The transductive LASSO

By an application of Theorem 5.1, the equivalence between **Program I**, applied with and $Z = X$, and the LASSO, is clear. However, Theorem 5.2 allows to extend the LASSO to the so-called transductive setting introduced by Vapnik [145].

In this setting, we have $Y = \overline{X}\overline{\beta}^* + \varepsilon$ where $\overline{X}$ is a matrix containing the observation vectors $\overline{x}_i$ and $\overline{\beta}^*$ is "sparse". However, we are not interested in the estimation of $\overline{\beta}^*$ or $\overline{X\beta}^*$: we have another set of m points - say $\overline{x}_{n+1},\ldots,\overline{x}_{n+m}$. We choose

$$\overline{Z} = (\overline{x}'_{n+1},\ldots,\overline{x}'_{n+m})',$$

satisfying $\ker \overline{Z} = \ker \overline{X}$ and so the objective is the estimation of $\overline{Z\beta}^*$, that is the prediction of the value of the regression function on a particular set of points. We have also $\overline{P}$ such that $\overline{X}'\overline{X}\overline{P} = \overline{Z}'\overline{Z}$ and $\overline{W}$ such that $(\overline{Z}'\overline{Z})\overline{W}(\overline{X}'\overline{X}) = (\overline{X}'\overline{X})$.

It is argued by Vapnik [145] that in this setting, a bound on the general performances of an estimator of $\overline{\beta}^*$ is (often) useless and that the statistician should focus on a particular method to estimate $\overline{Z\beta}^*$, that can be easier.

Note that a direct application of Theorem 5.2 (for example) would lead to an unsatisfying result as it would not assume that $\overline{\beta}^*$ is sparse, but $\overline{P\beta}^*$. The problem can be solved in the following way. Note that

$$Y = \overline{X\beta}^* + \varepsilon = (\overline{XP})(\overline{P}^{-1}\overline{\beta}^*) + \varepsilon = X\beta^* + \varepsilon$$

where we put $X = \overline{XP}$, $\beta^* = \overline{P}^{-1}\overline{\beta}^*$, and

$$\overline{Z\beta}^* = (\overline{ZP})(\overline{P}^{-1}\overline{\beta}^*) = Z\beta^*$$

where $Z = \overline{ZP}$. Note that the choice $W = \overline{P}^{-1}\overline{W}(\overline{P}^{-1})'$ satisfies $(Z'Z)W(X'X) = (X'X)$.

Note that ξ_j will still denote the j-th diagonal element of $X'X/n$. When we apply Theorem 5.2 to this setting, we have to make hypotheses about

$$\Omega = \frac{1}{n}(X'X)W(X'X) = \frac{1}{n}\overline{P}'(\overline{X}'\overline{X})\overline{P}\overline{P}^{-1}\overline{W}(\overline{P}^{-1})'\overline{P}'(\overline{X}'\overline{X})\overline{P}$$
$$= \frac{1}{n}\overline{P}'(\overline{X}'\overline{X})\overline{W}(\overline{X}'\overline{X})\overline{P} = \frac{1}{n}(\overline{Z}'\overline{Z}).$$

So we will not need any assumption about $\overline{X}$!

Corollary 5.3. *Let us assume that there is a constant $c > 0$ such that, for any $\alpha \in \mathbb{R}^p$ satisfying $\sum_{j:\overline{\beta}^*_j=0} \xi_j^{\frac{1}{2}} |\alpha_j| \leq 2 \sum_{j:\overline{\beta}^*_j \neq 0} \xi_j^{\frac{1}{2}} |\alpha_j|$, we have $\alpha'\alpha \leq (c/n)\alpha'(\overline{Z}'\overline{Z})\alpha$. Let $\hat{\beta}$ be any solution of the the quadratic program*

$$\underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|Z\beta\|_2^2 \qquad s.t. \quad \left\|\Xi^{-1} X' (Y - X\beta)\right\|_\infty \leq s$$

and let [4] $\tilde{\beta} = \overline{P}\hat{\beta}$. *Let us choose $\kappa > 2\sqrt{2}$ and $s = \kappa\sigma\sqrt{n\log(p)}$. Then with probability greater than $1 - p^{1-\frac{\kappa^2}{8}}$, we have the total error on the prevision of $x_{n+i}\beta^*$ for $1 \leq i \leq n$ that is given by*

$$\left\|\overline{Z}(\tilde{\beta} - \overline{\beta}^*)\right\|_2^2 \leq 16c\kappa^2\sigma^2 \log(p) \sum_{\overline{\beta}^* \neq 0} \xi_j.$$

Note that in the theorems about LASSO, there are always hypotheses about the matrix X that is *given* to the statistician. Here, the only hypotheses is about $\overline{Z}$ that is *chosen* by the statistician. For example, if $\overline{Z}'\overline{Z}$ is not enough well conditioned, it is possible to "add vectors" in $\overline{Z}$, namely, to choose a larger m. *This is a real improvement.* However, there is a *price to pay* for this improvement, as explained now.
Remark that the matrix $\overline{Z}$ is not normalized. If one is to impose such a normalization, the more natural thing to do is to impose that the diagonal elements of $\overline{Z}'\overline{Z}$ are constant. But in this case, one can check that (in general) *it is no longer possible to normalize X in such a way that $\xi_j = 1$* for any j. So, without any assumption about a link between $\overline{Z}$ and $\overline{X}$ that would allow to have more information on X, it is not possible to control $\sum_{\overline{\beta}^* \neq 0} \xi_j$ by $\|\overline{\beta}^*\|_0$.

5.5 Conclusion

Based on a geometrical remark by Alquier [4], we studied the family of estimators defined by

$$\underset{\beta \in DC(s)}{\operatorname{argmin}} \|\mathcal{M}\beta\|_q^q$$

that includes the LASSO, the Dantzig Selector, the Correlation Selector and the transductive LASSO in some particular cases: $q = 1$ and $q = 2$ for a quite general matrix $\mathcal{M}$, and $q \geq 1$ for the particular case $\mathcal{M} = (X'X)$.

[4] Equivalently, we could define

$$\tilde{\beta} = \arg\min_{\beta} \left\{ -2Y'X\beta + \beta'\overline{Z}'\overline{Z}\beta + 2s \|\Xi\beta\|_1 \right\}.$$

Future works could include the theoretical study of our estimator for a general matrix $\mathcal{M}$ and $q \notin \{1, 2\}$, as well as an extension of the LARS algorithm to compute efficiently the solutions of Program (5.3) and (5.4).

5.6 Proofs

5.6.1 Basic algebra results

In this subsection, we prove the basic algebra results claimed in the introduction.

Proof that the kernel condition implies the existence of P. As $(Z'Z)$ is symmetric, we can diagonalize it in an orthogonal basis, given by a matrix Q: there is a $q \in \{0, ..., p\}$ and $\lambda_1, ..., \lambda_q > 0$ with

$$(Z'Z) = Q' \left(\begin{array}{c|c} D & 0 \\ \hline 0 & 0 \end{array} \right) Q, \text{ with } D = \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_q \end{pmatrix}.$$

Remark also that this implies that $Ker(Z'Z) \perp Im(Z'Z)$. Now, remark that the kernel condition implies that $Ker(X'X) = Ker(Z'Z)$ and, because $(X'X)$ is symmetric, we have $Ker(X'X) \perp Im(X'X)$. This implies that $Im(Z'Z) = Im(X'X)$. So, $(X'X)$ can be "partially diagonalized" in the basis Q, in the sense that

$$(X'X) = Q' \left(\begin{array}{c|c} B & 0 \\ \hline 0 & 0 \end{array} \right) Q$$

where B is some invertible $q \times q$ matrix. Now, let us put

$$P = Q' \left(\begin{array}{c|c} B^{-1}D & 0 \\ \hline 0 & I_{p-q} \end{array} \right) Q.$$

We can easily check that P is invertible and that

$$(X'X)P = Q' \left(\begin{array}{c|c} B & 0 \\ \hline 0 & 0 \end{array} \right) QQ' \left(\begin{array}{c|c} B^{-1}D & 0 \\ \hline 0 & I_{p-q} \end{array} \right) Q = Q' \left(\begin{array}{c|c} D & 0 \\ \hline 0 & 0 \end{array} \right) Q = (Z'Z).$$

□ □

Proof of the existence of W. This proof uses the same arguments, we just put

$$W = Q' \left(\begin{array}{c|c} D^{-1} & 0 \\ \hline 0 & I_{p-q} \end{array} \right) Q$$

and we check that $(Z'Z)W(X'X) = (X'X)$ and that W is symmetric. □ □

Proof of $Z\beta^* = Z\beta^{**}$. We have $(Z'Z)\beta^* = (X'X)\gamma^* = (Z'Z)\beta^{**}$ by definition. So $(Z'Z)(\beta^* - \beta^{**}) = 0$ and so $(\beta^* - \beta^{**})'(Z'Z)(\beta^* - \beta^{**}) = 0$ and $Z(\beta^* - \beta^{**}) = 0$. □ □

5.6.2 Proof of Theorem 5.1

Proof. Let us remark that **Program I** can be written

$$\min_{\beta\in\mathbb{R}^p} \beta(Z'Z)\beta \qquad \text{s. t.} \quad \left\|\Xi^{-1}X'(Y-X\beta)\right\|_\infty \leq s \tag{5.19}$$

Let us write the Lagrangian of this program:

$$\mathcal{L}(\beta,\lambda,\mu) = \beta(Z'Z)\beta + \lambda'\Xi^{-1}\left[X'(X\beta-Y) - sE\right] + \mu'\Xi^{-1}\left[X'(Y-X\beta) - sE\right]$$

with $E=(1,...,1)'$, and for any j, $\lambda_j \geq 0$, $\mu_j \geq 0$ and $\lambda_j\mu_j = 0$. Any solution $\underline{\beta} = \underline{\beta}(\lambda,\mu)$ of **Program I** must satisfy

$$0 = \frac{\partial\mathcal{L}}{\partial\beta}(\underline{\beta},\lambda,\mu) = 2\underline{\beta}'(Z'Z) + (\lambda-\mu)'\Xi^{-1}(X'X),$$

so

$$(Z'Z)\underline{\beta} = (X'X)\frac{1}{2}\Xi^{-1}(\mu-\lambda).$$

Note that the conditions $\lambda_j \geq 0$, $\mu_j \geq 0$ and $\lambda_j\mu_j = 0$ means that there is a $\gamma_j \in \mathbb{R}$ such that $\gamma_j = \xi_j^{\frac{1}{2}}(\mu_j-\lambda_j)/2$, $|\gamma_j| = \xi_j^{\frac{1}{2}}(\lambda_j+\mu_j)/2$, and so $\lambda_j = 2(\gamma_j/\xi_j^{\frac{1}{2}})_-$ and $\mu_j = 2(\gamma_j/\xi_j^{\frac{1}{2}})_+$, where $(a)_+ = max(a;0)$ and $(a)_- = max(-a;0)$. Let also γ denote the vector which j-th component is exactly γ_j, we obtain:

$$(Z'Z)\underline{\beta} = (X'X)\gamma. \tag{5.20}$$

Note that this also implies that:

$$\underline{\beta}'(Z'Z)\underline{\beta} = \underline{\beta}'(X'X)\gamma = \underline{\beta}'(Z'Z)W(X'X)\gamma = \gamma'(X'X)W(X'X)\gamma.$$

Using these relations, the Lagrangian may be written:

$$\begin{aligned}\mathcal{L}(\underline{\beta},\lambda,\mu) &= \gamma'(X'X)W(X'X)\gamma + 2\gamma'X'Y - 2\gamma'(X'X)\underline{\beta} - 2s\sum_{j=1}^{p}\xi_j^{\frac{1}{2}}|\gamma_j| \\ &= 2\gamma'X'Y - \gamma'(X'X)W(X'X)\gamma - 2s\left\|\Xi\gamma\right\|_1\end{aligned}$$

Note that λ and β, and so γ, should maximize this value. Hence, γ is to minimize

$$-2\gamma'X'Y + \gamma'(X'X)W(X'X)\gamma + 2s\|\Xi\gamma\|_1 + Y'Y$$

Now, note that

$$Y'Y - 2\gamma'X'Y = \|Y-X\gamma\|_2^2 - \gamma'(X'X)\gamma$$

and then γ is also to minimize

$$\|Y-X\gamma\|_2^2 + 2s\left\|\Xi\gamma\right\|_1 + \gamma'\left[(X'X) - (X'X)W(X'X)\right]\gamma,$$

what is claimed in the theorem. Let $\underline{\gamma}$ denote a solution of this program. Equation (5.20) implies that if $\underline{\beta}$ is a solution of **Program I** then $(Z'Z)\underline{\beta} = (X'X)\underline{\gamma}$. Let $\overline{\beta}$ we another solution of **Program I**, note that we also have $(Z'Z)\overline{\beta} = (X'X)\underline{\gamma}$. Then $(Z'Z)(\underline{\beta}-\overline{\beta}) = 0$ so $(\underline{\beta}-\overline{\beta})$ belongs to $\ker Z$ and so to $\ker X$. This ends the proof. □ □

5.6.3 Proof of Lemma 5.1

Proof. Remember that Assumption (A2) implies among others that $\xi_j = 1$ for any j. Let $\alpha \in \mathbb{R}^p$ satisfy: $\sum_{j:\gamma_j^*=0} |\alpha_j| \leq 3 \sum_{j:\gamma_j^*\neq 0} |\alpha_j|$. We have:

$$\begin{aligned}
\alpha'\Omega\alpha &= \sum_{j:\gamma_j^*\neq 0} \Omega_{j,j}\alpha_j^2 + \sum_{j:\gamma_j^*=0} \sum_{k:\gamma_k^*=0} \Omega_{j,k}\alpha_j\alpha_k \\
&\quad +2 \sum_{j:\gamma_j^*\neq 0} \sum_{k:\gamma_k^*=0} \Omega_{j,k}\alpha_j\alpha_k + \sum_{j:\gamma_j^*\neq 0} \sum_{\substack{k:\,\gamma_k^*\neq 0 \\ k\neq j}} \Omega_{j,k}\alpha_k\alpha_j \\
&\geq \sum_{j:\gamma_j^*\neq 0} \Omega_{j,j}\alpha_j^2 + 2 \sum_{j:\gamma_j^*\neq 0} \sum_{k:\gamma_k^*=0} \Omega_{j,k}\alpha_j\alpha_k + \sum_{j:\gamma_j^*\neq 0} \sum_{\substack{k:\,\gamma_k^*\neq 0 \\ k\neq j}} \Omega_{j,k}\alpha_j\alpha_k.
\end{aligned}$$

So we have

$$\begin{aligned}
\sum_{j:\gamma_j^*\neq 0} \Omega_{j,j}\alpha_j^2 &\leq \alpha'\Omega\alpha - 2 \sum_{j:\gamma_j^*\neq 0} \sum_{k:\gamma_k^*=0} \Omega_{j,k}\alpha_j\alpha_k - \sum_{j:\gamma_j^*\neq 0} \sum_{\substack{k:\,\gamma_k^*\neq 0 \\ k\neq j}} \Omega_{j,k}\alpha_j\alpha_k \\
&\leq \alpha'\Omega\alpha + \left(\sup_{\gamma_j^*\neq 0} \sup_{k\neq j} |\Omega_{j,k}|\right) \left[2\left(\sum_{\gamma_j^*\neq 0} |\alpha_j|\right)\left(\sum_{\gamma_k^*=0} |\alpha_k|\right) + \left(\sum_{\gamma_j^*\neq 0} |\alpha_j|\right)^2\right] \\
&\leq \alpha'\Omega\alpha + 7\left(\sup_{\gamma_j^*\neq 0} \sup_{k\neq j} |\Omega_{j,k}|\right)\left(\sum_{\gamma_j^*\neq 0} |\alpha_j|\right)^2 = \alpha'\Omega\alpha + 7\rho\left(\sum_{\gamma_j^*\neq 0} |\alpha_j|\right)^2. \quad (5.21)
\end{aligned}$$

On the other hand, using the Cauchy-Schwarz inequality, we have

$$\left(\sum_{\gamma_j^*\neq 0} |\alpha_j|\right)^2 \leq \|\gamma^*\|_0 \sum_{\gamma_j^*\neq 0} \alpha_j^2 \leq \frac{\|\gamma^*\|_0}{\inf_{\gamma_j^*\neq 0} \Omega_{j,j}} \sum_{\gamma_j^*\neq 0} \Omega_{j,j}\alpha_j^2. \quad (5.22)$$

Combining (5.21) and (5.22), we obtain

$$\sum_{j:\gamma_j^*\neq 0} \Omega_{j,j}\alpha_j^2 \leq \frac{1}{1 - 7\frac{\|\gamma^*\|_0}{\inf_{\gamma_j^*\neq 0} \Omega_{j,j}}\rho} \alpha'\Omega\alpha.$$

Now, remember that we assumed that $\rho \leq \frac{\inf_{\gamma_j^*\neq 0} \Omega_{j,j}}{14\|\gamma^*\|_0}$ by hypotheses and conclude by

$$\sum_{\gamma_j^*\neq 0} \alpha_j^2 \leq \frac{1}{\inf_{\gamma_j^*\neq 0} \Omega_{j,j}} \sum_{j:\gamma_j^*\neq 0} \Omega_{j,j}\alpha_j^2 \leq \frac{2\alpha'\Omega\alpha}{\inf_{\gamma_j^*\neq 0} \Omega_{j,j}}.$$

□ □

5.6.4 A useful Lemma

Lemma 5.2. *Let $\Lambda_{n,p}$ be the random event defined by*

$$\Lambda_{n,p} = \left\{\forall j \in \{1, ..., p\}, \quad 2|V_j| \leq s\xi_j^{\frac{1}{2}}\right\}, \quad (5.23)$$

where $V_j = \sum_{i=1}^n x_{i,j}\varepsilon_i$. *Let us choose a* $\kappa > 2\sqrt{2}$ *and* $s = \kappa\sigma\sqrt{n\log(p)}$. *Then*

$$\mathbb{P}\left(\Lambda_{n,p}\right) \geq 1 - p^{1-\frac{\kappa^2}{8}}.$$

Proof. Remember that $\xi_j = \frac{1}{n}\sum_{i=1}^n x_{i,j}^2$. Since $V_j = \sum_{i=1}^n x_{i,j}\varepsilon_i \sim \mathcal{N}(0, n\xi_j\sigma^2)$, an elementary Gaussian inequality gives

$$\begin{aligned}\mathbb{P}\left(\max_{l=1,\dots,p} s^{-1}\xi_j^{-\frac{1}{2}}|V_l| \geq 2^{-1}\right) &\leq p \max_{l=1,\dots,p}\mathbb{P}\left(s^{-1}\xi_j^{-\frac{1}{2}}|V_l| \geq 2^{-1}\right)\\ &\leq p\exp\left(-\kappa^2\log(p)/8\right) = p^{1-\kappa^2/8}.\end{aligned}$$

□ □

5.6.5 Proof of Theorem 5.2

The proof follows the technique used by Bunea, Tsybakov, and Wegkamp [33]. We begin by a preliminary lemma.

Lemma 5.3. *Let us consider the regression model* (5.1). *Let* $\hat{\gamma}$ *be a solution of Program* (5.8). *Let us assume that* $\Lambda_{n,p}$, *the event defined in Lemma 5.2, is satisfied. Then*

$$\left\|Z(\hat{\beta} - \beta^*)\right\|_2^2 + s\left\|\Xi(\hat{\gamma} - \gamma^*)\right\|_1 \leq 4s\sum_{j:\gamma_j^*\neq 0}\xi_j^{\frac{1}{2}}\left|\hat{\gamma}_j - \gamma_j^*\right|. \tag{5.24}$$

Proof of Lemma 5.3. Let us remember the criterion (5.8):

$$\operatorname*{argmin}_{\gamma\in\mathbb{R}^p}\left\{\|Y - X\gamma\|_2^2 + 2s\left\|\Xi\gamma\right\|_1 + \gamma' M\gamma\right\}$$

with $Y = X\beta^* + \varepsilon$ and M is given by (5.12). First, let us prove that $X\beta^* = XW(X'X)\gamma^*$, we start from the relation $(Z'Z)\beta^* = (X'X)\gamma^* = (Z'Z)W(X'X)\gamma^*$ so $\beta^* - W(X'X)\gamma^* \in \ker(Z'Z) = \ker Z = \ker X$ and then $X\beta^* = XW(X'X)\gamma^*$. Therefore we have

$$\begin{aligned}\|Y - X\gamma\|_2^2 &= \|XW(X'X)\gamma^* - X\gamma + \varepsilon\|_2^2 = \|X[W(X'X) - I_p]\gamma^* + X(\gamma^* - \gamma) + \varepsilon\|_2^2\\ &= (\gamma^*)'[(X'X)W - I_p](X'X)[W(X'X) - Ip]\gamma^* + \|X\gamma^* - X\gamma\|_2^2\\ &\quad + \|\varepsilon\|_2^2 + 2\Big\{(\gamma^*)'[(X'X)W - I_p](X'X)(\gamma^* - \gamma)\\ &\qquad + \varepsilon' X[W(X'X) - I_p)\gamma^* + \varepsilon' X(\gamma^* - \gamma)\Big\}.\end{aligned}$$

Then, since $M = (X'X)W(X'X) - (X'X)$, we have

$$\begin{aligned}&\operatorname*{argmin}_{\gamma\in\mathbb{R}^p}\left\{\|Y - X\gamma\|_2^2 + 2s\left\|\Xi\gamma\right\|_1 + \gamma' M\gamma\right\}\\ &\qquad = \operatorname*{argmin}_{\gamma\in\mathbb{R}^p}\left\{\|X\gamma^* - X\gamma\|_2^2 + 2s\left\|\Xi\gamma\right\|_1 + \gamma' M\gamma - 2\varepsilon' X\gamma + 2(\gamma^*)' M(\gamma^* - \gamma)\right\}.\end{aligned}$$

Using now the definition of $\hat{\gamma}$ (it minimizes the above quantity) we obtain

$$\begin{aligned}\|X(\hat{\gamma}-\gamma^*)\|_2^2 &\leq 2s\sum_{j=1}^{p}\xi_j^{\frac{1}{2}}\left(|\gamma_j^*|-|\hat{\gamma}_j|\right)+2\sum_{i=1}^{n}\varepsilon_i x_i(\hat{\gamma}-\gamma^*)\\ &\quad+\Big[(\gamma^*)'M\gamma^*-\hat{\gamma}'M\hat{\gamma}-2(\gamma^*)'M(\gamma^*-\hat{\gamma})\Big]\\ &\leq 2s\sum_{j=1}^{p}\xi_j^{\frac{1}{2}}\left(|\gamma_j^*|-|\hat{\gamma}_j|\right)+2\sum_{i=1}^{n}\varepsilon_i x_i(\hat{\gamma}-\gamma^*)-(\gamma^*-\hat{\gamma})'M(\gamma^*-\hat{\gamma}).\end{aligned}$$

As a consequence, replacing M by its definition we obtain

$$(\gamma^*-\hat{\gamma})'(X'X)W(X'X)(\gamma^*-\hat{\gamma})\leq 2s\sum_{j=1}^{p}\xi_j^{\frac{1}{2}}\left(|\gamma_j^*|-|\hat{\gamma}_j|\right)+2\sum_{i=1}^{n}\varepsilon_i x_i(\hat{\gamma}-\gamma^*).$$

Note that

$$\begin{aligned}(\gamma^*-\hat{\gamma})'(X'X)W(X'X)(\gamma^*-\hat{\gamma}) &= (\beta^*-\hat{\beta})'(Z'Z)W(X'X)(\gamma^*-\hat{\gamma})\\ &= (\beta^*-\hat{\beta})'(X'X)(\gamma^*-\hat{\gamma})=(\beta^*-\hat{\beta})'(Z'Z)(\beta^*-\hat{\beta}),\end{aligned}\quad(5.25)$$

then our bound so far is

$$\left\|Z(\hat{\beta}-\beta^*)\right\|_2^2\leq 2s\sum_{j=1}^{p}\xi_j^{\frac{1}{2}}\left(|\gamma_j^*|-|\hat{\gamma}_j|\right)+2\sum_{i=1}^{n}\varepsilon_i x_i(\hat{\gamma}-\gamma^*).\quad(5.26)$$

Moreover, on the event $\Lambda_{n,p}$, we have

$$2\sum_{i=1}^{n}\varepsilon_i x_i\left(\hat{\gamma}-\gamma^*\right)=2\sum_{j=1}^{p}V_j\left(\hat{\gamma}_j-\gamma_j^*\right)\leq\sum_{j=1}^{p}s\xi_j^{\frac{1}{2}}\left|\hat{\gamma}_j-\gamma_j^*\right|.\quad(5.27)$$

It follows from (5.26) and (5.27) that

$$\begin{aligned}\left\|Z(\hat{\beta}-\beta^*)\right\|_2^2+s\sum_{j=1}^{p}\xi_j^{\frac{1}{2}}\left|\hat{\gamma}_j-\gamma_j^*\right| &\leq 2s\sum_{j=1}^{p}\xi_j^{\frac{1}{2}}\left|\hat{\gamma}_j-\gamma_j^*\right|+2s\sum_{j=1}^{p}\xi_j^{\frac{1}{2}}|\gamma_j^*|-2s\sum_{j=1}^{p}\xi_j^{\frac{1}{2}}|\hat{\gamma}_j|\\ &\leq 2s\sum_{j:\gamma_j^*\neq 0}\xi_j^{\frac{1}{2}}\left|\hat{\gamma}_j-\gamma_j^*\right|+2s\sum_{\gamma_j^*\neq 0}\xi_j^{\frac{1}{2}}\left(|\gamma_j^*|-|\hat{\gamma}_j|\right)\\ &\leq 4s\sum_{j:\gamma_j^*\neq 0}\xi_j^{\frac{1}{2}}\left|\hat{\gamma}_j-\gamma_j^*\right|.\end{aligned}$$

This is the result claimed in the lemma. □ □

We are now ready to give the

Proof of Theorem 5.2. We apply Lemmas 5.2 and 5.3 and state that Inequality (5.24)

$$\left\|Z(\hat{\beta}-\beta^*)\right\|_2^2+s\sum_{j=1}^{p}\xi_j^{\frac{1}{2}}\left|\hat{\gamma}_j-\gamma_j^*\right|\leq 4s\sum_{j:\gamma_j^*\neq 0}\xi_j^{\frac{1}{2}}\left|\hat{\gamma}_j-\gamma_j^*\right|.$$

holds with probability at least $1 - p^{1-\frac{\kappa^2}{8}}$. This equation implies in particular that

$$\sum_{j:\gamma_j^*=0} \xi_j^{\frac{1}{2}} \left|\hat{\gamma}_j - \gamma_j^*\right| \leq 3 \sum_{j:\gamma_j^*\neq 0} \xi_j^{\frac{1}{2}} \left|\hat{\gamma}_j - \gamma_j^*\right|,$$

then taking $\alpha = \hat{\gamma} - \gamma^*$ in Assumption (A1), we obtain

$$\begin{aligned}\left\|Z(\hat{\beta} - \beta^*)\right\|_2^2 + s\sum_{j=1}^{p} \xi_j^{\frac{1}{2}} \left|\hat{\gamma}_j - \gamma_j^*\right| &\leq 4s\sqrt{\left(\sum_{\gamma_j^*\neq 0} \xi_j\right)\left(\sum_{\gamma_j^*\neq 0} \left(\hat{\gamma}_j - \gamma_j^*\right)^2\right)} \\ &\leq 4s\sqrt{\left(\sum_{\gamma_j^*\neq 0} \xi_j\right) \frac{c}{n}(\hat{\gamma} - \gamma^*)'\Omega(\hat{\gamma} - \gamma^*)} = 4s\sqrt{\left(\sum_{\gamma_j^*\neq 0} \xi_j\right) \frac{c}{n}\|Z(\hat{\beta} - \beta^*)\|_2^2},\end{aligned}$$

where we used (5.25) in the last equality. As a consequence,

$$\left\|Z(\hat{\beta} - \beta^*)\right\|_2 \leq 4s\sqrt{\frac{c}{n}\sum_{\gamma_j^*\neq 0} \xi_j}$$

and so

$$\left\|Z(\hat{\beta} - \beta^*)\right\|_2^2 \leq \frac{16s^2 c}{n}\sum_{\gamma_j^*\neq 0} \xi_j = 16c\kappa^2\sigma^2 \log(p)\sum_{\gamma_j^*\neq 0} \xi_j$$

while

$$s\left\|\Xi(\hat{\gamma} - \gamma^*)\right\|_1 \leq 4s\sqrt{\frac{c}{n}\|Z(\hat{\beta} - \beta^*)\|_2^2 \sum_{\gamma_j^*\neq 0} \xi_j}$$

which implies that

$$\left\|\Xi(\hat{\gamma} - \gamma^*)\right\|_1 \leq 4\sqrt{\frac{16c^2\kappa^2\sigma^2 \log(p)}{n}}\sum_{\gamma_j^*\neq 0} \xi_j.$$

This ends the proof. □ □

5.6.6 Proof of Theorem 5.4

First, we give the following lemma.

Lemma 5.4. *We have, on the event $\Lambda_{n,p}$,*

$$\|Z(\tilde{\beta} - \beta^*)\|_2^2 \leq \frac{3s}{2}\left\|\Xi P\left(\tilde{\beta} - \beta^{**}\right)\right\|_1 \leq 3s\sum_{P_j\beta^{**}\neq 0} \xi_j^{\frac{1}{2}} \left|P_j\left(\tilde{\beta} - \beta^{**}\right)\right|. \tag{5.28}$$

Proof of Lemma 5.4. We have

$$\begin{aligned}\|Z(\tilde{\beta} - \beta^*)\|_2^2 &= \|Z(\tilde{\beta} - \beta^{**})\|_2^2 = (\tilde{\beta} - \beta^{**})'(Z'Z)(\tilde{\beta} - \beta^{**}) \\ &= [P(\tilde{\beta} - \beta^{**})]'(X'X)(\tilde{\beta} - \beta^{**}) = [\Xi P(\tilde{\beta} - \beta^{**})]'\Xi^{-1}(X'X)(\tilde{\beta} - \beta^{**}) \\ &\leq \|\Xi P(\tilde{\beta} - \beta^{**})\|_1 \|\Xi^{-1}(X'X)(\tilde{\beta} - \beta^{**})\|_\infty \leq \frac{3s}{2}\|\Xi P(\tilde{\beta} - \beta^{**})\|_1, \end{aligned} \tag{5.29}$$

where we use the Dantzig constraint in the last inequality. Note that, by definition of $\tilde{\beta}$,

$$0 \leq \|\Xi P\beta^{**}\|_1 - \|\Xi P\tilde{\beta}\|_1 = \sum_{P_j\beta^{**}\neq 0} \xi_j^{\frac{1}{2}} |P_j\beta^{**}| - \sum_{P_j\beta^{**}\neq 0} \xi_j^{\frac{1}{2}} \left|P_j\tilde{\beta}\right| - \sum_{P_j\beta^{**}=0} \xi_j^{\frac{1}{2}} \left|P_j\tilde{\beta}\right|$$
$$\leq \sum_{P_j\beta^{**}\neq 0} \xi_j^{\frac{1}{2}} \left|P_j\beta^{**} - P_j\tilde{\beta}\right| - \sum_{P_j\beta^{**}=0} \xi_j^{\frac{1}{2}} \left|P_j\beta^{**} - P_j\tilde{\beta}\right|,$$

that leads to Inequality (5.28). □ □

Proof of Theorem 5.4. We apply here Lemmas 5.2 and 5.4 and we obtain that with probability at least $1 - p^{1-\frac{\kappa^2}{8}}$, we have Inequality (5.28). Now, let us remark that

$$\begin{aligned}
\|Z(\beta^* - \tilde{\beta})\|_2^2 &\leq \frac{3s}{2}\|\Xi P\left(\beta^{**} - \tilde{\beta}\right)\|_1 \leq 3s \sum_{P_j\beta^{**}\neq 0} \xi_j^{\frac{1}{2}} \left|P_j\beta^{**} - P_j\tilde{\beta}\right| \\
&\leq 3s \sqrt{\left(\sum_{P_j\beta^{**}\neq 0} \xi_j\right)\left(\sum_{P_j\beta^{**}\neq 0} \left|P_j\beta^{**} - P_j\tilde{\beta}\right|^2\right)} \\
&\leq 3s \left(\sum_{P_j\beta^{**}\neq 0} \xi_j\right)^{\frac{1}{2}} \sqrt{\frac{c}{n}\|Z(\beta^* - \tilde{\beta})\|_2^2}.
\end{aligned}$$

So we have,

$$\|Z(\tilde{\beta} - \beta^*)\|_2^2 \leq 9s^2\frac{c}{n} \sum_{P_j\beta^{**}\neq 0} \xi_j,$$

and as a consequence

$$\frac{3s}{2}\left\|\Xi P\left(\beta^{**} - \tilde{\beta}\right)\right\|_1 \leq 3s \left(\sum_{P_j\beta^{**}\neq 0} \xi_j\right)^{\frac{1}{2}} \sqrt{\frac{c}{n}\|Z(\beta^* - \tilde{\beta})\|_2^2} \leq 9s^2\frac{c}{n} \sum_{P_j\beta^{**}\neq 0} \xi_j,$$

this ends the proof. □ □

5.6.7 Proof of Theorems 5.3 and 5.5

Let us remind that Assumption (A2), involved in both theorems, implies among others that $\xi_j = 1$ for any j, so Ξ is the identity matrix.

Proof of Theorem 5.3. We can rewrite the fact that $\hat{\beta} = P\hat{\gamma}$ satisfies the Dantzig constraint:

$$\left\|\Omega(\hat{\gamma} - \gamma^*) - \frac{X'\varepsilon}{n}\right\|_\infty \leq \frac{s}{n}. \tag{5.30}$$

Recall that $\Lambda_{n,p} = \{\max_{j=1,\dots,p} 2|V_j| \leq s\}$ with $V_j = X_j'\varepsilon$, then applying (5.30), we have on $\Lambda_{n,p}$ and for any $j \in \{1,\dots,p\}$,

$$\begin{aligned}
\left|\Omega_{j,j}(\hat{\gamma}_j - \gamma_j^*)\right| &= \left|\{\Omega(\hat{\gamma} - \gamma^*)\}_j - \sum_{\substack{k=1 \\ k\neq j}}^{p} \Omega_{j,k}(\hat{\gamma}_k - \gamma_k^*)\right| \\
&\leq \frac{s}{n} + \left|\frac{X_j'\varepsilon}{n}\right| + \sum_{\substack{k=1 \\ k\neq j}}^{p} |\Omega_{j,k}(\hat{\gamma}_k - \gamma_k^*)| \\
&\leq \frac{3s}{2n} + \sum_{\substack{k=1 \\ k\neq j}}^{p} |\Omega_{j,k}(\hat{\gamma}_k - \gamma_k^*)| \leq \frac{3s}{2n} + \|\hat{\gamma} - \gamma^*\|_1 \left(\sup_{k\neq j} |\Omega_{j,k}|\right)
\end{aligned}$$

which implies that

$$\|\hat{\gamma} - \gamma^*\|_\infty \leq \frac{1}{\inf_j \Omega_{j,j}} \left[\frac{3s}{2n} + \|\hat{\gamma} - \gamma^*\|_1 \left(\sup_j \sup_{k\neq j} |\Omega_{j,k}|\right)\right]. \tag{5.31}$$

Now, remind that $\sup_j \sup_{k\neq j} |\Omega_{j,k}| = \rho$ is upper bounded by Assumption (A2). Moreover, Assumption (A2) implies, by Lemma 5.1, that Assumption (A1) is satisfied with $c = 2/(\inf_j \Omega_{j,j})$, so we can apply Theorem 5.2 to upper bound $\|\hat{\gamma} - \gamma^*\|_1$. This leads to

$$\|\hat{\gamma} - \gamma^*\|_\infty \leq \frac{1}{\inf_j \Omega_{j,j}} \left[\frac{3s}{2n} + 16c\kappa\sigma\|\gamma^*\|_0\sqrt{\frac{\log(p)}{n}} \times \frac{\inf_j \Omega_{j,j}}{14\|\gamma^*\|_0}\right]$$

writing $c = 2/(\inf_j \Omega_{j,j})$ and $s = \kappa\sigma\sqrt{n\log(p)}$ we obtain:

$$\|\hat{\gamma} - \gamma^*\|_\infty \leq \frac{3\kappa\sigma}{\inf_j \Omega_{j,j}}\sqrt{\frac{\log(p)}{n}},$$

that is the inequality stated in Theorem 5.3. □ □

Proof of Theorem 5.5. The preceding proof is also valid for Theorem 5.5. The only difference is that, at Inequality (5.31), upper bound the l_1 norm using Theorem 5.4 instead of Theorem 5.2. This replaces the constant 16 by a 6. □ □

Chapter 6

Transductive versions of the LASSO and the Dantzig Selector

Abstract: We consider the linear regression problem, where the number p of covariates is possibly larger than the number n of observations $(x_i, y_i)_{i \leq i \leq n}$, under sparsity assumptions. On the one hand, several methods have been successfully proposed to perform this task, for example the LASSO introduced by Tibshirani [135] or the Dantzig Selector introduced by Candès and Tao [35]. On the other hand, consider new values $(x_i)_{n+1 \leq i \leq m}$. If one wants to estimate the corresponding y_i's, one should think of a specific estimator devoted to this task, referred by Vapnik [145] as a "transductive" estimator. This estimator may differ from an estimator designed to the more general task "estimate on the whole domain". In this work, we propose a generalized version both of the LASSO and the Dantzig Selector, based on the geometrical remarks about the LASSO by Alquier [4] and Alquier and Hebiri [7]. The "usual" LASSO and Dantzig Selector, as well as new estimators interpreted as transductive versions of the LASSO, appear as special cases. These estimators are interesting at least from a theoretical point of view: we can give theoretical guarantees for these estimators under hypotheses that are relaxed versions of the hypotheses required in the papers about the "usual" LASSO. These estimators can also be efficiently computed, with results comparable to the ones of the LASSO.

6.1 Introduction

In many modern applications, a statistician often have to deal with very large datasets. Regression problems may involve a large number of covariates p, possibly larger than the sample size n. In this situation, a major issue is dimension reduction, which can be performed through the selection of a small amount of relevant covariates. For this purpose, numerous regression methods have been proposed in the literature, ranging from the classical information criteria such as AIC (Akaike [2]) and BIC (Schwartz [127]) to the more recent sparse methods, known as the LASSO (Tibshirani [135]), and the Dantzig Selector (Candès and Tao [35]). Regularized regression methods have recently witnessed several developments due to the attractive feature of computational feasibility, even for high dimensional data (i.e., when the number of covariates p is large). We focus on the usual linear regression model:

$$y_i = x_i\beta^* + \varepsilon_i, \qquad i = 1, \dots, n, \tag{6.1}$$

where the design $x_i = (x_{i,1}, \dots, x_{i,p}) \in \mathbb{R}^p$ is deterministic, $\beta^* = (\beta_1^*, \dots, \beta_p^*)' \in \mathbb{R}^p$ is the unknown parameter and $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. centered Gaussian random variables with known variance σ^2. Let X denote the matrix with i-th line equal to x_i, and let X_j denote its j-th column, with $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, p\}$. So:

$$X = (x_1', \dots, x_n')' = (X_1, \dots, X_p).$$

For the sake of simplicity, we will assume that the observations are normalized in such a way that $X_j'X_j/n = 1$. We denote by Y the vector $Y = (y_1, \dots, y_n)'$.
For all $\alpha \leq 1$ and any vector $v \in \mathbb{R}^d$, we set $\|\cdot\|_\alpha$, the norm: $\|v\|_\alpha = (|v_1|^\alpha + \ldots + |v_d|^\alpha)^{1/\alpha}$. In particular $\|\cdot\|_2$ is the euclidean norm. Moreover for all $d \in \mathbb{N}$, we use the notation $\|v\|_0 = \sum_{i=1}^d \mathbb{1}(v_i \neq 0)$.

The problem of estimating the regression parameter in the high dimensional setting have been extensively studied in the statistical literature. Among others, the LASSO introduced by Tibshirani [135] (denoted by $\hat{\beta}^L$), the Dantzig Selector of Candès and Tao [35] (denoted by $\hat{\beta}^{DS}$) and the non-negative garrote (by Yuan and Lin [160], denoted by $\hat{\beta}^{NNG}$) have been proposed to deal with this problem for a large p, even for $p > n$. These estimators give very good practical results. For instance in the paper by Tibshirani [135], simulations and tests on real data have been provided for the LASSO. We also refer to the papers by Koltchinskii [91, 92], Meier, van de Geer, and Bühlmann [110], van de Geer [143], Dalalyan and Tsybakov [46] and Chesneau and Hebiri [44] for related work with different estimators: non-quadratic loss, penalties slightly different from ℓ_1 and random design.
From a theoretical point of view, Sparsity Inequalities (SI) have been proved for these estimators under different assumptions. That is upper bounds of order of $\mathcal{O}\left(\sigma^2\|\beta^*\|_0 \log(p)/n\right)$

for the errors $(1/n)\|X\hat{\beta} - X\beta^*\|_2^2$ and $\|\hat{\beta} - \beta^*\|_2^2$ have been derived, where $\hat{\beta}$ is one of the estimators mentioned above. In particular these bounds involve the number of non-zero coordinates in β^* (multiplied by $\log(p)$), instead of dimension p. Such bounds garanty that under some assumptions, $X\hat{\beta}$ and $\hat{\beta}$ are good estimators of $X\beta^*$ and β^* respectively. According to the LASSO $\hat{\beta}^L$, these SI are given for example by Bunea, Tsybakov, and Wegkamp [32] and Bickel, Ritov, and Tsybakov [17], whereas Candès and Tao [35] and Bickel, Ritov, and Tsybakov [17] provided SI for the Dantzig Selector $\hat{\beta}^{DS}$. On the other hand, Bunea [28] establishes conditions which ensure $\hat{\beta}^L$ and β^* have the same null coordinates. Analog results for $\hat{\beta}^{DS}$ can be found in the paper by Lounici [101].

Now, let us assume that we are given additional observations $x_i \in \mathbb{R}^p$ for $n+1 \leq i \leq m$ (with $m > n$), and introduce the matrix $Z = (x_1', \ldots, x_m')'$. Assume that the objective of the statistician is precisely to estimate $Z\beta^*$: namely, he cares about predicting what would be the labels attached to the additional x_i's. It is argued by Vapnik [145] that in such a case, a specific estimator devoted to this task should be considered: the transductive estimator. This estimator differs from an estimator tailored for the estimation of β^* or $X\beta^*$ like the LASSO. Indeed one usually builds an estimator $\hat{\beta}(X, Y)$ and then computes $Z\hat{\beta}(X, Y)$ to estimate $Z\beta^*$. The approach taken here is to consider estimators $\hat{\beta}(X, Y, Z)$ exploiting the knowledge of Z, and then to compute $Z\hat{\beta}(X, Y, Z)$.

Some methods in supervised classification or regression were successfully extended to the transductive setting, such as the well-known Support Vector Machines (SVM) by Vapnik [145], the Gibbs estimators by Catoni [39]. It is argued in the semi-supervised learning literature (see for example the paper by Chapelle, Schölkopf, and Zien [41] for a recent survey) that taking into account the information on the design given by the new additional x_i's has a stabilizing effect on the estimator.

In this chapter, we study a family of estimators which generalizes the LASSO and the Dantzig Selector. The considered family depends on a $q \times p$ matrix A, with $q \in \mathbb{N}$, whose choice allows to adapt the estimator to the objective of the statistician. The choice of the matrix A allows to cover transductive setting.

The rest of the chapter is organized as follows. In the next section, we motivate the use of the studied family of estimators through geometrical considerations stated by Alquier and Hebiri [7]. In Sections 6.3 and 6.4, we establish Sparsity Inequalities for these estimators. A discussion on the assumptions needed to prove the SI is also provided. In particular, it is shown that the estimators devoted to the transductive setting satisfy these SI with weaker assumptions that those needed by the LASSO or the Dantzig Selector, when $m > p > n$. That is, when the number of news points is large enough. The implementation of our estimators and some numerical experiments are the purpose of Section 6.5. The

results clearly show that the use of a transductive version of the LASSO may improve the performance of the estimation. All proofs of the theoretical results are postponed to Section 6.7.

6.2 Preliminaries

In this section we state geometrical considerations (projections on a confidence region) for the LASSO and the Dantzig Selector. These motivate the introduction of our estimators. Finally we discuss the different objectives considered in this chapter.

Let us remind that a definition of the LASSO estimate is given by

$$\arg\min_{\beta\in\mathbb{R}^p}\left\{\|Y-X\beta\|_2^2+2\lambda\|\beta\|_1\right\}. \tag{6.2}$$

A dual form (Osborne, Presnell, and Turlach [118]) of this program is also of interest:

$$\begin{cases} \arg\min_{\beta\in\mathbb{R}^p}\|X\beta\|_2^2 \\ \\ s.t.\,\|X'(Y-X\beta)\|_\infty\leq\lambda; \end{cases} \tag{6.3}$$

actually it is proved by Alquier [4] that any solution of Program 6.3 is a solution of Program 6.2 and that the set $\{X\beta\}$ is the same where β is taken among all the solutions of Program 6.2 or among all the solutions of 6.3. So both programs are equivalent in terms of estimating $X\beta^*$.

Now, let us remind the definition of the Dantzig Selector:

$$\begin{cases} \arg\min_{\beta\in\mathbb{R}^p}\|\beta\|_1 \\ \\ s.t.\,\|X'(Y-X\beta)\|_\infty\leq\lambda. \end{cases} \tag{6.4}$$

Alquier [4] observed that both Programs 6.3 and 6.4 can be seen as a projection of the null vector $\mathbf{0_p}$ onto the region $\{\beta:\|X'(Y-X\beta)\|_\infty\leq\lambda\}$ that can be interpreted as a confidence region, with confidence $1-\eta$, for a given λ that depends on η (see Lemma 6.1 here for example). The difference between the two programs is the distance (or semi-distance) used for the projection.

Based on these geometrical considerations, we proposed by Alquier and Hebiri [7] to study the following transductive estimator:

$$\begin{cases} \arg\min_{\beta\in\mathbb{R}^p}\|Z\beta\|_2^2 \\ \\ s.t.\,\|X'(Y-X\beta)\|_\infty\leq\lambda; \end{cases} \tag{6.5}$$

that is a projection on the same confidence region, but using a distance adapted to the transductive estimation problem. We proved a Sparsity Inequality for this estimator exploiting a novel sparsity measure.

In this chapter, we propose a generalized version of the LASSO and of the Dantzig Selector, based on the same geometrical remark. More precisely for $q \in \mathbb{N}^*$, let A be a $q \times p$ matrix. We propose two general estimators, $\hat{\beta}_{A,\lambda}$ (extension of the LASSO, based on a generalization of Program 6.2) and $\tilde{\beta}_{A,\lambda}$ (transductive Dantzig Selector, generalization of Program 6.4). These novel estimators depend on two tuning parameters: $\lambda > 0$ is a regularization parameter, it plays the same role as the tuning parameter involved in the LASSO, and the matrix A that will allow to adapt the estimator to the objective of the statistician. More particularly, depending on the choice of the matrix A, this estimator can be adapted to one of the following objectives:

- **denoising objective:** the estimation of $X\beta^*$, that is a denoised version of Y. For this purpose, we consider the estimator $\hat{\beta}_{A,\lambda}$, with $A = X$. In this case, the estimator will actually be equal to the LASSO $\hat{\beta}^L$ and $\tilde{\beta}_{A,\lambda}$, with the same choice $A = X$ will be equal to the Dantzig Selector;

- **transductive objective:** the estimation of $Z\beta^*$, by $\hat{\beta}_{A,\lambda}$ or $\tilde{\beta}_{A,\lambda}$, with $A = \sqrt{n/m}Z$. We will refer the corresponding estimators as the "Transductive LASSO" and "Transductive Dantzig Selector";

- **estimation objective:** the estimation of β^* itself, by $\hat{\beta}_{A,\lambda}$, with $A = \sqrt{n}I$. In this case, it appears that both estimators are well defined only in the case $p < n$ and are equal to a soft-thresholded version of the usual least-square estimator.

For both estimators and all the above objectives, we prove SI (Sparsity Inequalities). Moreover, we show that these estimators can easily be computed.

6.3 The "easy case": $\mathrm{Ker}(X) = \mathrm{Ker}(Z)$

In this section, we deal with the "easy case", where $\mathrm{Ker}(A) = \mathrm{Ker}(X)$ (think of $A = X$, $A = \sqrt{n}I$ or $A = \sqrt{n/m}Z$). This setting is natural at least in the case $p < n$ where both kernels are equal to $\{0\}$ in general. We provide SI (Sparsity Inequality, Theorem 6.1) for the studied estimators, based on the techniques developed by Bickel, Ritov, and Tsybakov [17].

6.3.1 Definition of the estimators

Definition 6.1. *For a given parameter $\lambda \geq 0$ and any matrix A such that $Ker(A) = Ker(X)$, we consider the estimator given by*

$$\hat{\beta}_{A,\lambda} \in \arg\min_{\beta \in \mathbb{R}^p} \Big\{ -2Y'X(\widetilde{X'X})^{-1}(A'A)\beta + \beta'(A'A)\beta + 2\lambda\|\Xi_A\beta\|_1 \Big\},$$

where $(\widetilde{X'X})^{-1}$ *is exactly* $(X'X)^{-1}$ *if* $(X'X)$ *is invertible, and any pseudo-inverse of this matrix otherwise, and where* Ξ_A *is a diagonal matrix whose* (j,j)*-th coefficient is* $\xi_j^{\frac{1}{2}}(A)$ *with* $\xi_j(A) = \frac{1}{n}[(A'A)(\widetilde{X'X})^{-1}(A'A)]_{j,j}$.

Remark 6.1. *Equivalently we have*

$$\hat{\beta}_{A,\lambda} \in \arg\min_{\beta\in\mathbb{R}^p}\Big\{\Big\|\tilde{Y}_A - A\beta\Big\|_2^2 + 2\lambda\|\Xi_A\beta\|_1\Big\},$$

where $\tilde{Y}_A = A(\widetilde{X'X})^{-1}X'Y$.

Actually, we are going to consider three particular cases of this estimator in this work, depending on the objective of the statistician:

- **denoising objective:** the LASSO, denoted here by $\hat{\beta}_{X,\lambda}$, given by

$$\hat{\beta}_{X,\lambda} \in \arg\min_{\beta\in\mathbb{R}^p}\Big\{\|Y - X\beta\|_2^2 + 2\lambda\|\beta\|_1\Big\}$$
$$= \arg\min_{\beta\in\mathbb{R}^p}\big\{-2Y'X\beta + \beta'X'X\beta + 2\lambda\|\beta\|_1\big\}$$

  (note that in this case, $\Xi_X = I$ since X is normalized);

- **transductive objective:** the Transductive LASSO, denoted here by $\hat{\beta}_{\sqrt{n/m}Z,\lambda}$, given by

$$\hat{\beta}_{\sqrt{\frac{n}{m}}Z,\lambda} \in \arg\min_{\beta\in\mathbb{R}^p}\Big\{\frac{n}{m}\Big\|\tilde{Y}_Z - Z\beta\Big\|_2^2 + 2\lambda\|\Xi_{\frac{n}{m}Z'Z}\beta\|_1\Big\};$$

- **estimation objective:** $\hat{\beta}_{\sqrt{n}I,\lambda}$, defined by

$$\hat{\beta}_{\sqrt{n}I,\lambda} \in \arg\min_{\beta\in\mathbb{R}^p}\Big\{n\Big\|\tilde{Y}_I - \beta\Big\|_2^2 + 2\lambda\|\Xi_{\sqrt{n}I}\beta\|_1\Big\}.$$

Let us give the analogous definition for an extension of the Dantzig Selector.

Definition 6.2. *For a given parameter* $\lambda > 0$ *and any matrix* A *such that* $Ker(A) = Ker(X)$*, we consider the estimator given by*

$$\tilde{\beta}_{A,\lambda} = \begin{cases} \arg\min_{\beta\in\mathbb{R}^p}\|\beta\|_1 \\ \\ s.t.\ \Big\|\Xi_A^{-1}A'A((\widetilde{X'X})^{-1}X'Y - \beta)\Big\|_\infty \leq \lambda. \end{cases} \tag{6.6}$$

Here again, we are going to consider three cases, for $A = X$, $A = \sqrt{n/m}Z$ and $A = \sqrt{n}I$, and it is easy to check that for $A = X$ we have exactly the usual definition of the Dantzig Selector (Program 6.4). Moreover, here again, note that we can rewrite this estimator:

$$\tilde{\beta}_{A,\lambda} = \begin{cases} \arg\min_{\beta\in\mathbb{R}^p}\|\beta\|_1 \\ \\ s.t.\ \Big\|\Xi_A^{-1}A'(\tilde{Y}_A - A\beta)\Big\|_\infty \leq \lambda. \end{cases}$$

The following proposition provides an interpretation of our estimators when $A = \sqrt{n}I$.

Proposition 6.1. *Let us assume that $(X'X)$ is invertible. Then $\hat{\beta}_{\sqrt{n}I,\lambda} = \tilde{\beta}_{\sqrt{n}I,\lambda}$ and this is a soft-thresholded least-square estimator: let us put $\hat{\beta}^{LSE} = (X'X)^{-1}X'Y$ then $\hat{\beta}_{\sqrt{n}I,\lambda}$ is the vector obtained by replacing the j-th coordinate $b_j = \hat{\beta}_j^{LSE}$ of $\hat{\beta}^{LSE}$ by $sgn(b_j)\left(|b_j| - \lambda\xi_j(nI)/n\right)_+$, where we use the standard notation $sgn(x) = +1$ if $x \geq 0$, $sgn(x) = -1$ if $x < 0$ and $(x)_+ = \max(x, 0)$.*

Proposition 6.2 deals with a dual definition of the estimator $\hat{\beta}_{A,\lambda}$.

Proposition 6.2. *When* $\mathrm{Ker}(A) = \mathrm{Ker}(X)$*, the solutions β of the following program:*

$$\begin{cases} \arg\min_{\beta\in\mathbb{R}^p} \|A\beta\|_2^2 \\ \\ s.t. \left\|\Xi_A^{-1}A'((\tilde{Y}_A - A\beta)\right\|_\infty \leq \lambda \end{cases}$$

all satisfy $X\beta = X\hat{\beta}_{A,\lambda}$ and $A\beta = A\hat{\beta}_{A,\lambda}$.

Proofs can be found in Section 6.7, page 150.

6.3.2 Theoretical results

Let us first introduce our main assumption. This assumption is stated with a given $p \times p$ matrix M and a given real number $x > 0$.

Assumption $H(M, x)$**:** there is a constant $c(M) > 0$ such that, for any $\alpha \in \mathbb{R}^p$ such that $\sum_{j:\beta_j^*=0} \xi_j(M)\,|\alpha_j| \leq x \sum_{j:\beta_j^*\neq 0} \xi_j(M)\,|\alpha_j|$ we have

$$\alpha' M\alpha \geq c(M)n \sum_{j:\beta_j^*\neq 0} \alpha_j^2. \tag{6.7}$$

First, let us explain briefly the meaning of this hypothesis. In the case, where M is invertible, the condition

$$\alpha' M\alpha \geq c(M)n \sum_{j:\beta_j^*\neq 0} \alpha_j^2$$

is always satisfied for any $\alpha \in \mathbb{R}^p$ with $c(M)$ larger than the smallest eigenvalue of M/n. However, for the LASSO, we have $M = (X'X)$ and M cannot be invertible if $p > n$. Even in this case, Assumption $H(M, x)$ may still be satisfied. Indeed, the assumption requires that Inequality (6.7) holds only for a small for a small subset of $\mathbb{R}^p$ determined by the condition $\sum_{j:\beta_j^*=0} \xi_j(M)\,|\alpha_j| \leq x \sum_{j:\beta_j^*\neq 0} \xi_j(M)\,|\alpha_j|$. For $M = (X'X)$, this assumption becomes exactly the one taken by Bunea, Tsybakov, and Wegkamp [32]. In that paper, the necessity of such an hypothesis is also discussed.

Theorem 6.1. *Let us assume that Assumption $H(A'A, 3)$ is satisfied and that $Ker(A) = Ker(X)$. Let us choose $0 < \eta < 1$ and $\lambda = 2\sigma\sqrt{2n\log(p/\eta)}$. With probability at least $1-\eta$ on the draw of Y, we have simultaneously*

$$\left\| A\left(\hat{\beta}_{A,\lambda} - \beta^*\right)\right\|_2^2 \leq \frac{72\sigma^2}{c(A'A)} \log\left(\frac{p}{\eta}\right) \sum_{j:\beta_j^* \neq 0} \xi_j(A),$$

and

$$\left\| \Xi_A\left(\hat{\beta}_{A,\lambda} - \beta^*\right)\right\|_1 \leq \frac{24\sqrt{2}\sigma}{c(A'A)} \left(\frac{\log(p/\eta)}{n}\right)^{\frac{1}{2}} \sum_{j:\beta_j^* \neq 0} \xi_j(A).$$

In particular, the first inequality gives

- if Assumption $H(X'X, 3)$ is satisfied, with probability at least $1-\eta$,

$$\frac{1}{n}\left\| X\left(\hat{\beta}_{X,\lambda} - \beta^*\right)\right\|_2^2 \leq \frac{72\sigma^2}{nc(X'X)} \|\beta^*\|_0 \log\left(\frac{p}{\eta}\right);$$

- if Assumption $H(\frac{n}{m}Z'Z, 3)$ is satisfied, and if $Ker(Z) = Ker(X)$, with probability at least $1-\eta$,

$$\frac{1}{m}\left\| Z\left(\hat{\beta}_{Z,\lambda} - \beta^*\right)\right\|_2^2 \leq \frac{72\sigma^2}{nc(\frac{n}{m}Z'Z)} \sum_{j:\beta_j^* \neq 0} \xi_j\left(\sqrt{n/m}Z\right) \log\left(\frac{p}{\eta}\right);$$

- and if $(X'X)$ is invertible, with probability at least $1-\eta$,

$$\left\| \hat{\beta}_{\sqrt{n}I,\lambda} - \beta^*\right\|_2^2 \leq \frac{72\sigma^2}{nc(nI)} \sum_{j:\beta_j^* \neq 0} \xi_j(nI) \log\left(\frac{p}{\eta}\right).$$

This result shows that each of these three estimators satisfy at least a SI for the task it is designed for. For example, the LASSO is proved to have "good" performance for the estimation of $X\beta^*$ and the Transductive LASSO is proved to have good performance for the estimation of $Z\beta^*$. However we cannot assert that, for example, the LASSO performs better than the Transductive LASSO for the estimation of $Z\beta^*$.

Remark 6.2. *For $A = X$, the particular case of our result applied to the LASSO is quite similar to the result given by Bunea, Tsybakov, and Wegkamp [32] on the LASSO. Actually, Theorem 6.1 can be seen as a generalization of the result by Bunea, Tsybakov, and Wegkamp [32] and it should be noted that the proof used to prove Theorem 6.1 uses arguments introduced by Bunea, Tsybakov, and Wegkamp [32].*

Remark 6.3. *As soon as $A'A$ is better determined than $X'X$, Assumption $H(A, x)$ is less restrictive than $H(X'X, x)$. In particular, in the case where $m > n$, Assumption $H((n/m)Z'Z, x)$ is expected to be less restrictive than Assumption $H(X'X, x)$.*

Now we give the analogous result for the estimator $\tilde{\beta}_{A,\lambda}$.

Theorem 6.2. *Let us assume that Assumption $H(A'A,1)$ is satisfied and that $Ker(A) = Ker(X)$. Let us choose $0 < \eta < 1$ and $\lambda = 2\sigma\sqrt{2n\log(p/\eta)}$. With probability at least $1-\eta$ on the draw of Y, we have simultaneously*

$$\left\|A\left(\tilde{\beta}_{A,\lambda} - \beta^*\right)\right\|_2^2 \leq \frac{72\sigma^2}{c(A'A)} \log\left(\frac{p}{\eta}\right) \sum_{j:\beta_j^* \neq 0} \xi_j(A),$$

and

$$\left\|\Xi_A\left(\tilde{\beta}_{A,\lambda} - \beta^*\right)\right\|_1 \leq \frac{12\sqrt{2}\sigma}{c(A'A)} \left(\frac{\log(p/\eta)}{n}\right)^{\frac{1}{2}} \sum_{j:\beta_j^* \neq 0} \xi_j(A).$$

6.4 An extension to the general case

In this section, we only deal with the transductive setting, $A = \sqrt{n/m}Z$. Let us remind that in such a framework, we observe X which consists of some observations x_i associated to labels Y_i in Y, for $i \in \{1,\dots,n\}$. Moreover we have additional observations x_i for $i \in \{n+1,\dots,m\}$ with $m > n$. We also recall that Z contains all the x_i for $i \in \{1,\dots,m\}$ and that the objective is to estimate the corresponding labels Y_i, let us put $\tilde{Y} = (Y_1,\dots,Y_m)'$.

6.4.1 General remarks

Let us have look at the definition of $\hat{\beta}_{\sqrt{n/m}Z,\lambda}$, for example as given in Remark 6.1:

$$\hat{\beta}_{\sqrt{\frac{n}{m}}Z,\lambda} \in \arg\min_{\beta\in\mathbb{R}^p}\left\{\frac{n}{m}\left\|\tilde{Y}_Z - Z\beta\right\|_2^2 + 2\lambda\|\Xi_{\frac{n}{m}Z'Z}\beta\|_1\right\},$$

where actually $\tilde{Y}_Z = Z\left(\widetilde{X'X}\right)^{-1}XY$ can be interpreted as a preliminary estimator of $\tilde{Y}$. Hence, in any case, we propose the following procedure.
Let us assume that, depending on the context, the user has a natural (and not necessary efficient) estimator of $\tilde{Y} = (Y_1,\dots,Y_{n+m})'$. Note this estimator $\check{Y}$.

Definition 6.3. *The Transductive LASSO is given by:*

$$\hat{\beta}_{\check{Y},\sqrt{\frac{n}{m}}Z,\lambda} \in \arg\min_{\beta\in\mathbb{R}^p}\left\{\frac{n}{m}\left\|\check{Y} - Z\beta\right\|_2^2 + 2\lambda\|\Xi_{\frac{n}{m}Z'Z}\beta\|_1\right\},$$

and the Transductive Dantzig Selector is defined as:

$$\tilde{\beta}_{\check{Y},\sqrt{\frac{n}{m}}Z,\lambda} = \begin{cases} \arg\min_{\beta\in\mathbb{R}^p}\|\beta\|_1 \\ \\ s.t. \left\|\frac{n}{m}\Xi^{-1}_{\sqrt{n/m}Z}Z'(\check{Y} - Z\beta)\right\|_\infty \leq \lambda. \end{cases}$$

In the next subsection, we propose a context where we have a natural estimator $\check{Y}$ and give a SI on this estimator.

6.4.2 An example: small labeled dataset, large unlabeled dataset

The idea of this example is to consider the case where the examples x_i for $1 \leq i \leq n$ are "representative" of the large populations x_i for $1 \leq i \leq m$.

Consider, $Z = (x_1', \ldots, x_m')'$ where the $x_i's$ are the points of interest: we want to estimate $\tilde{Y} = Z\beta^*$. However, we just have a very expensive and noisy procedure, that, given a point x_i, returns $Y_i = x_i\beta^* + \varepsilon_i$, where the ε_i's are $\mathcal{N}(0, \sigma^2)$ independent random variables. In such a case, the procedure cannot be applied for the whole dataset $Z = (x_1', \ldots, x_m')'$. We can only make a deal with a "representative" sample of size n. A typical case could be $n < p < m$.

First, let us introduce a slight modification of our main hypothesis. It is also stated with a given $p \times p$ matrix M and a given real number $x > 0$.

Assumption $H'(M, x)$**:** there is a $c(M) > 0$ such that, for any $\alpha \in \mathbb{R}^p$ such that $\sum_{j:\beta_j^*=0} |\alpha_j| \leq x \sum_{j:\beta_j^*\neq 0} |\alpha_j|$ we have

$$\alpha' M \alpha \geq c(M) n \sum_{j:\beta_j^*\neq 0} \alpha_j^2.$$

We can now state our main result.

Theorem 6.3. *Let us assume that Assumption $H'((n/m)Z'Z, 1)$ is satisfied. Let us choose $0 < \eta < 1$ and $\lambda_1 = \lambda_2 = 10^{-1}\sigma\sqrt{2n\log(p/\eta)}$. Moreover, let us assume that*

$$\forall u \in \mathbb{R}^p \text{ with } \|u\|_1 \leq \|\beta^*\|_1, \quad \left\| \left((X'X) - \frac{n}{m}(Z'Z) \right) u \right\|_\infty < \frac{\sigma}{10}\sqrt{2n \log\left(\frac{p}{\eta}\right)}. \tag{6.8}$$

Let $\check{Y}_{\lambda_1} = Z\tilde{\beta}_{X,\lambda_1}$ be a preliminary estimator of $\tilde{Y}$, based on ths Dantzig Selector given by (6.6) *(with $A = X$). Then define the Transductive LASSO by*

$$\hat{\beta}^*_{\frac{n}{m}Z,20\lambda_2} = \begin{cases} \arg\min_{\beta\in\mathbb{R}^p} \frac{n}{m}\|Z\beta\|_2^2 \\ \\ s.t. \left\|\frac{n}{m}Z'(\check{Y}_{\lambda_1} - Z\beta)\right\|_\infty \leq 20\lambda_2, \end{cases}$$

and the Transductive Dantzig Selector

$$\tilde{\beta}^*_{\frac{n}{m}Z,\lambda_2} = \begin{cases} \arg\min_{\beta\in\mathbb{R}^p} \|\beta\|_1 \\ \\ s.t. \left\|\frac{n}{m}Z'(\check{Y}_{\lambda_1} - Z\beta)\right\|_\infty \leq \lambda_2. \end{cases}$$

With probability at least $1 - \eta$ on the draw of Y, we have simultaneously

$$\frac{1}{m}\left\|Z(\tilde{\beta}^*_{\frac{n}{m}Z,\lambda_2} - \beta^*)\right\|_2^2 \leq \frac{16\sigma^2}{nc((n/m)Z'Z)} \log\left(\frac{p}{\eta}\right) \|\beta^*\|_0,$$

$$\left\|\tilde{\beta}^*_{\frac{n}{m}Z,\lambda_2} - \beta^*\right\|_1 \leq \frac{8\sigma}{c((n/m)Z'Z)} \left(\frac{\log(p/\eta)}{n}\right)^{\frac{1}{2}} \|\beta^*\|_0,$$

and moreover, if $H'((n/m)Z'Z, 5)$ *is also satisfied,*

$$\frac{1}{m}\left\|Z(\hat{\beta}^*_{\frac{n}{m}Z,20\lambda_2}-\beta^*)\right\|_2^2 \leq \frac{88\sigma^2}{nc((n/m)Z'Z)}\log\left(\frac{p}{\eta}\right)\|\beta^*\|_0,$$

$$\left\|\hat{\beta}^*_{\frac{n}{m}Z,20\lambda_2}-\beta^*\right\|_1 \leq \frac{54\sigma}{c((n/m)Z'Z)}\left(\frac{\log(p/\eta)}{n}\right)^{\frac{1}{2}}\|\beta^*\|_0.$$

First, let us remark that the preliminary estimator $\check{Y}_{\lambda_1}$ is defined using the Dantzig Selector $\tilde{\beta}_{X,\lambda_1}$. We could give exactly the same kind of results using a the LASSO $\hat{\beta}_{X,\lambda_1}$ as a preliminary estimator.

Now, let us give a look at the new hypothesis, Inequality (6.8). We can interpret this condition as the fact that the x_i's for $1 \leq i \leq n$ are effectively representative of the wide population: so $X'X/n$ is "not too far" from $Z'Z/m$. We will end this section by a result that proves that this is effectively the case in a typical situation.

Proposition 6.3. *Assume that* $m = kn$ *for an integer value* $k \in \mathbb{N}\setminus\{0,1\}$*. Let us assume that* X *and* Z *are build in the following way: we have a population* $\chi_1 = (\chi_{1,1},\ldots,\chi_{1,p}) \in \mathbb{R}^p, \ldots, \chi_m \in \mathbb{R}^p$ *(the points of interest). Then, we draw uniformly without replacement,* n *of the* χ_i*'s to be put in* X*: more formally, but equivalently, we draw uniformly a permutation* σ *of* $\{1,\ldots,m\}$ *and we put* $X = (x'_1,\ldots,x'_n)' = (\chi'_{\sigma(1)},\ldots,\chi'_{\sigma(n)})'$ *and* $Z = (x'_1,\ldots,x'_m)' = (\chi'_{\sigma(1)},\ldots,\chi'_{\sigma(m)})'$.
Let us assume that for any $(i,j) \in \{1,\ldots,m\}\times\{1,\ldots,p\}$, $\chi^2_{i,j} < \kappa$ *for some* $\kappa > 0$*, and that* $p \geq 2$*. Then, with probability at least* $1-\eta$*, for any* $u \in \mathbb{R}^p$*,*

$$\left\|\left(X'X - \frac{n}{m}Z'Z\right)u\right\|_\infty \leq \|u\|_1 \frac{2\kappa k}{k-1}\sqrt{2\log\frac{p}{\eta}}.$$

In particular, if we have

$$\|u\|_1 \leq \|\beta^*\|_1 \text{ and } \kappa \leq \frac{k-1}{10\,k}\frac{\sigma}{\|\beta^*\|_1}$$

then we have

$$\left\|\left(X'X - \frac{n}{m}Z'Z\right)u\right\|_\infty \leq \frac{\sigma}{10}\sqrt{2n\log\left(\frac{p}{\eta}\right)}.$$

Let us just mention that the assumption $m = kn$ is not restrictive. It has been introduced for the sake of simplicity.

6.5 Experimental results

Implementation. Since the paper by Tibshirani [135], several effective algorithms to compute the LASSO have been proposed and studied (for instance Interior Points methods by Kim, Koh, Lustig, Boyd, and Gorinevsky [88], LARS by Efron, Hastie, Johnstone,

and Tibshirani [61], Pathwise Coordinate Optimization by Friedman, Hastie, Höfling, and Tibshirani [67], Relaxed Greedy Algorithms by Huang, Cheang, and Barron [80]). For the Dantzig Selector, a linear method was proposed in the first paper by Candès and Tao [35]. The LARS algorithm was also successfully extended by James, Radchenko, and Lv [84] to compute the Dantzig Selector.

Then there are many algorithms to compute $\hat{\beta}_{A,\lambda}$ and $\tilde{\beta}_{A,\lambda}$, when $A = X$. Thanks to Proposition 6.1, it is also clear that we can easily find an efficient algorithm for the case $A = \sqrt{n}I$.

The general form of the estimators $\hat{\beta}_{A,\lambda}$ and $\tilde{\beta}_{A,\lambda}$ given by Definitions 6.1 and 6.2, allows to use one of the algorithms mentioned previously to compute our estimator in two cases. For example, from Remark 6.1, we have:

$$\hat{\beta}_{A,\lambda} \in \arg\min_{\beta\in\mathbb{R}^p}\left\{\left\|\tilde{Y}_A - A\beta\right\|_2^2 + 2\lambda\|\Xi_A\beta\|_1\right\},$$

then we just have to compute $\tilde{Y}_A$, to put $B = A\Xi_A^{-1}$, to use any program that computes the LASSO to determine

$$\hat{\gamma} \in \arg\min_{\gamma\in\mathbb{R}^p}\left\{\left\|\tilde{Y}_A - B\gamma\right\|_2^2 + 2\lambda\|\gamma\|_1\right\}$$

and then to put $\hat{\beta}_{A,\lambda} = \Xi_A^{-1}\gamma$.

In the rest of this section, we compare the LASSO and the transductive LASSO on the classical toy example introduced by Tibshirani [135] and used as a benchmark.

Data description. In the model proposed by Tibshirani, we have

$$Y_i = x_i\beta^* + \varepsilon_i$$

for $i \in \{1,\dots,n\}$, $\beta^* \in \mathbb{R}^p$ and the ε_i are i.i.d. $\mathcal{N}(0,\sigma^2)$. Finally, the $(x_i)_{i\in\{1,\dots,m\}}$ are generated from a probability distribution: they are independent and identically distributed

$$x_i \sim \mathcal{N}\left(\begin{pmatrix}0\\ \vdots\\ 0\end{pmatrix}, \begin{pmatrix}1 & \rho & \dots & \dots & \rho^{p-1}\\ \rho & 1 & \rho & \dots & \rho^{p-2}\\ \vdots & \ddots & \ddots & \ddots & \vdots\\ \rho^{p-2} & \dots & \rho & 1 & \rho\\ \rho^{p-1} & \dots & \dots & \rho & 1\end{pmatrix}\right),$$

for a given $\rho \in]-1,1[$.

As in the paper by Tibshirani [135], we set $p = 8$. In a first experiment, we take $(n,m) = (7,10)$, $\rho = 0.5$, $\sigma = 1$ and $\beta^* = (3, 1.5, 0, 0, 2, 0, 0, 0)$ ("sparse"). Then, in order to check the robustness of the results, we consider successively $\rho = 0.5$ by $\rho = 0.9$ (correlated variables), $\sigma = 1$ by $\sigma = 3$ (noisy case), $\beta^* = (3, 1.5, 0, 0, 2, 0, 0, 0)$ by $\beta^* = (5, 0, 0, 0, 0, 0, 0, 0)$ ("very

sparse" case), $(n,m) = (7,10)$ by $(n,m) = (7,20)$ (larger unlabeled set), $(n,m) = (20,30)$ ($p < n$, easy case) and finally $(n,m) = (20,120)$.

We use the version of the Transductive LASSO proposed in Section 6.4: for a given λ_1, we first compute the LASSO estimator $\hat{\beta}_{X,\lambda_1}$. In the sequel, the Transductive LASSO is given by

$$\hat{\beta}^{TL}(\lambda_1,\lambda_2) = \begin{cases} \arg\min_{\beta\in\mathbb{R}^p} \frac{n}{m}\|Z\beta\|_2^2 \\ \\ s.t. \left\| \frac{n}{m} Z'(Z\hat{\beta}_{X,\lambda_1} - Z\beta) \right\|_\infty \leq \lambda_2, \end{cases}$$

for a given λ_2. We compare this two step procedure with the procedure obtained using the usual LASSO only: $\hat{\beta}^L(\lambda) = \hat{\beta}_{X,\lambda}$ for a given λ that may differ from λ_1. In both cases, the solutions are computed using PCO algorithm. We compute $\hat{\beta}^L(\lambda)$ and $\hat{\beta}^{TL}(\lambda_1,\lambda_2)$ for $(\lambda,\lambda_1,\lambda_2) \in \Lambda^3$ where $\Lambda^3 = \{1.2^k, k = -50,-49,\ldots,30\}$. In the next subsection, we examine the performance of each estimator according to the value of the regularization parameters.

Results. We illustrate here some of the results obtained in the considered cases.

Case $(n,m) = (7,10)$, $\rho = 0.5$, $\sigma = 1$ *and* β^* *"sparse":*
We simulated 100 experiments and studied the distribution of

$$PERF(X) = \frac{\min_{(\lambda_1,\lambda_2)\in\Lambda^2} \|X(\hat{\beta}^{TL}(\lambda_1,\lambda_2) - \beta^*)\|_2^2}{\min_{\lambda\in\Lambda} \|X(\hat{\beta}^L(\lambda) - \beta^*)\|_2^2},$$

$$PERF(Z) = \frac{\min_{(\lambda_1,\lambda_2)\in\Lambda^2} \|Z(\hat{\beta}^{TL}(\lambda_1,\lambda_2) - \beta^*)\|_2^2}{\min_{\lambda\in\Lambda} \|Z(\hat{\beta}^L(\lambda) - \beta^*)\|_2^2},$$

and

$$PERF(I) = \frac{\min_{(\lambda_1,\lambda_2)\in\Lambda^2} \|\hat{\beta}^{TL}(\lambda_1,\lambda_2) - \beta^*\|_2^2}{\min_{\lambda\in\Lambda} \|\hat{\beta}^L(\lambda) - \beta^*\|_2^2},$$

over all the experiments.

For example, we plot (Figure 6.1) the histogram of $PERF(X)$ (actually, the three distributions where quite similar). We observe that in 50% of the simulations, $\min_{(\lambda_1,\lambda_2)\in\Lambda^2} \|X(\hat{\beta}^{TL}(\lambda_1,\lambda_2) - \beta^*)\|_2^2 = \min_{(\lambda_1,0)\in\Lambda^2} \|X(\hat{\beta}^{TL}(\lambda_1,0) - \beta^*)\|_2^2 = \min_{\lambda\in\Lambda} \|X(\hat{\beta}^L(\lambda) - \beta^*)\|_2^2$. In these cases, the Transductive LASSO does not improve at all the LASSO. But in the others 50%, the Transductive LASSO actually improve the LASSO, and the improvement is sometimes really important. We give an overview of the results in Table 6.1.

The other cases :
The following conclusions emerge of the experiments: first, $\beta^* = (5,0,\ldots,0)$ leads to a more significative improvement of the Transductive LASSO compared to the LASSO (Table 6.1).

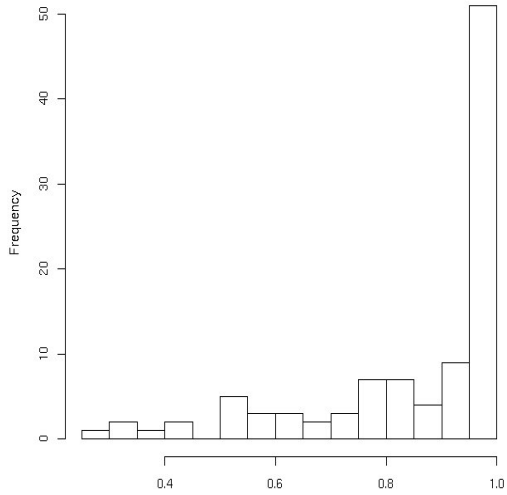

Figure 6.1: Histogram of $PERF(X)$ with $(n, m) = (7, 10)$, $\rho = 0.5$, $\sigma = 1$ and $\beta^* = (3, 1.5, 0, 0, 2, 0, 0, 0)$.

This good performance of the Transductive LASSO can also be observed when $(n, m) = (7, 10)$ and $(n, m) = (7, 20)$. However in the case $n > p$ (easy case), i.e., $(n, m) = (20, 30)$ and $(n, m) = (20, 120)$, the improvement of the Transductive LASSO with respect to the LASSO becomes less significant (Table 6.1).
Finally, ρ and σ have of course a significant influence on the performance of the LASSO. However these parameters do not seem to have any influence on the relative performance of the Transductive LASSO with respect to the LASSO (see for instant the three last rows in Table 6.1, where we kept $(n, m) = (20, 30)$).
Quite surprisingly, the relative performance of both estimators does not strongly depend on the estimation objective β^*, $X\beta^*$ or $Z\beta^*$, but on the particular experiment we deal with. According to the realized study and for all the objectives, the Transductive LASSO performs better than the LASSO in about 50% of the experiments. Otherwise, $\lambda_1 = 0$ is the optimal tuning parameter and then, the LASSO and the Transductive LASSO are equivalent.

Also surprising is that as often as not, the minimum in

$$\min_{(\lambda_1,\lambda_2)\in\Lambda^2} \|X(\hat{\beta}^{TL}(\lambda_1, \lambda_2) - \beta^*)\|_2^2 < \min_{(\lambda_1,0)\in\Lambda^2} \|X(\hat{\beta}^{TL}(\lambda_1, 0) - \beta^*)\|_2^2,$$

does not significantly depend on λ_1 for a very large range of values λ_1. This is quite interesting for a practitioner as it means that when we use the Transductive LASSO, we deal with only a singular unknown tuning parameter (that is λ_2) and not two.

Discussion on the regularization parameter. Finally, we would like to point out the importance of the tuning parameter λ (in a general term). Figure 6.2 illustrates a graph

Table 6.1: Evaluation of the mean ME and the quantile Q_3 of order 0.3 of $PERF(I)$, $PERF(X)$ and $PERF(Z)$. In these experiments, σ always equals 1. The case *sparse* corresponds to $\beta^* = (3, 1.5, 0, 0, 2, 0, 0, 0)$ while the case *very sparse* corresponds to $\beta^* = (5, 0, 0, 0, 0, 0, 0, 0)$.

				$PERF(I)$		$PERF(X)$		$PERF(Z)$	
β^*	(n, m)	ρ	σ	ME	Q_3	ME	Q_3	ME	Q_3
VERY SPARSE	$(7, 10)$	0.5	1	0.74	0.71	0.76	0.71	0.75	0.70
SPARSE	$(7, 10)$	0.5	1	0.83	0.76	0.86	0.80	0.88	0.88
SPARSE	$(7, 20)$	0.5	1	0.84	0.79	0.84	0.81	0.88	0.89
SPARSE	$(20, 30)$	0.5	1	0.91	0.90	0.93	0.93	0.93	0.95
SPARSE	$(20, 30)$	0.9	1	0.91	0.93	0.94	0.95	0.93	0.96
SPARSE	$(20, 30)$	0.5	3	0.90	0.89	0.92	0.92	0.92	0.93

of a typical experiment. There are two curves on this graph, that represent the quantities $(1/n)\|X(\hat{\beta}^L(\lambda) - \beta^*)\|_2^2$ and $(1/m)\|Z(\hat{\beta}^L(\lambda) - \beta^*)\|_2^2$ with respect to λ. We observe that both functions do not reach their minimum value for the same value of λ (the minimum is highlighted on the graph by a dot), even if these minimum are quite close.

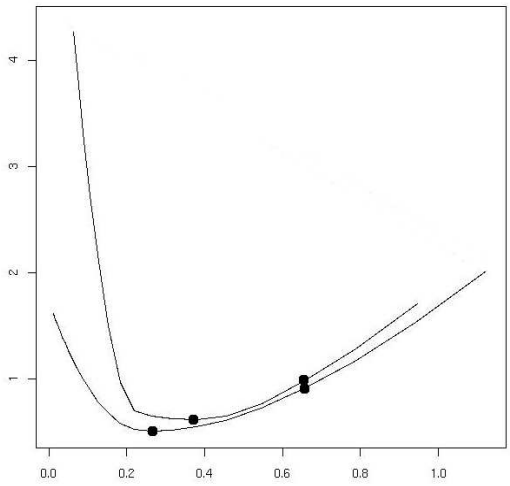

Figure 6.2: Performance vs. λ.

Since we consider variable selection methods, the identification of the true support $\{j : \beta_j^* \neq 0\}$ of the vector β^* is also in concern. One expects that the estimator $\hat{\beta}$ and the true vector β^* share the same support at least when n is large enough. This is known as the variable selection consistency problem and it has been considered for the LASSO estimator in several works (see the papers by Bunea [28] Meinshausen and Bühlmann [112], Meinshausen and Yu [114], Wainwright [152] and Zhao and Yu [164]). Recently, Lounici

[101] provided the variable selection consistency of the Dantzig Selector. Other popular selection procedures, based on the LASSO estimator, such as the Adaptive LASSO by Zou [166], the SCAD by Fan and Li [63], the S-LASSO by Hebiri [78] and the Group-LASSO by Bach [12], have also been studied under a variable selection point of view. Following our previous work (Alquier and Hebiri [7]), it is possible to provide such results for the Transductive LASSO.

The variable selection task has also been illustrated in Figure 6.2. We reported the minimal value of λ for which the LASSO estimator identifies correctly the non zero components of β^*. This value of λ is quite different from the values that minimizes the prediction losses. This observation is recurrent in almost all the experiments: the estimation $X\beta^*$, $Z\beta^*$ and the support of β^* are three different objectives and have to be treated separately. We cannot expect in general to find a choice for λ which makes the LASSO, for instance, has good performance for all the mentioned objective simultaneously.

6.6 Conclusion

In this chapter, we propose an extension of the LASSO and the Dantzig Selector for which we provide theoretical results with less restrictive hypothesis than in previous works. These estimators have a nice interpretation in terms of transductive prediction. Moreover, we study the practical performance of the proposed transductive estimators on simulated data. It turns out that the benefit using such methods is emphasized when the model is sparse and particularly when the samples sizes (n labeled points and m unlabeled points) and dimension p are such that $n < p < m$.

6.7 Proofs

In this section, we state the proofs of our main results.

6.7.1 Proof of Propositions 6.1 and 6.2

Proof of Proposition 6.1. Let us assume that $(X'X)$ is invertible. Then just remark that the criterion minimized by $\hat{\beta}_{\sqrt{n}I,\lambda}$ is just

$$n\left\|\hat{\beta}^{LSE}-\beta\right\|_2^2+2\lambda\|\Xi_{nI}\beta\|_1=\sum_{j=1}^{p}\left\{\left[\hat{\beta}_j^{LSE}-\beta_j\right]^2+\frac{2\lambda\xi_j(\sqrt{n}I)}{n}|\beta_j|\right\}.$$

So we can optimize with respect to each coordinate β_j individually. It is quite easy to check that the solution is, for β_j,

$$sgn\left(\hat{\beta}_j^{LSE}\right)\left(\left|\hat{\beta}_j^{LSE}\right|-\frac{\lambda\xi_j(\sqrt{n}I)}{n}\right)_+.$$

The proof for $\hat{\beta}_{\sqrt{n}I,\lambda}$ is also easy as it solves

$$\begin{cases} \arg\min_{\beta\in\mathbb{R}^p} \|\beta\|_1 \\ s.t. \left\| n\Xi_{nI}^{-1}(\hat{\beta}^{LSE} - \beta) \right\|_\infty \leq \lambda. \end{cases}$$

□

Proof of Proposition 6.2. Let us write the Lagrangian of the program

$$\begin{cases} \arg\min_{\beta\in\mathbb{R}^p} \|A\beta\|_2^2 \\ s.t. \left\| \Xi_A^{-1}(A'A)(\widetilde{(X'X)}^{-1}X'Y - \beta) \right\|_\infty \leq \lambda, \end{cases}$$

$$\mathcal{L}(\beta,\gamma,\mu) = \beta(Z'Z)\beta + \gamma' \left[\Xi_A^{-1}(A'A)(\widetilde{(X'X)}^{-1}X'Y - \beta) - \lambda E\right] + \mu' \left[\Xi_A^{-1}(A'A)(\beta - \widetilde{(X'X)}^{-1}X'Y) - \lambda E\right]$$

with $E = (1,\ldots,1)'$, and for any j, $\gamma_j \geq 0$, $\mu_j \geq 0$ and $\gamma_j\mu_j = 0$. Any solution $\underline{\beta} = \underline{\beta}(\gamma,\mu)$ must satisfy

$$0 = \frac{\partial\mathcal{L}}{\partial\beta}(\underline{\beta},\lambda,\mu) = 2\underline{\beta}(A'A) + (\gamma-\mu)\Xi_A^{-1}(A'A)$$

so

$$(A'A)\underline{\beta} = (A'A)\Xi_A^{-1}\frac{\mu-\gamma}{2}.$$

Note that the conditions $\gamma_j \geq 0$, $\mu_j \geq 0$ and $\gamma_j\mu_j = 0$ means that there is a $\zeta_j \in \mathbb{R}$ such that $\zeta_j = \xi_j^{\frac{1}{2}}(A)(\mu_j - \gamma_j)/2$, $|\zeta_j| = \xi_j^{\frac{1}{2}}(A)(\gamma_j + \mu_j)/2$, and so $\gamma_j = 2(\zeta_j/\xi_j^{\frac{1}{2}}(A))_-$ and $\mu_j = 2(\zeta_j/\xi_j^{\frac{1}{2}}(A))_+$, where $(a)_+ = max(a;0)$ and $(a)_- = max(-a;0)$. Let also ζ denote the vector which j-th component is exactly ζ_j, we obtain

$$(A'A)\underline{\beta} = (A'A)\zeta,$$

or, using the condition $\mathrm{Ker}(A) = \mathrm{Ker}(X)$, $X\underline{\beta} = X\zeta$ and $A\underline{\beta} = A\zeta$. This leads to

$$\mathcal{L}(\underline{\beta},\gamma,\mu) = -2Y'X\widetilde{(X'X)}^{-1}(A'A)\zeta + \zeta'(A'A)\zeta + 2\lambda\|\Xi_A\zeta\|_1,$$

and note that the first order condition also implies that γ and μ (and so ζ) maximize $\mathcal{L}$. This ends the proof. □

6.7.2 A useful Lemma

The following lemma will be used in the proofs of Theorems 6.1 and 6.2.

Lemma 6.1. *Let us put* $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$. *If* $\mathrm{Ker}(A) = \mathrm{Ker}(X)$ *we have, with probability at least* $1 - \eta$,

$$\forall j \in \{1, \ldots, p\}, \left| \left[A'A(\widetilde{X'X})^{-1} X'\varepsilon \right]_j \right| \leq \xi_j(A)\sigma \sqrt{2n \log \frac{p}{\eta}},$$

or, in other words,

$$\|\Xi_A^{-1}(A'A)((\widetilde{X'X})^{-1}X'Y - \beta^*)\|_\infty \leq \sigma \sqrt{2n \log \frac{p}{\eta}}.$$

Proof of the lemma. By definition, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ and so

$$(A'A)(\widetilde{X'X})^{-1}X'\varepsilon \sim \mathcal{N}(0, \sigma^2 (A'A)(\widetilde{X'X})^{-1}(A'A)).$$

So, for all j, $[(A'A)(\widetilde{X'X})^{-1}X'\varepsilon]_j$ comes from a $\mathcal{N}(0, \sigma^2 \xi_j^2(A))$ distribution. This implies the first point, the second one is trivial using $Y = X\beta^* + \varepsilon$. □

6.7.3 Proof of Theorems 6.1 and 6.2

Proof of Theorem 6.1. By definition of $\hat{\beta}_{A,\lambda}$ we have

$$\begin{aligned} &- 2Y'X(\widetilde{X'X})^{-1}(A'A)\hat{\beta}_{A,\lambda} + \left(\hat{\beta}_{A,\lambda}\right)'(A'A)\hat{\beta}_{A,\lambda} + 2\lambda\|\Xi_{A'A}\hat{\beta}_{A,\lambda}\|_1 \\ &\qquad \leq 2Y'X(\widetilde{X'X})^{-1}(A'A)\beta^* + (\beta^*)'(A'A)\beta^* + 2\lambda\|\Xi_A\beta^*\|_1. \end{aligned}$$

Since $Y = X\beta^* + \varepsilon$, we obtain

$$\begin{aligned} &2(\beta^*)'X'X(\widetilde{X'X})^{-1}(A'A)\left(\beta^* - \hat{\beta}_{A,\lambda}\right) + \left(\hat{\beta}_{A,\lambda}\right)'(A'A)\hat{\beta}_{A,\lambda} - (\beta^*)'(A'A)\beta^* \\ &\qquad + 2\varepsilon'X(\widetilde{X'X})^{-1}(A'A)\left(\beta^* - \hat{\beta}_{A,\lambda}\right) \leq 2\lambda\|\Xi_A\beta^*\|_1 - 2\lambda\|\Xi_A\hat{\beta}_{A,\lambda}\|_1. \end{aligned}$$

Now, if $\mathrm{Ker}(X) = \mathrm{Ker}(A)$ then we have $X'X(\widetilde{X'X})^{-1}(A'A) = (A'A)$ and then the previous inequality leads to

$$\begin{aligned} &\left(\beta^* - \hat{\beta}_{A,\lambda}\right)'(A'A)\left(\beta^* - \hat{\beta}_{A,\lambda}\right) \\ &\qquad \leq 2\varepsilon'X(\widetilde{X'X})^{-1}(A'A)\left(\hat{\beta}_{A,\lambda} - \beta^*\right) + 2\lambda\|\Xi_A\beta^*\|_1 - 2\lambda\|\Xi_A\hat{\beta}_{A,\lambda}\|_1. \quad (6.9) \end{aligned}$$

Now we have to work on the term $2\varepsilon'X(\widetilde{X'X})^{-1}(A'A)\left(\hat{\beta}_{A,\lambda} - \beta^*\right)$. Note that

$$\begin{aligned} 2\varepsilon'X(\widetilde{X'X})^{-1}(A'A)\left(\hat{\beta}_{A,\lambda} - \beta^*\right) &= 2\sum_{j=1}^p \left(\hat{\beta}_{A,\lambda} - \beta^*\right)_j \left[(A'A)(\widetilde{X'X})^{-1}X'\varepsilon\right]_j \\ &\leq 2\sum_{j=1}^p \left|\left(\hat{\beta}_{A,\lambda} - \beta^*\right)_j\right| \left|\left[(A'A)(\widetilde{X'X})^{-1}X'\varepsilon\right]_j\right| \\ &\leq 2\sigma\sqrt{2n\log\left(\frac{p}{\eta}\right)} \sum_{j=1}^p \xi_j^{\frac{1}{2}}(A)\left|\left(\hat{\beta}_{A,\lambda}\right)_j - \beta_j^*\right| \end{aligned}$$

with probability at least $1-\eta$, by Lemma 6.1. We plug this result into Inequality (6.9) (and replace λ by its value $2\sigma\sqrt{2n\log(p/\eta)}$) to obtain

$$\left(\beta^* - \hat{\beta}_{A,\lambda}\right)' (A'A) \left(\beta^* - \hat{\beta}_{A,\lambda}\right) \leq 2\sigma\sqrt{2n\log\left(\frac{p}{\eta}\right)} \sum_{j=1}^{p} \xi_j^{\frac{1}{2}}(A) \left\{ \left| \left(\hat{\beta}_{A,\lambda}\right)_j - \beta_j^* \right| + 2\left(|\beta_j^*| - \left| \left(\hat{\beta}_{A,\lambda}\right)_j \right| \right) \right\}$$

and then

$$\begin{aligned}
& \left(\beta^* - \hat{\beta}_{A,\lambda}\right)' (A'A) \left(\beta^* - \hat{\beta}_{A,\lambda}\right) \\
& +2\sigma\sqrt{2n\log\left(\frac{p}{\eta}\right)} \sum_{j=1}^{p} \xi_j^{\frac{1}{2}}(A) \left| \left(\hat{\beta}_{A,\lambda}\right)_j - \beta_j^* \right| \\
\leq\ & 4\sigma\sqrt{2n\log\left(\frac{p}{\eta}\right)} \sum_{j=1}^{p} \xi_j^{\frac{1}{2}}(A) \left\{ \left| \left(\hat{\beta}_{A,\lambda}\right)_j - \beta_j^* \right| + |\beta_j^*| - \left| \left(\hat{\beta}_{A,\lambda}\right)_j \right| \right\} \\
=\ & 4\sigma\sqrt{2n\log\left(\frac{p}{\eta}\right)} \sum_{j:\beta_j^*\neq 0} \xi_j^{\frac{1}{2}}(A) \left\{ \left| \left(\hat{\beta}_{A,\lambda}\right)_j - \beta_j^* \right| + |\beta_j^*| - \left| \left(\hat{\beta}_{A,\lambda}\right)_j \right| \right\} \\
\leq\ & 8\sigma\sqrt{2n\log\left(\frac{p}{\eta}\right)} \sum_{j:\beta_j^*\neq 0} \xi_j^{\frac{1}{2}}(A) \left| \left(\hat{\beta}_{A,\lambda}\right)_j - \beta_j^* \right|. \qquad (6.10)
\end{aligned}$$

This implies, in particular, that $\beta^* - \hat{\beta}_{A,\lambda}$ is an admissible vector α in Assumption $H(A'A, 3)$ because

$$\sum_{j=1}^{p} \xi_j^{\frac{1}{2}}(A) \left| \left(\hat{\beta}_{A,\lambda}\right)_j - \beta_j^* \right| \leq 4 \sum_{j:\beta_j^*\neq 0} \xi_j^{\frac{1}{2}}(A) \left| \left(\hat{\beta}_{A,\lambda}\right)_j - \beta_j^* \right|.$$

On the other hand, thanks to Inequality (6.10), we have

$$\begin{aligned}
& \left(\beta^* - \hat{\beta}_{A,\lambda}\right)' (A'A) \left(\beta^* - \hat{\beta}_{A,\lambda}\right) \\
& \leq 6\sigma\sqrt{2n\log\left(\frac{p}{\eta}\right)} \sum_{j:\beta_j^*\neq 0} \xi_j^{\frac{1}{2}}(A) \left| \left(\hat{\beta}_{A,\lambda}\right)_j - \beta_j^* \right| \\
& \leq 6\sigma\sqrt{2n \sum_{j:\beta_j^*\neq 0} \left[\left(\hat{\beta}_{A,\lambda}\right)_j - \beta_j^* \right]^2 \sum_{j:\beta_j^*\neq 0} \xi_j(A) \log\left(\frac{p}{\eta}\right)} \\
& \leq 6\sigma\sqrt{\frac{2}{c(A'A)} \left(\beta^* - \hat{\beta}_{A,\lambda}\right)' (A'A) \left(\beta^* - \hat{\beta}_{A,\lambda}\right) \sum_{j:\beta_j^*\neq 0} \xi_j(M) \log\left(\frac{p}{\eta}\right)}, \qquad (6.11)
\end{aligned}$$

where we used Assumption $H(A'A, 3)$ for the last inequality. Then

$$\left(\beta^* - \hat{\beta}_{A,\lambda}\right)' (A'A) \left(\beta^* - \hat{\beta}_{A,\lambda}\right) \leq 72 \frac{\sigma^2}{c(A'A)} \log\left(\frac{p}{\eta}\right) \sum_{j:\beta_j^*\neq 0} \xi_j(A). \qquad (6.12)$$

A similar reasoning as in (6.11) leads to

$$2\sigma\sqrt{2n\log\left(\frac{p}{\eta}\right)}\sum_{j=1}^{p}\xi_j^{\frac{1}{2}}(A)\left|\left(\hat{\beta}_{A,\lambda}\right)_j-\beta_j^*\right| \\ \leq 8\sigma\sqrt{\frac{2}{c(A'A)}\left(\beta^*-\hat{\beta}_{A,\lambda}\right)'(A'A)\left(\beta^*-\hat{\beta}_{A,\lambda}\right)\sum_{j:\beta_j^*\neq 0}\xi_j(M)\log\left(\frac{p}{\eta}\right)}.$$

Finally, combine this last inequality with (6.12) to obtain the desired bound for $\left\|\Xi_A\left(\beta^*-\hat{\beta}_{A,\lambda}\right)\right\|_1$. This ends the proof. □

Proof of Theorem 6.2. We have

$$\begin{aligned}(\tilde{\beta}_{A,\lambda}-\beta^*)'(A'A)(\tilde{\beta}_{A,\lambda}-\beta^*) &= [\Xi_A(\tilde{\beta}_{A,\lambda}-\beta^*)]'\Xi_A^{-1}(A'A)(\tilde{\beta}_{A,\lambda}-\beta^*)\\ &\leq \|\Xi_A(\tilde{\beta}_{A,\lambda}-\beta^*)\|_1\|\Xi_A^{-1}(A'A)(\tilde{\beta}_{A,\lambda}-\beta^*)\|_\infty\\ &\leq \|\Xi_A(\tilde{\beta}_{A,\lambda}-\beta^*)\|_1\Big\{\|\Xi_A^{-1}(A'A)((\widetilde{X'X})^{-1}X'Y-\beta^*)\|_\infty\\ &\qquad +\|\Xi_A^{-1}(A'A)((\widetilde{X'X})^{-1}X'Y-\tilde{\beta}_{A,\lambda})\|_\infty\Big\}, \quad (6.13)\end{aligned}$$

by the constraint in the definition on $\tilde{\beta}_{A,\lambda}$ we have

$$\|\Xi_A^{-1}(A'A)((\widetilde{X'X})^{-1}X'Y-\tilde{\beta}_{A,\lambda})\|_\infty\leq\lambda,$$

while Lemma 6.1 implies that for $\lambda=2\sigma\sqrt{2n\log(p/\eta)}$ we have

$$\|\Xi_A^{-1}(A'A)((\widetilde{X'X})^{-1}X'Y-\beta^*)\|_\infty\leq\frac{\lambda}{2},$$

with probability at least $1-\eta$; and so:

$$(\tilde{\beta}_{A,\lambda}-\beta^*)'(A'A)(\tilde{\beta}_{A,\lambda}-\beta^*)\leq\frac{3\lambda}{2}\|\Xi_A(\tilde{\beta}_{A,\lambda}-\beta^*)\|_1.$$

Moreover note that, by definition,

$$\begin{aligned}0\leq\|\Xi_A\beta^*\|_1-\|\Xi_A\tilde{\beta}_{A,\lambda}\|_1 &= \sum_{\beta_j^*\neq 0}\xi_j^{\frac{1}{2}}(A)\left|\beta_j^*\right|-\sum_{\beta_j^*\neq 0}\xi_j^{\frac{1}{2}}(A)\left|(\tilde{\beta}_{A,\lambda})_j\right|-\sum_{\beta_j^*=0}\xi_j^{\frac{1}{2}}(A)\left|(\tilde{\beta}_{A,\lambda})_j\right|\\ &\leq \sum_{\beta_j^*\neq 0}\xi_j^{\frac{1}{2}}(A)\left|\beta_j^*-(\tilde{\beta}_{A,\lambda})_j\right|-\sum_{\beta_j^*=0}\xi_j^{\frac{1}{2}}(A)\left|\beta_j^*-(\tilde{\beta}_{A,\lambda})_j\right|,\end{aligned}$$

this implies that $\beta^* - (\tilde{\beta}_{A,\lambda})$ is an admissible vector in the relation that defines Assumption $H(A'A, 1)$. Let us combine this result with Inequality (6.13), we obtain

$$\begin{aligned}
(\tilde{\beta}_{A,\lambda} - \beta^*)'(A'A)(\tilde{\beta}_{A,\lambda} - \beta^*) &\leq \frac{3\lambda}{2}\|\Xi_A(\beta^* - \tilde{\beta}_{A,\lambda})\|_1 \\
&\leq 3\lambda \sum_{\beta_j^* \neq 0} \xi_j^{\frac{1}{2}}(A) \left|\beta_j^* - (\tilde{\beta}_{A,\lambda})_j\right| \\
&\leq 3\lambda \sqrt{\left(\sum_{\beta_j^* \neq 0} \xi_j(A)\right)\left(\sum_{\beta_j^* \neq 0} \left|\beta_j^* - (\tilde{\beta}_{A,\lambda})_j\right|^2\right)} \\
&\leq 3\lambda \left(\sum_{\beta_j^* \neq 0} \xi_j(A)\right)^{\frac{1}{2}} \sqrt{\frac{1}{nc(A'A)}(\tilde{\beta}_{A,\lambda} - \beta^*)'(A'A)(\tilde{\beta}_{A,\lambda} - \beta^*)}. \quad (6.14)
\end{aligned}$$

So we have,

$$(\tilde{\beta}_{A,\lambda} - \beta^*)'(A'A)(\tilde{\beta}_{A,\lambda} - \beta^*) \leq 9\lambda^2 \frac{1}{nc(A'A)} \sum_{\beta_j^* \neq 0} \xi_j(A),$$

and as a consequence, Inequality (6.14) gives the upper bound on $\|\Xi_A(\tilde{\beta}_{A,\lambda} - \beta^*)\|_1$, and this ends the proof. □

6.7.4 Proof of Theorem 6.3

Proof of Theorem 6.3. The proof is almost the same as in the previous case. For the sake of simplicity, let us write $\tilde{\beta}^*$ instead of $\tilde{\beta}^*_{\sqrt{n/m}Z,\lambda_2}$ and the same for $\hat{\beta}^*$. We first give a look at the Dantzig Selector:

$$\begin{aligned}
\frac{n}{m}\left(\tilde{\beta}^* - \beta^*\right)' Z'Z \left(\tilde{\beta}^* - \beta^*\right) &\leq \left\|\tilde{\beta}^* - \beta^*\right\|_1 \left\|\frac{n}{m} Z'Z \left(\tilde{\beta}^* - \beta^*\right)\right\|_\infty \\
&\leq \left\|\tilde{\beta}^* - \beta^*\right\|_1 \left\{\left\|\frac{n}{m} Z'\left(Z\tilde{\beta}^* - \check{Y}_{\lambda_1}\right)\right\|_\infty + \left\|\frac{n}{m} Z'\left(Z\beta^* - \check{Y}_{\lambda_1}\right)\right\|_\infty\right\} \\
&\leq \left\|\tilde{\beta}^* - \beta^*\right\|_1 \left\{\left\|\frac{n}{m} Z'\left(Z\tilde{\beta}^* - \check{Y}_{\lambda_1}\right)\right\|_\infty + \left\|X'\left(X\beta^* - Y\right)\right\|_\infty \right. \\
&\quad \left. + \left\|X'\left(X\tilde{\beta}_{X,\lambda_1} - Y\right)\right\|_\infty + \left\|\left(\frac{n}{m} Z'Z - X'X\right)\left(\beta^* - \tilde{\beta}_{X,\lambda_1}\right)\right\|_\infty\right\}. \quad (6.15)
\end{aligned}$$

By Lemma 6.1, for $\lambda_1 = 10^{-1}\sigma\sqrt{2n\log(p/\eta)}$ we have

$$\|X'Y - X'X\beta^*\|_\infty \leq 10\lambda_1,$$

with probability at least $1 - \eta$. On the other hand, we have

$$\|\beta^* - \tilde{\beta}_{X,\lambda_1}\|_1 \leq \|\beta^*\|_1 + \|\tilde{\beta}_{X,\lambda_1}\|_1 \leq 2\|\beta^*\|_1,$$

by definition of the Dantzig Selector. Then, let $u = (\beta^* - \tilde{\beta}_{X,\lambda_1})/2$ and use Inequality (6.8) for this specific u. This ensures that

$$\left\| \left(\frac{n}{m} Z'Z - X'X \right) \left(\beta^* - \tilde{\beta}_{X,\lambda_1} \right) \right\|_\infty \leq 2\lambda_1. \tag{6.16}$$

The definition of the Dantzig Selector also implies that

$$\left\| X' \left(X\tilde{\beta}_{X,\lambda_1} - Y \right) \right\|_\infty \leq \lambda_1,$$

and finally the definition of the estimator leads to

$$\left\| \frac{n}{m} Z' \left(Z\tilde{\beta}^* - \check{Y}_{\lambda_1} \right) \right\|_\infty \leq \lambda_2 = \lambda_1,$$

and as a consequence, Inequality (6.15) becomes

$$\frac{n}{m} \left(\tilde{\beta}^* - \beta^* \right)' Z'Z \left(\tilde{\beta}^* - \beta^* \right) \leq 14\lambda_1 \left\| \tilde{\beta}^* - \beta^* \right\|_1.$$

Using the fact that $\|\tilde{\beta}^*\|_1 \leq \|\beta^*\|_1$ gives

$$\begin{aligned}
\frac{n}{m} \left(\tilde{\beta}^* - \beta^* \right)' Z'Z \left(\tilde{\beta}^* - \beta^* \right) &\leq 14\lambda_1 \left\| \tilde{\beta}^* - \beta^* \right\|_1 \leq 28\lambda_1 \sum_{\beta_j^* \neq 0} \left| \beta_j^* - (\tilde{\beta}^*)_j \right| \\
&\leq 28\lambda_1 \sqrt{\left| \left\{ j : \beta_j^* \neq 0 \right\} \right| \left(\sum_{\beta_j^* \neq 0} \left| \beta_j^* - (\tilde{\beta}^*)_j \right|^2 \right)} \\
&\leq 28\lambda_1 \left| \left\{ j : \beta_j^* \neq 0 \right\} \right|^{\frac{1}{2}} \sqrt{\frac{1}{nc(n/m(Z'Z))} \frac{n}{m} \left(\tilde{\beta}^* - \beta^* \right)' Z'Z \left(\tilde{\beta}^* - \beta^* \right)}. \quad (6.17)
\end{aligned}$$

To establish the last inequality, we used Assumption $H'((n/m)Z'Z, 1)$. Then we have,

$$\frac{n}{m} \left(\tilde{\beta}^* - \beta^* \right)' Z'Z \left(\tilde{\beta}^* - \beta^* \right) \leq 28^2 \lambda_1^2 \left| \left\{ j : \beta_j^* \neq 0 \right\} \right| \frac{1}{nc(n/m(Z'Z))},$$

This inequality, combined with (6.17), end the proof for the Dantzig Selector.

Now, let us deal with the LASSO case. The dual form of the definition of the estimator leads to

$$\begin{aligned}
&- 2\frac{n}{m} \check{Y}_{\lambda_1} Z\hat{\beta}^* + \frac{n}{m} (\hat{\beta}^*)' Z'Z\hat{\beta}^* + 40\lambda_2 \|\hat{\beta}^*\|_1 \\
&\qquad \leq -2\frac{n}{m} \check{Y}_{\lambda_1} Z\beta^* + \frac{n}{m} (\beta^*)' Z'Z\beta^* + 40\lambda_2 \|\beta^*\|_1
\end{aligned}$$

and so

$$\begin{aligned}
&- 2\frac{n}{m} \tilde{\beta}_{X,\lambda_1} Z'Z\hat{\beta}^* + \frac{n}{m} (\hat{\beta}^*)' Z'Z\hat{\beta}^* + 40\lambda_2 \|\hat{\beta}^*\|_1 \\
&\qquad \leq -2\frac{n}{m} \tilde{\beta}_{X,\lambda_1} Z'Z\beta^* + \frac{n}{m} (\beta^*)' Z'Z\beta^* + 40\lambda_2 \|\beta^*\|_1.
\end{aligned}$$

As a consequence,

$$\frac{n}{m}\left(\hat{\beta}^*-\beta^*\right)' Z'Z\left(\hat{\beta}^*-\beta^*\right) \leq 2\frac{n}{m}\left(\hat{\beta}^*-\beta^*\right)' Z'Z\left(\tilde{\beta}_{X,\lambda_1}-\beta^*\right)+40\lambda_2\left(\|\beta^*\|_1-\|\hat{\beta}^*\|_1\right).$$

Now, we try to upper bound $\left(\hat{\beta}^*-\beta^*\right)' Z'Z\left(\tilde{\beta}_{X,\lambda_1}-\beta^*\right)$. We remark that

$$\begin{aligned}
&\frac{n}{m}\left(\hat{\beta}^*-\beta^*\right)' Z'Z\left(\tilde{\beta}_{X,\lambda_1}-\beta^*\right) \leq \left\|\hat{\beta}^*-\beta^*\right\|_1 \left\|\frac{n}{m}(Z'Z)\left(\tilde{\beta}_{X,\lambda_1}-\beta^*\right)\right\|_\infty \\
\leq\ & \left\|\hat{\beta}^*-\beta^*\right\|_1 \Bigg[\left\|\left(\left(\frac{n}{m}Z'Z-X'X\right)\left(\tilde{\beta}_{X,\lambda_1}-\beta^*\right)\right\|_\infty \\
&+\left\|X'X\left(\tilde{\beta}_{X,\lambda_1}-\beta^*\right)\right\|_\infty\Bigg] \leq 13\lambda_1\left\|\hat{\beta}^*-\beta^*\right\|_1,
\end{aligned}$$

where we used (6.16) and the fact that

$$\left\|X'X\left(\tilde{\beta}_{X,\lambda_1}-\beta^*\right)\right\|_\infty \leq \left\|X'\left(X\tilde{\beta}_{X,\lambda_1}-Y\right)\right\|_\infty+\|X'\varepsilon\|_\infty \leq \lambda_1+10\lambda_1=11\lambda_1.$$

Then we have

$$\frac{n}{m}\left(\hat{\beta}^*-\beta^*\right)' Z'Z\left(\hat{\beta}^*-\beta^*\right) \leq 26\lambda_1\left\|\hat{\beta}^*-\beta^*\right\|_1+40\lambda_2\left(\|\beta^*\|_1-\|\hat{\beta}^*\|_1\right),$$

and so

$$\frac{n}{m}\left(\hat{\beta}^*-\beta^*\right)' Z'Z\left(\hat{\beta}^*-\beta^*\right)+14\lambda_1\left\|\hat{\beta}^*-\beta^*\right\|_1 \leq 40\lambda_1\left(\left\|\hat{\beta}^*-\beta^*\right\|_1+\|\beta^*\|_1-\|\hat{\beta}^*\|_1\right).$$

Up to a multiplying constant, the rest of the proof of Theorem 6.3 is the same as the last lines in the proof of Theorem 6.1. Then we omit it here. □

6.7.5 Proof of Proposition 6.3

Proof of Proposition 6.3. First, let us remark that

$$\left\|\left(X'X-\frac{n}{m}Z'Z\right)u\right\|_\infty = n\sup_{1\leq i\leq p}\sum_{j=1}^{p}u_j\left(\frac{X_i'X_j}{n}-\frac{Z_i'Z_j}{m}\right) \leq n\|u\|_1\sup_{1\leq i,j\leq p}\left|\frac{X_i'X_j}{n}-\frac{Z_i'Z_j}{m}\right|.$$

Now, using the "exchangeable-distribution inequality" by Catoni [39] we obtain, for a given pair (i,j), for any $\tau > 0$, with probability at least $1-\eta$,

$$\frac{X_i' X_j}{n} - \frac{Z_i' Z_j}{m} \leq \frac{\tau k^2}{2n(k+1)^2} \left(\frac{1}{m} \sum_{k=1}^{m} X_{i,k}^2 X_{j,k}^2 \right) + \frac{\log \frac{1}{\eta}}{\tau}$$

$$\leq \frac{\tau k^2 \kappa^2}{2n(k+1)^2} + \frac{\log \frac{1}{\eta}}{\tau} = \frac{\kappa k}{k-1} \sqrt{\frac{2 \log \frac{1}{\eta}}{n}},$$

for $\tau = (\log(1/\eta)(k-1)2n/k\kappa^2)^{1/2}$ and so, by a union bound argument, with probability at least $1-\eta$, for any pair (i,j),

$$\left| \frac{X_i' X_j}{n} - \frac{Z_i' Z_j}{m} \right| \leq \frac{\kappa k}{k-1} \sqrt{\frac{2 \log \frac{2p^2}{\eta}}{n}} \leq \frac{2\kappa k}{k-1} \sqrt{\frac{2 \log \frac{p}{\eta}}{n}},$$

(where we used $p \geq 2$). □

Chapter 7

Estimation with Sparse Conformal Predictors

Abstract: Conformal predictors, introduced by Vovk, Gammerman, and Shafer [150], serve to build prediction intervals by exploiting a notion of conformity of the new data point with previously observed data. In the present paper, we propose a novel method for constructing prediction intervals for the response variable in multivariate linear models. The main emphasis is on sparse linear models, where only few of the covariates have significant influence on the response variable even if their number is very large. Our approach is based on combining the principle of conformal prediction with the ℓ_1 penalized least squares estimator (LASSO). The resulting confidence set depends on a parameter $\varepsilon > 0$ and has a coverage probability larger than or equal to $1 - \varepsilon$. The numerical experiments reported in the paper show that the length of the confidence set is small. Furthermore, as a by-product of the proposed approach, we provide a data-driven procedure for choosing the LASSO penalty. The selection power of the method is illustrated on simulated data.

7.1 Introduction

Consider observations $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$ for $i \geq 1$ from a linear regression model $y_i = x_i'\beta + \xi_i$, where $\beta \in \mathbb{R}^p$ is the unknown parameter and the ξ_i's are the noise variables. Suppose we have already collected the dataset $\mathcal{E}_n = ((x_1, y_1), \ldots, (x_{n-1}, y_{n-1}), x_{new})$ where $x_{new} \in \mathbb{R}^p$ denotes a new observation. Our goal is to predict the label y_{new} corresponding to x_{new} based on $\mathcal{E}_n$ and then exploiting the information in x_{new}. This setup is known as the transduction problem (Vapnik [146]). Our estimation strategy is based on local arguments in order to produce a better estimation for y_{new} (Györfi, Kohler, Krzyżak, and Walk [73]). More precisely, we will follow the approach of *conformal prediction* presented by Vovk, Gammerman, and Shafer [150] which relies on two key ideas: one is to provide a confidence prediction (namely, a confidence set containing y_{new} with high probability) and the other is to account for the similarity of the new data x_{new} compared to the previously observed x_i's. The notion of conformal predictor was first described by Vovk, Gammerman, and Saunders [149]. Moreover, Vovk, Gammerman, and Shafer [150] illustrate this approach on the example of ridge regression. Along the paper, this predictor will be referred to as Conformal Ridge Predictor[1] (CoRP). In the present contribution, we propose to adapt conformal predictors to the sparse linear regression model, that is a model where the regression vector $\beta \in \mathbb{R}^p$ contains only a few of nonzero components. We introduce a novel conformal predictor called the *Conformal Lasso Predictor* (CoLP) which takes into account the sparsity of the model. Its construction is based on the LASSO estimator (Tibshirani [135]). The LASSO estimator for linear regression corres! ponds to an ℓ_1-penalized least square estimator and it has been extensively studied over the last few years (Knight and Fu [90], Meinshausen and Bühlmann [112], Bunea, Tsybakov, and Wegkamp [33] and Zhao and Yu [164], among others) and several modifications have been proposed (Zou [166], Yuan and Lin [159], Zou and Hastie [167], Tibshirani, Saunders, Rosset, Zhu, and Knight [136] and Hebiri [78] among others). One attractive aspect of the LASSO is that it aims both to provide accurate estimating while enjoying variable selection when the model is sparse. In the approach considered in the present paper, the resulting Conformal Lasso Predictor has a large coverage probability and are small in term of its length in the same time. When we deal with regularized methods like the Ridge or the LASSO estimators, the choice of the penalty is an important task. Contrary to the Conformal Ridge Predictor for which no rule was established to pick the Ridge-penalty (Vovk, Gammerman, and Shafer [150]), the construction of the Conformal Lasso Predictor provides a data-driven way for choosing the LASSO-penalty. Moreover, it turn out that this choice is adapted to variable selection as supported by the numerical

[1]The Conformal Ridge Predictor was called the Ridge Regression Confidence Machine in the book of Vovk, Gammerman, and Shafer [150].

experiments.

The paper is organized as follows. We concisely introduce conformal prediction and the LASSO procedure in Section 7.2 and Section 7.3 respectively. In Section 7.4, we give the explicit form of the Conformal Lasso Predictor. An algorithm producing the CoLP is presented in Section 7.5. Then in Section 7.6 we discuss a generalization of the Conformal Lasso Predictor to other selection-type procedures; we call these generalized procedures *Sparse Conformal Predictors*. Finally, in Section 7.7, we illustrate the performance of Sparse Conformal Predictors through some numerical experiments.

7.2 Conformal prediction

Let us briefly describe the approach based on conformal prediction developed in the book by Vovk, Gammerman, and Shafer [150] where they develop the idea of *conformal* prediction. In order to predict the label y_{new} of a new observation $x_n = x_{new}$, the similarity of pairs of the form (x_{new}, y), where $y \in \mathbb{R}$, to the former observations (x_i, y_i) for $i = 1, \ldots, n-1$ is exploited. This is the purpose of introducing a *nonconformity score* $\alpha(y) = (\alpha_1(y), \ldots, \alpha_n(y))'$ which is based on $\mathcal{E}_n$. Each component α_i describes the efficiency of explaining the observation (x_i, y_i) by a procedure based on the augmented sample $\{(x_1, y_1), \ldots, (x_{n-1}, y_{n-1}), (x_{new}, y)\}$. In order to obtain a relative information between different nonconformity scores α_i, we shall use the notion of *p-value*, as introduced by Vovk, Gammerman, and Shafer [150], defined as:

$$p(y) = \frac{1}{n} \left| \{ i \in \{1, \ldots, n\} : \alpha_i(y) \geq \alpha_n(y) \} \right| , \tag{7.1}$$

where for any set $\mathcal{A}$, we denote its cardinality by $|\mathcal{A}|$. The above quantity lies between $1/n$ and 1. Moreover, we note that the smaller this p-value is, the less likely the tested pair (x_{new}, y) is (in other words, y is an outlier when associated to x_{new}). An explicit form of the nonconformity score and the p-value will be given in Section 7.4 when we will adapt it to the CoLP.

Remark 7.1. *The notion of p-value introduced in the present paper differs from the classical one. To make the connection with hypothesis testing in mathematical statistics (Casella and Berger [36]), consider the following hypotheses:*

$$\begin{cases} H_0 : & \text{the pair } (x_{new}, y) \text{ is conformal,} \\ H_1 : & \text{the pair } (x_{new}, y) \text{ is not conformal.} \end{cases}$$

Assume the observation $Y = y$ is given. The function $p(y)$ permits to construct a statistical test procedure with critical region $\mathcal{R}_\varepsilon = \{y : p(y) \leq \varepsilon\}$ and H_0 is rejected if $y \in \mathcal{R}_\varepsilon$.

A nice feature of this nonconformity score is that it can be related to the confidence of the prediction for y_{new}. We now recall the concept of conformal predictor introduced by Vovk, Gammerman, and Shafer [150]. Set $\varepsilon \in (0,1)$. Given the new observation x_{new}, we search for a subset $\Gamma^\varepsilon = \Gamma^\varepsilon(\mathcal{E}_n)$ of $\mathbb{R}$, in which the expected value of y_{new} lies with a probability of $1-\varepsilon$. The conformal predictor Γ^ϵ is defined as the set of labels $y \in \mathbb{R}$ such that $p(y) > \varepsilon$. *In other words,* Γ^ε *consists of labels* y *which make the pair* (x_{new}, y) *more conformal than a proportion* ε *of the previous pairs* (x_i, y_i) *for* $i = 1, \ldots, n-1$. Note moreover that the smaller ε, the more confident the predictor. That is to say, for any ε_1, $\varepsilon_2 > 0$:

$$\Gamma^{\varepsilon_1} \subset \Gamma^{\varepsilon_2} \quad \text{whenever } \varepsilon_1 \geq \varepsilon_2 \ .$$

In the present analysis, apart from prediction, we develop an approach for selecting relevant variables. For this reason, we consider three criteria measuring the quality of our procedure: *validity*, *accuracy*, and *selection*. The first two were introduced by Vovk, Nouretdinov Ilia, and Gammerman [151]. The fact that we consider the issue of sparsity leads us to include the selection power of the predictor.

Validity. This criterion accounts for the power of conformal prediction. The simplest approach is to count the number of times where y_n does not belong to the set Γ^ε. We take the notation:

$$\text{err}_n^\varepsilon = \begin{cases} 1 & \text{if } y_n \notin \Gamma^\varepsilon(\mathcal{E}_n) \\ 0 & \text{otherwise.} \end{cases}$$

Note that in an on-line perspective, one focuses on the cumulative error $\text{ERR}_n^\varepsilon = \sum_{i=1}^n \text{err}_i^\varepsilon$. Asymptotic validity properties of this cumulative error have been studied by Vovk [147] and Vovk, Gammerman, and Shafer [150, chapters 2 and 8]. In the present work, we will be interested in evaluating the error err_n^ε for a fixed n, rather than the cumulative one.

Accuracy. The length of the confidence predictor provides a natural measure of the accuracy. We will see that such a measure is adapted to the variable selection purpose. Note that other choices are possible. We shall discuss this point in Section 7.5.

Selection. Finally, in the case of sparse linear regression, it is important to include a measure of the capacity of the estimator to select relevant variables, namely those for which the regression parameter β has nonzero components.

7.3 The LASSO Procedure

The LASSO estimator (Tibshirani [135]) has originally been introduced in the linear regression model:

$$y_i = x_i'\beta^* + \xi_i, \qquad i = 1, \dots, n-1 \tag{7.2}$$

where the design $x_i = (x_{i,1}, \dots, x_{i,p})' \in \mathbb{R}^p$ is deterministic, $\beta^* = (\beta_1^*, \dots, \beta_p^*)' \in \mathbb{R}^p$ is the unknown regression vector and the ξ_i's are independent and identically distributed (i.i.d.) centered Gaussian random variables with known variance σ^2. Then the goal is to use the observations to provide an approximation of the label y_{new} of a new observation x_{new} through the estimation of the regression vector β^*. The LASSO estimator is defined as follows:

$$\hat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n-1} \left(y_i - x_i'\beta\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{7.3}$$

where $\lambda \geq 0$ is a tuning parameter. Based on $\hat{\beta}_\lambda$, an estimation of the response y_{new} of the new observation $x_n = x_{new}$ is produced by $\hat{\mu}_\lambda = x_{new}'\hat{\beta}_\lambda$. For a large enough λ, the LASSO estimator is sparse. That is many components of $\hat{\beta}_\lambda$ equal zero. Therefore we can naturally define a sparsity (or active) set as $\mathcal{A}_\lambda = \{j \in \{1, \dots, p\} : \hat{\beta}_\lambda \neq 0\}$. A LASSO modification of the LARS algorithm (Efron, Hastie, Johnstone, and Tibshirani [61]) can iteratively provide approximations of the LASSO estimator for a few values of the tuning parameters $\lambda = \lambda_0, \dots, \lambda_K$ such that $\infty = \lambda_0 > \dots > \lambda_K = 0$ (the indices refer to the algorithm steps and K denotes the last step). These points are the so-called *transition points*.

From now on, let us write $\hat{\beta}_k$ and $\mathcal{A}_k$ for the LASSO estimator $\hat{\beta}_\lambda$ and the sparsity set $\mathcal{A}_\lambda$ evaluated at the transition point $\lambda = \lambda_k$. Obviously, the estimator $\hat{\beta}_k$ is an $|\mathcal{A}_k|$-dimensional vector where $|\mathcal{A}_k|$ is the cardinality of the set $\mathcal{A}_k$. Furthermore, we denote by s_k the $|\mathcal{A}_k|$-dimensional sign vector whose components are the signs of the components of the LASSO estimator evaluated at the transition point λ_k (i.e., $(s_k)_j = 1$ if $(\hat{\beta}_k)_j > 0$, $(s_k)_j = -1$ if $(\hat{\beta}_k)_j < 0$ where $j \in \mathcal{A}_k$). Finally, let us denote by $\mathbf{x}_k$, the $(n-1) \times |\mathcal{A}_k|$ matrix whose columns are the variables X_j, with indices $j \in \mathcal{A}_k$. For each λ_k, we assume that the matrix $(\mathbf{x}_k'\mathbf{x}_k)^{-1}$ is invertible. Here are some characteristics of the LARS algorithm and we refer to [2] for more details:

i) At each iteration of the algorithm (i.e., at each transition point), only one variable $X_j = (x_{1,j}, \dots, x_{n-1,j})'$, $j = 1, \dots, p$ is added (or deleted) to the construction of the estimator according to its correlation with the current residual. The algorithm begins with only one variable and ends up with the ordinary least square (OLS) estimator[2].

[2]When $p > n$, the LARS cannot select all p variables. It is limited by the sample size n. In such a case, the last iteration does not correspond to the OLS.

ii) For each $\lambda \in (\lambda_{k+1}, \lambda_k]$, the LASSO estimator can be expressed in the following form:

$$\hat{\beta}_\lambda(\mathbf{y}, \mathbf{x}_k, s_k) = (\mathbf{x}_k'\mathbf{x}_k)^{-1}(\mathbf{x}_k'\mathbf{y} - \frac{\lambda}{2}\, s_k), \tag{7.4}$$

where $\mathbf{y} = (y_1, \ldots, y_{n-1})'$. Note that (7.4) is obtained by minimizing (7.3) over the set $\mathcal{A}_k$. Let us also mention that the set $\mathcal{A}_k$ and the sign vector s_k remain unchanged when λ varies in the interval $(\lambda_{k+1}, \lambda_k]$.

iii) As highlighted by (7.4), the LASSO estimator is piecewise linear in λ and linear in $\mathbf{y}$ for every fixed λ (Rosset and Zhu [126]). Using the LASSO modification of the LARS algorithm, this property helps us to provide the regularization path of the LASSO estimator, which is defined as $\{\hat{\beta}_\lambda : \ \lambda \in [0, \infty)\}$ (each point of the regularization path corresponds to the evaluation of the regression vector estimator for a given value of λ). Indeed, the slope of the LASSO regularization path changes at a finite number of points which coincide with the transition points $\lambda_1, \ldots, \lambda_K$.

iv) Piecewise linearity is an important property of the LASSO modification of the LARS algorithm. Indeed, let $\lambda \in (\lambda_{k+1}, \lambda_k]$ where λ_{k+1} and λ_k are two transition points. In this interval, the LASSO estimator $\hat{\beta}_\lambda$ uses the same variables (variables with indices in $\mathcal{A}_k$). By using (7.4), it is easy to see Zou, Hastie, and Tibshirani [169] that the linearity of the LASSO estimator implies that, for any $\lambda \in (\lambda_{k+1}, \lambda_k]$:

$$\sum_{i=1}^{n-1} \left(y_i - x_i'\hat{\beta}_\lambda\right)^2 > \sum_{i=1}^{n-1} \left(y_i - x_i'\hat{\beta}_{\lambda_{k+1}}\right)^2.$$

This last observation indicates that the transition points are the most interesting points in the regularization path.

All these nice properties encourage the use of the LASSO as a selection procedure. In the sequel, we will consider the LASSO modification of the LARS algorithm which provides an approximate solution to the LASSO.

Remark 7.2. *Through the paper, one should keep in mind the analogy between each iteration k of the modification of the LARS algorithm and its corresponding tuning parameter value λ_k. Decrease of tuning parameter λ is reflected through the increase of the number of iterations of the modification of the LARS algorithm.*

7.4 Sparse predictor with conformal Lasso

For the reasons exposed above, we focus on the transition points $\lambda_1, \ldots, \lambda_K$ and construct conformal predictors for each of these λ_k. We then propose to select the best conformal predictor among them according to its performance in terms of accuracy (cf. Section 7.2).

Now let us detail the construction of the CoLP for each λ_k. To this end, denote by $X_j = (x_{1,j}, \ldots, x_{n-1,j}, x_{new,j})'$, $j = 1, \ldots, p$ the augmented variable j. Define the augmented matrix $\widetilde{\mathbf{x}} = (x_1, \ldots, x_{n-1}, x_{new})' = (X_1, \ldots, X_p)$ and the augmented response vector $\widetilde{\mathbf{y}} = (y_1, \ldots, y_{n-1}, y)'$ where y is a candidate value for y_{new}. Using the notation introduced in Section 7.3, for the fixed λ_k, we also define the LASSO estimator $\hat{\beta}_k(\widetilde{\mathbf{y}}, \widetilde{\mathbf{x}}_k, s_k)$ from expression (7.4) with the augmented data. From now on, we denote this estimator by $\hat{\beta}_k$. Define $\hat{\mu}_k := \widetilde{\mathbf{x}}_k \hat{\beta}_k$. Moreover, the matrix $\mathbf{H}_k$ will be the $n \times n$ projection matrix onto the subspace generated by $\widetilde{\mathbf{x}}_k$ and $\mathbf{I}$ identity matrix of the same size. For each λ_k, we define a corresponding nonconformity score $\alpha^k = \left(\alpha_1^k, \ldots, \alpha_n^k\right)'$ by:

$$\begin{aligned} \alpha^k(y) &:= |\widetilde{\mathbf{y}} - \hat{\mu}_k| = |\left(\mathbf{I} - \mathbf{H}_k\right)\widetilde{\mathbf{y}} + \frac{\lambda_k}{2}\widetilde{\mathbf{x}}_k \left(\widetilde{\mathbf{x}}_k'\widetilde{\mathbf{x}}_k\right)^{-1} s_k| \\ &= |A_k + B_k\, y|, \end{aligned}$$

where $|\cdot|$ is meant here componentwise and

$$\begin{cases} A_k = (a_1^k, \ldots, a_n^k)' := \left(\mathbf{I} - \mathbf{H}_k\right)(y_1, \ldots, y_{n-1}, 0)' + \frac{\lambda_k}{2}\widetilde{\mathbf{x}}_k \left(\widetilde{\mathbf{x}}_k'\widetilde{\mathbf{x}}_k\right)^{-1} s_k, \\ B_k = (b_1^k, \ldots, b_n^k)' := \left(\mathbf{I} - \mathbf{H}_k\right)(0, \ldots, 0, 1)', \end{cases} \tag{7.5}$$

Note that each component $\alpha_i^k(y)$ is piecewise linear with respect to y. Then the corresponding p-value $p_k(y)$ as defined by (7.1) clearly can change only at points y where the sign of $\alpha_i^k(y) - \alpha_n^k(y)$ changes. Hence, we do not have to evaluate all the possible values of y. We only focus on points y for which the i-th nonconformity measure $\alpha_i^k(y)$ equals $\alpha_n^k(y)$. For this purpose, we define, for each observation $i \in \{1, \ldots, n\}$

$$S_i^k = \left\{ y : \alpha_i^k(y) \geq \alpha_n^k(y) \right\}, \tag{7.6}$$

which corresponds to the range of values y such that the new pair (x_{new}, y) has a better conformity score than the i-th pair (x_i, y_i). Moreover, let l_i^k and u_i^k denote two real defined respectively as

$$l_i^k = \min\{-\frac{a_i^k - a_n^k}{b_i^k - b_n^k}; -\frac{a_i^k + a_n^k}{b_i^k + b_n^k}\}, \qquad \text{and} \qquad u_i^k = \max\{-\frac{a_i^k - a_n^k}{b_i^k - b_n^k}; -\frac{a_i^k + a_n^k}{b_i^k + b_n^k}\}, \tag{7.7}$$

where a_i^k and b_i^k are given by (7.5).

Proposition 7.1. *Let us fix a $k \in \{1, \ldots, K\}$ and an $i \in \{1, \ldots, n-1\}$. Assume that both b_i^k and b_n^k are non-negative. Then*

i) if $b_i^k \neq b_n^k$, we have either $S_i^k = [l_i^k; u_i^k]$ or $S_i^k = (-\infty; l_i^k] \cup [u_i^k; -\infty)$, with l_i^k and u_i^k given by (7.7).

ii) if $b_i^k = b_n^k \neq 0$, then $l_i^k = u_i^k = -\frac{a_i^k + a_n^k}{2b_n^k}$ and we have either $S_i^k = (-\infty; l_i^k]$ or $S_i^k = [l_i^k; -\infty)$. Moreover if $a_i^k = a_n^k$, we have $S_i^k = \mathbb{R}$.

iii) if $b_i^k = b_n^k = 0$, we have either $S_i^k = \mathbb{R}$ or $S_i^k = \emptyset$.

The assumption that all the b_i^k are non-negative does not make loose any generality as one can multiply a_i^k, b_i^k and c_i^k by -1 if $b_i^k < 0$. With this definition of S_i^k, we may rewrite the definition of the conformal predictor as follows:

$$\Gamma_k^\varepsilon = \{y : \sum_{i=1}^n \mathbb{I}(\alpha_i^k(y) \geq \alpha_n^k(y)) \geq n\varepsilon\} = \{y : \sum_{i=1}^n \mathbb{I}(S_i^k)(y) \geq n\varepsilon\}, \tag{7.8}$$

where $\mathbb{I}(\cdot)$ stands for the indicator function. This approach leads to a whole collection of confidence intervals $\Gamma_1^\varepsilon, \ldots, \Gamma_K^\varepsilon$. We propose below a strategy for choosing one particular Γ_k^ε, the performance of which will be studied through numerical simulations.

It is worth mentioning that in view of Vovk [148, Theorem 1] (see also Vovk, Gammerman, and Shafer [150, Proposition 2.3 page 26]), each of predictor Γ_k^ε would have a coverage probability at least equal to $1 - \varepsilon$, if the corresponding value λ_k of the tuning parameter were deterministic. In fact, the following result holds.

Proposition 7.2. *Fix the significance level $\varepsilon \in (0, 1)$ and the tuning parameter $\lambda > 0$. Let $\hat{\beta}_{\lambda,n}(y)$ be the Lasso estimate for the augmented dataset $(\tilde{\mathbf{y}}, \tilde{\mathbf{x}})$ and let us define $\alpha^\lambda(y) = |\tilde{\mathbf{y}} - \tilde{\mathbf{x}}\hat{\beta}_{\lambda,n}(y)|$. Then, the conformal predictor*

$$\Gamma_\lambda^\varepsilon = \Big\{y : \sum_{i=1}^n \mathbb{I}(\alpha_i^\lambda(y) \geq \alpha_n^\lambda(y)) \geq n\varepsilon\Big\},$$

satisfies

$$\mathbb{P}(y_{new} \in \Gamma_k^\varepsilon) \geq 1 - \varepsilon,$$

for any $n \in \mathbb{N}$.

Actually, in the proof of Proposition 7.2 detailed by Vovk [148], one needs the exchangeability of the pairs $(x_1, y_1), \ldots, (x_{n-1}, y_{n-1}), (x_n, y)$ in the definition of the predictor. This property is not fulfilled when the tuning parameter λ is chosen in the set $\{\lambda_1, \ldots, \lambda_K\}$ of Lasso's transition points, since the elements of this set depend only on the first $n - 1$ observations and not on (x_n, y). We believe that under some additional assumptions a

result similar to Proposition 7.2 can be obtained for the predictor Γ_k^ε as well, for each $k = 1, \ldots, K$. This is the topic of an ongoing work. In the present paper, we content ourselves by proposing a data-driven choice of the conformal predictor from the collection of predictors $\{\Gamma_k^\varepsilon; 1 \leq k \leq K\}$ and by exploring its empirical properties.

Remark 7.3. *Of course, one can also apply the well-known sample splitting technique for choosing the values $\lambda_1, \ldots, \lambda_K$ based on a first sample, and then use the methodology described below for selecting the data-driven predictor based on a second sample which is assumed to be independent of the first sample. However, this technique is not attractive from the practical standpoint, that is why we do not develop this approach.*

As discussed above, we believe that all the predictors Γ_k^ϵ share nearly the $1-\varepsilon$ validity property, which is supported by our empirical study. We suggest to select among them the one which has the smallest Lebesgue measure. We denote this confidence set by Γ_{opt}^ϵ, that is

$$\Gamma_{opt}^\varepsilon = \Gamma_\nu^\varepsilon, \qquad \nu = \operatorname*{argmin}_k |\Gamma_k^\varepsilon|. \tag{7.9}$$

In general, since ν is a random variable, the $1-\varepsilon$ validity of all Γ_k^ε would not imply the $1-\varepsilon$ validity of Γ_{opt}^k, but only $1-K\varepsilon$ validity. However, $1-K\varepsilon$ is a worst case majorant obtained by a simple application of the union bound, whereas numerical examples we considered (some of them are reported below) suggest that the validity is much better than $1-K\varepsilon$ and could even be equal to $1-\varepsilon$ when $p \leq n$.

7.5 Implementation

We provide here a three-step algorithm which enables us to easily construct the CoLP. We start in **Step 1** by applying the LASSO modification of the LARS algorithm to the dataset $((x_1, y_1), \ldots, (x_{n-1}, y_{n-1}))$. This step provides all transition points $\lambda_1, \ldots, \lambda_K$, the corresponding design matrices $\mathbf{x}_k$ and sign vectors s_k for $k = 1, \ldots, K$. Then, in **Step 2**, we construct the conformal predictor Γ_k^ε associated to each λ_k. Thanks to Proposition 7.1, for each λ_k, we can construct the sets S_i^k for $i = 1 \ldots, n$ defined by (7.6). We use these sets in order to construct the conformal predictor Γ_k^ε. To do this, we take advantage from the fact that the function $y \mapsto \sum_{i=1}^n \mathbb{I}(S_i^k(y))$ is piecewise constant. Furthermore, the endpoints of the intervals where this function is constant belong to the set of the all endpoints of intervals forming the sets S_i^k. Thus, to determine Γ_k^ε, we sort the set U consisting of the all endpoints of the intervals described in Proposition 1 and include an interval having as endpoints two successive elements of U in Γ_k^ε if the center of this interval belongs to at least $[n\varepsilon]$ sets S_i^k.

Algorithme 1 : Lasso Conformal Predictor

Step 1: Run the LASSO modification of the LARS algorithm on the data set $((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{n-1}, y_{n-1}))$

Step 2: Construct the Conformal Lasso Predictors for each $\lambda_k \in \{\lambda_1, \ldots, \lambda_K\}$

begin

Step 2a: Initialization : Define A_k and B_k as in (7.5). Set $U^k \longleftarrow \emptyset$

Step 2b: Harmonization

for $i = 1$ **to** n **do**

if $b_i^k < 0$ **then**

$a_i^k = -a_i^k$ and $b_i^k = -b_i^k$

end

end

Step 2c: Actualize the set U^k

for $i = 1$ ***to*** n **do**

if $b_i^k \neq b_n^k$ **then**

Add l_i^k and u_i^k (7.7) to U^k

end

if $b_i^k = b_n^k \neq 0$ *and* $a_i^k \neq a_n^k$ **then**

Add $l_i^k = u_i^k$ (7.7) to U^k

end

end

Step 2d: Sort U^k. Let $m \longleftarrow |U^k|$. Then $y_{(0)} \longleftarrow -\infty$ and $y_{(m+1)} \longleftarrow +\infty$

Step 2e: Evaluate N_j^k for $j = 1, \ldots, m$. Initialize $N_j^k \longleftarrow 0$. Then actualize

for $i = 1$ **to** n **do**

for $j = 1$ **to** m **do**

if $|a_i^k + b_i^k\, y| \geq |a_n^k + b_n^k\, y|$ *for* $y \in (y_{(j)}, y_{(j+1)})$ **then**

Increment $N_j^k = N_j^k + 1$

end

end

end

Step 2f: For a fixed threshold $\varepsilon > 0$, output the conformal predictor

$$\Gamma_k^\varepsilon = \cup_{j:\frac{N_j}{n} > \varepsilon} [y_{(j)}, y_{(j+1)}]$$

end

Step 3: Output the Conformal Lasso Predictor Γ_{opt}^ε as the smallest (w.r.t. their Lebesgue measure) confidence set among the constructed conformal predictors

Finally, in a **Step 3**, we provide the CoLP, says Γ_{opt}^ε, which is defined as the smallest confidence set, according to its Lebesgue measure, among the constructed conformal predictors Γ_k^ε, $k = 1, \ldots, K$. According to Proposition 7.2, each Γ_k^ε is valid. Moreover the criterion for choosing the CoLP is adapted to variable selection as conformal predictors constructed here for different values of λ_k, $k = 1, \ldots, K$ bring into play different variables. This is illustrated in Figure 7.1 (left) where we constructed the conformal predictors when $n = 300$. One can observe that all the conformal predictors are valid since they contain

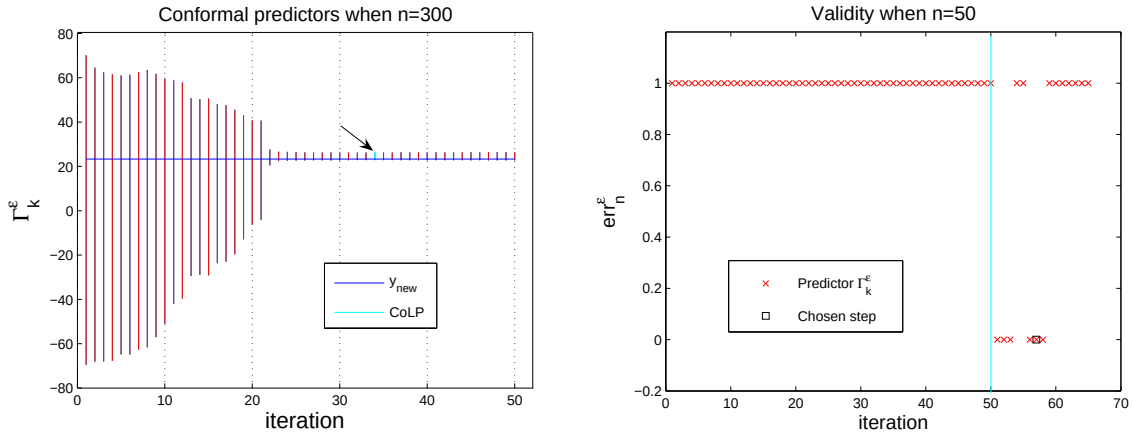

Figure 7.1: *Left*: Conformal predictors Γ_k^ε evolution through the iterations of the LASSO modification of the LARS algorithm when $n = 300$ (the first iteration corresponds to λ_{max} and the last one corresponds to λ_{min}). The CoLP is drawn in cyan and corresponds to the 34-th iteration. The horizontal blue line corresponds to the value of y_{new}. *Right*: Validity analysis (err_n^ε) of the conformal predictors Γ_k^ε through the iterations of the LASSO modification of the LARS algorithm when $n = 50$ (the first iteration corresponds to λ_{max} and the last one corresponds to λ_{min}). The CoLP is marked by a black square and corresponds to the 57-th iteration. The vertical line represents a separation between a stable and an unstable zone.

the true value of the label y_{new}. Hence our construction is suitable when the sample size is larger than the number of variables (i.e., $n > p$) but may be not appropriated when $p \geq n$. Figure 7.1 (right) shows an example where almost all the constructed conformal predictors Γ_k^ε, $k = 1, \ldots, K$, using the above algorithm are valid. Only six are not. One of them is the selected CoLP (iteration 57 in Figure 7.1 (right)) which corresponds to the smallest predictor. In such cases ($p \geq n$), a correction can be made and other choices for the accuracy measure are possible. We discuss this criterion in Section 7.7. Let us add that we only illustrated the validity of the conformal predictors in Figure 7.1 (right) as the unstable zone (on the right side of the vertical line) makes the representation hard to be analyzed. More details are given in Section 7.7.

Remark 7.4. *In **Step 1** of Algorithm 1, we use the LARS algorithm for its ability to generate a small number of tuning parameter values of interest. It is an important aspect as it considerably reduces the computational cost. On-line versions could be implemented by plugging in an on-line version of the LASSO solution as in the paper of Garrigues and El Ghaoui [70]. The analysis of such on-line versions is the object of work under progress.*

7.6 Extension to others procedures

In this section we generalize the construction of the confidence predictor to a family of estimators which includes selection-type procedures as the Elastic-Net (Zou and Hastie [167]) and the Smooth-Lasso (Hebiri [78]). As for CoLP (Section 7.4), we are interested in

two properties of estimators: the *piecewise linearity w.r.t. the response y* (to easily compute the nonconformity scores α_i, $i = 1, \ldots, n$), and the *piecewise linearity w.r.t the tuning parameter λ (Rosset and Zhu [126])* (to reduce computational effort by using a modification of the LARS algorithm).
We use the same notation as in Section 7.3 for the LASSO estimator. Set $\hat{\beta}$ to be an estimator of the regression vector β based on $\mathbf{x}$ and $\mathbf{y}$. Let also s be the sign vector of the estimator $\hat{\beta}$. On the other hand, using the notation in Section 7.4, we set $\hat{\mu} = \widetilde{\mathbf{x}}\hat{\beta}$ where this time $\hat{\beta}$ is based on the augmented dataset $\widetilde{\mathbf{x}}$ and $\widetilde{\mathbf{y}}$.

Assumption 1. *The estimator $\hat{\mu}$ can be written as:*

$$\hat{\mu} = u(\widetilde{\mathbf{x}}, s)\widetilde{\mathbf{y}} + v(\widetilde{\mathbf{x}}, s), \tag{7.10}$$

where $u(\cdot)$ and $v(\cdot)$ are piecewise constant functions w.r.t. $\widetilde{\mathbf{y}}$.

As soon as Assumption 1 holds, we can construct a conformal predictor corresponding to the estimator $\hat{\mu}$. Then many estimators can be considered. The CoLP and CoRP obviously belong to this class of predictors and we introduce here the Conformal Elastic Net Predictor (CENeP) which is a conformal predictor constructed based on the Elastic-Net modification of the LARS instead of the LASSO one (**Step1** in Algorithm 1). This predictor is defined by $u(\widetilde{\mathbf{x}}, s) = \widetilde{\mathbf{x}}_k(\widetilde{\mathbf{x}}_k'\widetilde{\mathbf{x}}_k + \mu_k\mathbf{I}_k)^{-1}\widetilde{\mathbf{x}}_k'$ and $v(\widetilde{\mathbf{x}}, s) = -\lambda_k\widetilde{\mathbf{x}}_k(\widetilde{\mathbf{x}}_k'\widetilde{\mathbf{x}}_k)^{-1}s_k$ where λ_k and μ_k correspond respectively to the LASSO and Ridge tuning parameters in the definition of the Elastic-Net estimator and $\mathbf{I}_k$ is the $|\mathcal{A}_k| \times |\mathcal{A}_k|$ identity matrix (Zou and Hastie [167]). In the same way, we can define the Conformal Smooth Lasso Predictor (CoSmoLaP) based on a Smooth-Lasso modification of the LARS algorithm (Hebiri [78]). Here $u(\widetilde{\mathbf{x}}, s) = \widetilde{\mathbf{x}}_k(\widetilde{\mathbf{x}}_k'\widetilde{\mathbf{x}}_k + \mu_k\mathbf{J}_k)^{-1}\widetilde{\mathbf{x}}_k'$ and $v(\widetilde{\mathbf{x}}, s) = -\lambda_k\widetilde{\mathbf{x}}_k(\widetilde{\mathbf{x}}_k'\widetilde{\mathbf{x}}_k)^{-1}s_k$. The difference between the CoSmoLaP definition the CENeP one is the identity matrix $\mathbf{I}_k$ which is replaced by the $|\mathcal{A}_k| \times |\mathcal{A}_k|$ matrix $\mathbf{J}_k$ whose components are such that $(\mathbf{J}_k)_{i,i} = 1$ if $i = 1$ or $i = |\mathcal{A}_k|$ and $(\mathbf{J}_k)_{i,i} = 2$ otherwise. Moreover for $(i, j) \in \{1, \ldots, \mathcal{A}_k\}^2$ with $i \neq j$, we have $(\mathbf{J}_k)_{i,j} = -1$ if $|i - j| = 1$ and zero otherwise. Note that the definition of $\mathbf{J}_k$ makes the CoSmoLaP more appropriated to model with successive correlation between successive variables.

As for CoLP, we can define the nonconformity score of an expected label y associated to the estimator $\hat{\mu}$ as follows:

$$\begin{aligned}(\alpha_1(y), \ldots, \alpha_n(y))' &:= |\widetilde{\mathbf{y}} - \hat{\mu}| \\ &= |\left(\mathbf{I} - u(\widetilde{\mathbf{x}}, s)\right)\widetilde{\mathbf{y}} - v(\widetilde{\mathbf{x}}, s)| \\ &= |A + B\,y|,\end{aligned}$$

with

$$\begin{cases} A = (a_1, \ldots, a_n)' := (\mathbf{I} - u(\widetilde{\mathbf{x}}, s))\,(y_1, \ldots, y_{n-1}, 0)' - v(\widetilde{\mathbf{x}}, s), \\ B = (b_1, \ldots, b_n)' := (\mathbf{I} - u(\widetilde{\mathbf{x}}, s))\,(0, \ldots, 0, 1)', \end{cases}$$

and $\mathbf{I}$ is the $n \times n$ identity matrix. The quantities A and B are the analogues of A_k and B_k respectively, when we considered the CoLP at the transition point λ_k, $k = 1, \ldots, K$. Then replacing A_k and B_k by respectively A and B in **Step 2.a** of Algorithm 1, we obtain the conformal predictors associated to the estimator $\hat{\mu}$.

Note that the dependency in the tuning parameter, noted λ, can be included in $u(\widetilde{\mathbf{x}}, s)$ (as for CoRP) or $v(\widetilde{\mathbf{x}}, s)$ or in both of them (as for the CoLP). For instance, in the construction of the CoLP, this dependency is underlined in the matrix $\widetilde{\mathbf{x}}_k$ and the sign vector s_k as they were computed by the LARS algorithm for a specified value λ_k of the tuning parameter λ.

Computational cost of the construction of conformal predictors has also to be considered. Three main points interfere. First, one run of the LARS algorithm requires the same cost as the computation of the least square estimation. Then we have to consider the number of conformal predictors we have to construct: each value of the tuning parameter λ provides a conformal predictor Γ_λ using the algorithm described in Section 7.5. The final conformal predictor Γ_{opt} is then the one with the minimal length. As for the CoRP, the main problem is: how many λ's do we have to test? One way is to use a grid of value for λ which lets open the problem of the choice of the grid and the window of this grid.
On the other hand, we saw how the LARS algorithm permits to reduce considerably the number of tuning parameters to be considered. Indeed the grid of tuning parameters values is directly described by the transition points $\lambda_1, \ldots, \lambda_K$ obtained from the run of the LARS algorithm. Finally, let us consider *the construction of the conformal predictor itself:* this point has been treated in Vovk et al. Vovk, Gammerman, and Shafer [150, Chapter 2.3 and 4.1]. It turns out that sparse conformal predictors and the CoLP requires computation time $\mathcal{O}(n^2)$ and can be reduced to $\mathcal{O}(n \log(n))$.

7.7 Experimental Results

In the section we present the experimental performances of the Sparse Conformal Predictors (SCP) w.r.t. their validity, their accuracy and also their selection power. As benchmark, we use the CoRP[3] for its validity and accuracy and the original LASSO and Elastic-Net

[3]We construct the CoRP associated to same tuning parameters as the CoLP (i.e., the transition points λ_k observed in Section 7.5). Note that the performance would not be inflected as conformal predictors according to this method are almost embedded and changes sensitively while the tuning parameter varies. See Vovk, Gammerman, and Shafer [150, page 39] for more details.

estimators for their selection[4] power.

We consider three SCPs: the Conformal Lasso Predictor (CoLP was introduced in Sections 7.4 and 7.5) and the Conformal Elastic Net Predictor (CENeP was described in Section 7.6). The last SCP called Conformal Ridge Lasso Predictor (CoRLaP) is a mix of the CoRP and the CoLP. To construct the CoRLaP, we use the variables selected by the LASSO modification of the LARS algorithm (**Step 1** in Algorithm 1 described in Section 7.5). Then we use these variables to construct a CoRP. This conformal predictor can be seen as a restricted CoRP. All conformal predictors are constructed with confidence level $1 - \varepsilon = 90\%$.

7.7.1 Simulation Experiments

We consider four simulations from the linear regression model

$$y = \mathbf{X}'\beta + \sigma\xi, \quad \xi \sim \mathcal{N}(0,1),\ \mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_{50})' \in \mathbb{R}^{50},$$

with $\beta \in \mathbb{R}^{50}$. Hence $p = 50$ through the simulations. Noise level σ and the sample size n are let free. They will be specified during experiments.

Example **(a)** $[n/\sigma]$: *Very Sparse and Correlated.* Here only β_1 is nonzero and equals 5. Moreover, the design correlations matrix Σ is described by $\Sigma_{j,k} = \exp(-|j-k|)$ for $(j,k) \in \{15, \ldots, 35\}^2$ and $\Sigma_{j,k} = \mathbb{I}(j = k)$ otherwise where $\mathbb{I}(\cdot)$ is the indicator function.

Example **(b)** $[n/\sigma]$: *Sparse and Correlated.* The correlations are defined as in Example (a) and the regression vector is given by $\beta_j = -5 + 0.2j$ for $j = 1, \ldots, 5$; $\beta_j = 4 + 0.2j$ for $j = 10, \ldots, 25$ and zero otherwise.

Example **(c)** $[n/\sigma]$: *Sparse and Highly correlated.* We have $\beta_j = 5$ for $j \in \{1, \ldots, 15\}$ and zero otherwise. We construct three groups of correlated variables: $\Sigma_{j,k}$ is close to 1 when (j,k) belongs to $\{1, \ldots, 5\}^2$, $\{6, \ldots, 10\}^2$ or $\{11, \ldots, 15\}^2$; $\Sigma_{j,k} = 1$ for $(j,k) \in \{16, \ldots, p\}^2$ if $j = k$ and zero otherwise.

Example **(d)** $[n/\sigma]$: *Non Sparse and correlated.* Here $\beta_j = 3 + 0.2j$ for $j \in \{1, \ldots, p\}$ and the correlations are described by $\Sigma_{j,k} = \exp(-|j-k|)$ for $(j,k) \in \{1, \ldots, p\}^2$.

[4] We use a BIC-type criterion to select the optimal tuning parameter. Such a criterion is adapted to variable selection.

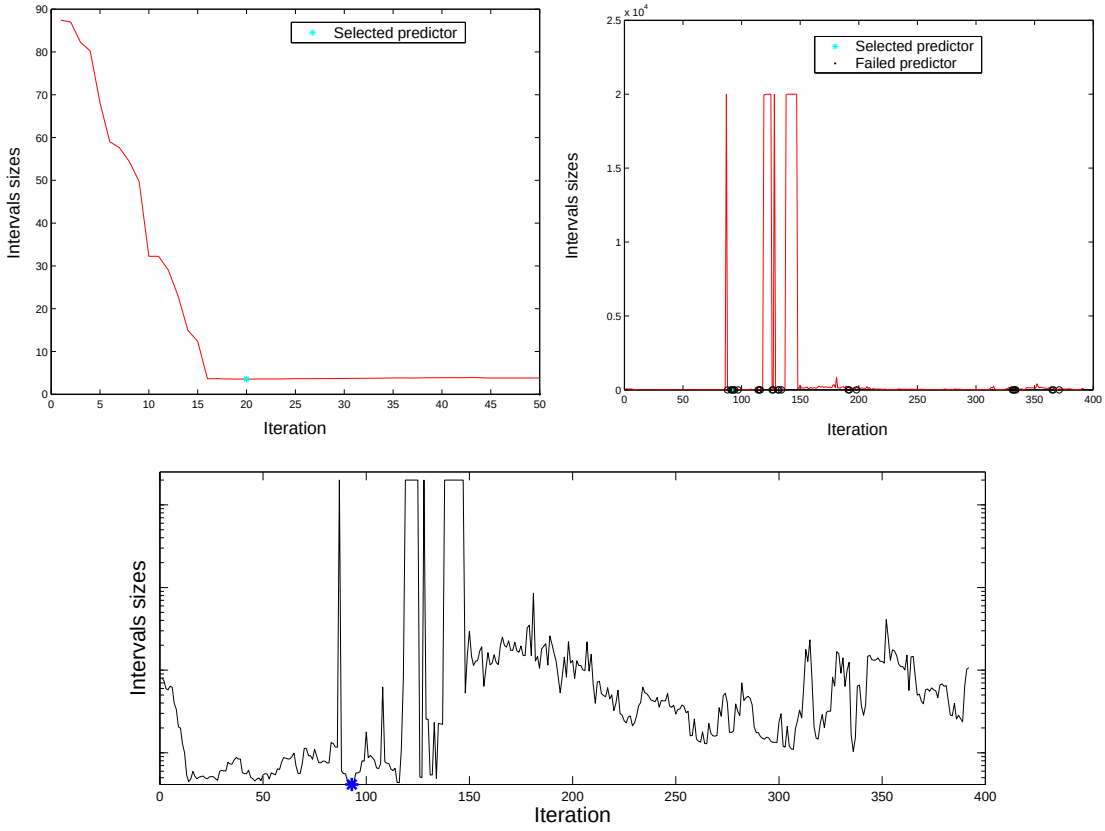

Figure 7.2: Analysis of conformal predictors length (y-axis) through the LASSO modification of the LARS algorithm iterations (x-axis: the first iteration corresponds to λ_{max} and the last one corresponds to λ_{min}) in Example (c)[300/1] (top left) and in Example (c)[50/1] (top right). The iteration associated to the CoLP is marked by a blue star. Predictors which are non valid are marked by a black circle. The panel of bottom shows the lengths of intervals in a logarithmic scale.

We consider separately the three points of interest: accuracy, validity and selection.

Accuracy. First of all, let us consider the length of the predictors Γ^{ε}_k, $k = 1, \ldots, K$ obtained at the end of **Step 2** in Algorithm 1 described in Section 7.5. We remind that each of these predictors is associated to an iteration of a modification of the LARS algorithm, that is the transition points λ_k, $k = 1, \ldots, K$. Figure 7.2 illustrates the predictors lengths for the construction of the CoLP, when applied to Example (c)$[n/1]$ with $n = 300$ and $n = 50$. When $n = 300$, we note that the length of the Γ^{ε}_ks sensitively changes from one iteration to the following and that the larger predictor has a reasonable length compared to the smallest one (about 10 times larger). Then the construction is stable. We also observe that in the neighborhood of the optimal iteration (that is iteration 20), the conformal predictors have approximately the same size. Such an observation can also be made when we take a look at Figure 7.1 (left)

Table 7.1: Validity frequencies [with precision ±95%] of the CoRP, CoLP, CoRLaP, CENeP, the Early-Stopped CoLP and the 2-PN CoLP based on 1000 replications.

Example	σ	CoRP	CoLP	CoRLaP	CENeP
(a)[300/σ]	1	0.897± 0.019	0.876± 0.020	0.854± 0.022	0.878± 0.020
	7	0.894± 0.019	0.908± 0.018	0.894± 0.019	0.899± 0.019
	15	0.893± 0.019	0.893 ± 0.019	0.879± 0.020	0.887± 0.020
(b)[300/σ]	1	0.901± 0.018	0.875± 0.020	0.869± 0.021	0.874± 0.021
(c)[300/σ]	1	0.900± 0.019	0.900± 0.019	0.891± 0.019	0.901± 0.018
(d)[300/σ]	1	0.892± 0.019	0.895± 0.019	0.895± 0.019	0.895± 0.019
(a)[50/σ]	3	0.887± 0.020	0.668± 0.029	0.414± 0.030	0.789± 0.025
(a)[20/σ]	3	0.865± 0.021	0.596± 0.030	0.304± 0.028	0.685± 0.029
Example	σ	CoRP	CoLP	Stopped-CoLP	2-PN-CoLP
(a)[50/σ]	7	0.853± 0.022	0.620± 0.030	0.815± 0.024	0.881± 0.020
(b)[50/σ]	1	0.875± 0.020	0.558± 0.031	0.814± 0.024	0.907 ± 0.018
(c)[20/σ]	15	0.875± 0.020	0.608± 0.030	0.769± 0.026	0.893± 0.019
(d)[20/σ]	1	0.900± 0.019	0.602± 0.030	0.793± 0.025	0.892± 0.019

when applied to Example (b)[300/1]. On the other hand, when $n = 50$, it appears that the predictors length grows drastically at some iteration (around iteration 85). We even can not compare the lengths of the bigger and smaller predictors (more than 10^4 times larger). In the same time, it seems that the construction becomes unstable as violent variations often happen after this iteration 85. We will consider in the next point the validity of these predictors. However let us mention that in Example (c)[50/1], the CoLP which is the smallest Γ_k^ε and then the selected predictor is not valid (in Figure 7.2 (right), the selected predictor at iteration 93 is not valid). This aspect can also be observed in Figure 7.1 (right) (the graph corresponds to Example (b)[50/1]) where the selected CoLP at iteration 57 is not valid. Similar violent variations of the corresponding predictors lengths would have been observed after iteration 49 if we have provided a graph as Figure 7.2 (right).

Validity. Now, we consider the validity of the selected predictors (cf. **Step 3** in Algorithm 1). As shown in Table 7.1, we observe that variations on the noise level, the variables correlations and the sparsity of the model do to not perturb the validity whereas the sample size relatively to the dimension p does. When $n = 300 > p$, all the procedures seem to be quite similar and produce good predictors. In the other cases, i.e., when $n = p = 50$ and $n = 20 < p$, the selected confidence predictors have worst performance than expected (validity with smaller proportion than $1-\varepsilon = 90\%$). Moreover, Sparse Confidence Predictors perform worst than the CoRP as observed in Table 7.1. As pointed in the accuracy part, one explication can be observed in Fig-

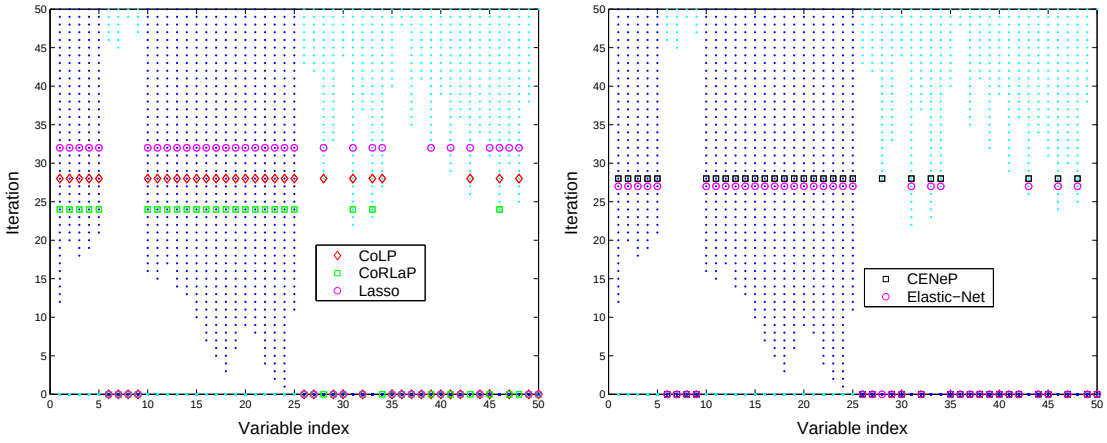

Figure 7.3: Variable selection analysis for the CoLP, the CoRLaP and the CENeP in Example (b)[300/1] (variables 1 to 5 and 10 to 25 are relevant; see variables in dark blue on the plot). On the left, we consider the CoLP and the CoRLaP selected variables (x-axis) with respect to the LASSO modification of the LARS algorithm iterations (y-axis: the first iteration corresponds to λ_{max} and the last one corresponds to λ_{min}). On the right, we consider the CENeP selected variables (x-axis) with respect to the Elastic-Net modification of the LARS algorithm iterations (y-axis: the first iteration corresponds to λ_{max} and the last one corresponds to λ_{min}). The selected iteration is marked by red diamonds for the CoLP, green squares for CoRLaP and black squares for the CENeP.

ure 7.2 as the selected predictor which also is not valid (iteration 93) corresponds to an iteration in the unstable zone (that is, after iteration 85). Then in order to reduce the gap between SCP and CoRP in the cases $p \geq n$, we suggest to modify the selection criterion in **Step 3** in two ways. i) *Early Stopping CoLP:* do not consider (and do not construct) all the conformal predictors Γ^{ε}_{k}. Stop the construction of the predictors Γ^{ε}_{k} as soon as the length of Γ^{ε}_{k} (predictor at iteration k) has a length at least 10 times larger than $\Gamma^{\varepsilon}_{k-1}$; ii) *N Previous Neighbors CoLP:* we can enforce the Early Stopping rule by considering as final predictor: $\Gamma^{\varepsilon}_{opt} = \bigcup_{j:\, 0 \leq k-j < N} \Gamma^{\varepsilon}_{j}$, where k is the index of the (selected) smallest predictor and N is the number of neighbors we consider. Note that this method does not alter selection properties as Γ^{ε}_{k} is usually constructed with more variables than Γ^{ε}_{j}, $j < k$. It further does not alter a lot the accuracy as the Early Stopping rule ensures that we are in stable zone (cf. Figure 7.2 (right) and Figure 7.1 (right)). Table 7.1 sums up the performances of the early-stopped CoLP and the 2-PN CoLP in term of validity. We observe the good adaptation of both methods to the case $p = n$ and we remark that 2-PN CoLP nicely produce valid predictor even in the case $p > n$. This improvement in the term of validity can also be illustrated by Figure 7.1 (right) where we observe that in Example (b)[50/1], the early-stopped CoLP is valid whereas the original CoLP is not.

Selection. The selection ability of Sparse Conformal Predictors is here in concern. First,

note that the selected variables in SCPs are directly linked to the selection ordering through the iterations of the LASSO or Elastic-Net modification of the LARS algorithm. Then, if the used modification of the LARS algorithm fails to recover the true model, we can not hope to get a predictor which contains only the true variables. Figure 7.3 illustrates the evolution of the variable selection of CoLP, CoRLaP and the LASSO on one hand and the CENeP and the Elastic-Net on the other hand, in Example (b)[300/1]. It turns out that CoLP and CENeP select larger model that expected (that is, some noise variables are selected), as the LASSO and the Elastic-Net do. Moreover CoRLaP uses to select a smaller subset of variables than the CoLP. Then it often produces a better variable selection performance than the other methods. It often provides closer model to the true one. Compared to the LASSO, it seems that the CoLP and the CoRLaP perform better in this example. However, we can not conclude the superiority of the CoLP on the LASSO in term of variable selection. A similar conclusion can be given when we compare the CENeP and the Elastic-Net. Nevertheless, the CENeP seems to select little larger models than the Elastic-Net. Finally, analogously to the superiority of the Elastic-Net compared to the LASSO, we can remark that the CENeP manages to have better selection performances compared to the CoLP and the CoRLaP when a group structure may exist between different variables (for instance in Example (d)$[n/\sigma]$). This is due to the LASSO modification of the LARS algorithm which uses to select some noise variables before relevant ones in such cases.

7.7.2 Real data

We applied SCPs on 150 randomly permutations of the House Boston dataset[5], in which we randomly choose one row to be the new pair (x_{new}, y_{new}). The original dataset consists of 506 observations with 13 variables. When we consider variable selection, we note that almost all SCPs are constructed without the variable $X_7 = (x_{1,7}, \ldots, x_{505,7})$. This variable is selected with frequencies lower than 3%. The CoRLaP also does not consider the variable X_3 as relevant with a frequence equal to 17%. Conforming to Section 7.7.1, we would better consider X_3 irrelevant as the CoRLaP uses to produce better performance when variable selection is in concern. Then we conclude that the proportion of non-retail business acres per town and the proportion of owner-occupied units built prior to 1940 do not interfere in the value of owner-occupied homes. We also can notice that variable selection sligtly improved accuracy of conformal predictors in all presented experiments. Here, we can for instance remark that the median lengths of the CoLP, the CoRLaP and the CENeP are

[5]The data and their description are available at http://archive.ics.uci.edu/ml/datasets/Housing.

respectively 13.61, 13.50 and 13.58, whereas CoRP length is 14.45.

7.8 Conclusion

We presented Sparse Conformal Predictors, a family of ℓ_1 regularized conformal predictors. We focused on LASSO and Elastic-Net versions of these Sparse Conformal Predictors. We illustrated their performance in term of accuracy, validity and variable selection. We concluded that such Sparse Conformal Predictors are valid and nicely exploit the sparsity of the model when the sample size is larger than the the number of variables (i.e, when $n > p$). We also provided a way to adopt these sparse predictors to the case $p \geq n$ through a pair of rules we called Early Stopping and N Previous Neighbors rules.
Several extensions of this work can be explored such as the construction of SCP with Adaptive LASSO (Zou [166]) and they will be investigated in future work.

Annexe A

Pénalisations non convexes

Dans la section 2.1.3 du chapitre 2, nous avons énoncé des méthodes d'estimation par pénalisation ℓ_0. Celles-ci sont consistantes en *sélection* ou en *prédiction*, sans hypothèses sur la matrice de Gram $\Psi^n = \frac{X'X}{n}$. En revanche, elles sont difficilement calculables en pratique sans imposer de lourdes hypothèses sur le modèle (variables ordonnées, ensembles de sparsité emboîtés, petit nombre d'ensembles de sparsité considérés,etc.). Dans le section 2.2 du chapitre 2, nous avons présenté le Lasso (2.15), une méthode de sélection de variables définie avec une pénalité ℓ_1. Il s'agit d'une *convexification* du critère ℓ_0, grâce à quoi il existe des algorithmes fournissant rapidement une solution (cf. Annexe B). L'estimateur Lasso a de surcroît de bonnes performances théoriques. Toutefois, l'ensemble des résultats établis pour cet estimateur nécessitent des hypothèses sur la matrice de Gram Ψ^n.

Dans le cadre des problèmes de régression linéaire avec un paramètre de régression sparse, nous pouvons également considérer des méthodes d'estimation intermédiaires ayant pour objectif de réduire ou supprimer l'hypothèse sur la matrice de Gram imposée par le Lasso, tout en proposant une implémentation en un temps raisonnable.

Nous nous proposons l'étude d'estimateurs pénalisés de la forme :

$$\hat{\beta}(\lambda_n) \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|Y - X\beta\|_n^2 + \operatorname{pen}(\beta, n), \tag{A.1}$$

où la fonction de pénalité $\operatorname{pen}(\beta, n)$ sera précisée selon la méthode considérée dans les prochaines sections.

A.1 Poids exponentiels

Les estimateurs à poids exponentiels (EPE) ont largement été étudiés ces dernières années (Alquier [5], Audibert [9], Catoni [38, 37, 39], Juditsky, Rigollet, et Tsybakov [87], Leung et Barron [99], Nemirovski [116], Tsybakov [139], Yang [158] et Yuditskiĭ, Nazin, Tsybakov, et Vayatis [161]). Nous nous intéressons ici, aux EPE adaptés au contexte de sparsité.

Dans leurs travaux, Dalalyan et Tsybakov [46, 47, 48] montrent que les EPE, bien que non sparses, exploitent la sparsité du modèle et vérifient des IS en *prédiction*. Ces IS s'obtiennent en choisissant convenablement la loi a priori du paramètre β^* à estimer. Pour capturer la sparsité du modèle, cette loi a priori doit avoir une queue de distribution lourde. En outre, les auteurs considèrent la loi de Student à 3 degrés de liberté. Formellement, l'EPE considéré, noté $\hat{\beta}^{EPE}$, est tel que :

$$\hat{\beta}_j^{EPE} = \int_{\mathbb{R}^p} \beta_j \overline{\pi}(d\beta) \quad j = 1, \dots, p, \tag{A.2}$$

où $\overline{\pi}$ est la densité a posteriori définie par :

$$q(d\beta) = \frac{\exp\left(-n\|Y - X\beta\|_n^2/\gamma\right) f(\beta) d\beta}{\int_{\mathbb{R}^p} \exp\left(-n\|Y - Xb\|_n^2/\gamma\right) f(b) db},$$

avec $\gamma > 0$ un paramètre de température, et f la densité de la loi a priori de β^*. La densité f est alors choisie de telle sorte que $f(\beta) = \prod_{j=1}^p \tau^{-1} f_0(\beta_j/\tau)$ où f_0 est la densité de la loi de Student t_3 à 3 degrés de liberté et τ est un paramètre positif de dispersion. Avec un tel choix, les auteurs fournissent des IS sans hypothèses sur la matrice de Gram Ψ^n. Dans leurs derniers travaux, Dalalyan et Tsybakov [48] proposent également une méthode de calcul de l'estimateur en un temps raisonnable par le biais d'une diffusion de Langevin. Pour définir cet estimateur, les auteurs décrivent explicitement tous les paramètres inconnus.

A.2 Pénalité de type Kullback et entropie

Les méthodes considérées dans la section A.1 exploitent la sparsité de β^* sans pour autant que l'estimateur soit sparse. En outre, l'estimateur $\hat{\beta}^{EPE}$ défini par (A.2) peut être vu comme le minimiseur de l'estimateur des Moindres Carrés avec une pénalisation type Kullback. Il peut s'écrire :

$$\hat{\beta}^{EPE}(\lambda_n) \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_n^2 + \lambda_n \sum_{j=1}^p \log\left(1 + \frac{\beta_j}{\tau}\right).$$

L'EPE n'est pas la solution d'un problème convexe, ce qui rend celui-ci difficile à résoudre sous cette forme.

Un autre estimateur obtenu par la minimisation d'un critère du type (A.1) avec une pénalité en lien avec la précédente est étudié par Koltchinskii [92]. L'estimateur est défini par le critère (A.1) où la pénalité est de type entropique, elle s'écrit :

$$\text{pen}(\beta, n) = -H(\beta) = \sum_{j=1}^{p} \beta_j \log(\beta_j),$$

et est donc convexe. Dans cette expression, $H(\beta)$ désigne l'entropie de β. Koltchinskii [92] établit, sous des hypothèses sur la matrice de Gram Ψ^n, des IS, et met en évidence le fait qu'un estimateur non sparse peut atteindre des vitesses de convergence presque optimales, i.e. égales à la vitesse optimale à un terme logarithmique près.

A.3 Pénalités presque sans biais

Nous présentons dans ce paragraphe deux pénalités définies par morceaux, permettant d'améliorer l'estimateur Lasso en réduisant son biais. L'approche consiste à combiner au mieux les avantages de l'estimateur Lasso et ceux des moindres carrés.

La première pénalité est étudiée par Fan [62] qui suggère d'enlever le biais nuisible du Lasso dans le cas de l'estimation par ondelettes (dans ce cas le Lasso devient équivalent au seuillage doux (cf. Section 2.2)). L'analyse est ensuite affinée par Fan et Li [63] et par Fan et Peng [64] qui introduisent et étudient en détails la pénalité *Smoothly Clipped Absolute Deviation Penalty* (SCAD). La deuxième pénalité est introduite par Zhang [162], qui apporte quelques améliorations à la pénalité SCAD et définit la pénalité *Minimax Concave Penalty* (MCP).

Nous présentons à présent plus en détails les améliorations sur l'analyse de la pénalité apportées par les auteurs précités à l'estimateur Lasso. Supposons que la pénalité $\text{pen}(\beta, n)$ dans le critère (A.1), soit définie de la façon suivante :

$$\text{pen}(\beta, n) = \lambda_n \sum_{j=1}^{p} p(\beta_j),$$

où $p(\cdot)$ est une fonction positive croissante sur $[0, +\infty[$. Les propriétés sur lesquelles les auteurs se focalisent sont :

1. *Sélection.* Pour obtenir des estimateurs sparses, la pénalité doit être non différentiable à l'origine. La pénalité utilisée ici au voisinage de 0, est la pénalité ℓ_1 ; c'est-à-dire $p(\beta_j) = |\beta_j|$ lorsque $|\beta_j|$ est petit.
2. *Presque sans biais.* L'introduction d'une pénalité induit généralement un biais d'estimation nuisible (Fan et Li [63]). Pour les méthodes que nous considérons ici, ce biais

est supprimé pour les grandes valeurs de $|\beta_j|$. Autrement dit, $p(\beta_j) = 0$ quand $|\beta_j|$ dépasse un certain seuil. Pour ces valeurs de $|\beta_j|$, l'estimateur devient semblable à l'estimateur des moindres carrés.

3. *Continuité.* Il est préférable que la pénalité construite soit continue (éventuellement par morceaux). Cette propriété assure une stabilité de la solution pour de petites variations du paramètre β_j.

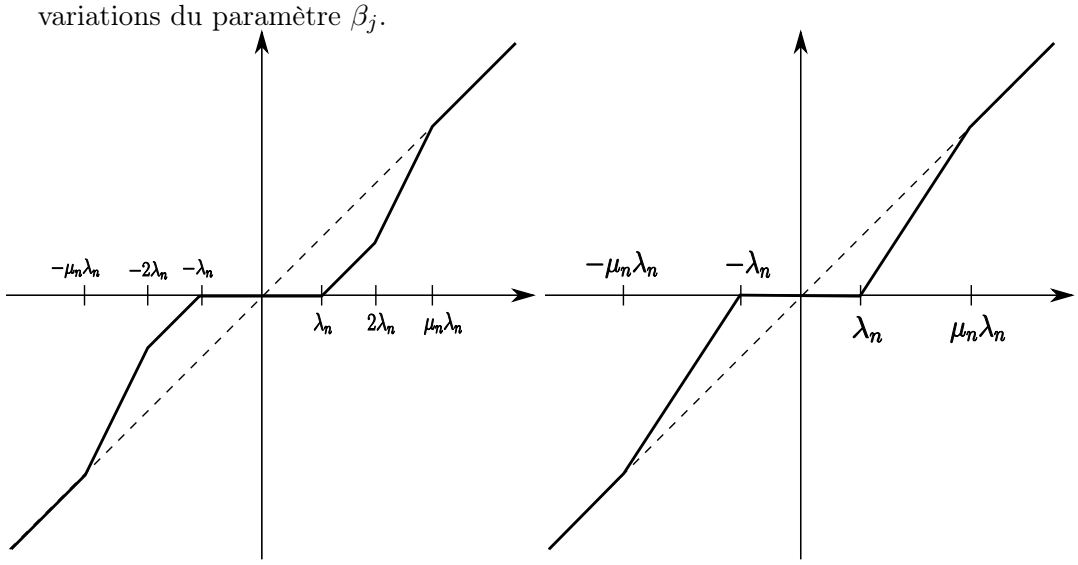

FIG. A.1 – Fonctions de seuillage SCAD (à gauche) et de seuillage MCP (à droite).

Répondre de ces trois propriétés de sélection, de continuité et de sans biais impose aux pénalités SCAD et MCP de ne pas être convexes (voir Figure A.1 - gauche). La pénalité de l'estimateur SCAD est définie par :

$$p(\beta_j) = \int_0^{\beta_j} \min\left\{1, \frac{(\mu_n - x/\lambda_n)_+}{\mu_n - 1}\right\} dx, \quad \mu_n > 2,$$

où μ_n est un deuxième paramètre de la méthode. Les propriétés théoriques comme la consistance en sélection de variables de cet estimateur ont été considérées lorsque $p \leq n$ par Fan et Peng [64]. De plus, des algorithmes itératifs ont été développés par Hunter et Li [83] et Zou et Li [168] pour approcher la solution SCAD. Notons que, le critère SCAD n'étant pas convexe, la solution n'est pas unique.

Par la suite Zhang [162] précise le travail de Fan et Li [63] en suggérant une pénalité qui partage les mêmes propriétés. Il définit ainsi la pénalité MCP :

$$p(\beta_j) = \int_0^{\beta_j} \left(1 - \frac{x}{\mu_n\lambda_n}\right)_+ dx,$$

où $\mu_n > 0$ est un paramètre de régularisation qui contrôle la complexité algorithmique du critère. En effet, ce paramètre est directement lié à la convexité de la pénalité. De par sa

définition, la pénalité MCP réduit le nombre de points où celle-ci change de linéarité (un seul point de changement de linéarité alors que la pénalité SCAD en a deux (cf. Figure A.1 - droite)). Ceci réduit davantage le biais associé à l'estimateur MCP, ainsi que la complexité algorithmique de la méthode.
Pour cette étude de la pénalité MCP, Zhang [162] propose un algorithme, utilisable pour $p \geq n$ sous l'hypothèse de sparsité. Son algorithme calcule les solutions MCP de manière itérative en ajoutant une variable à chaque étape. L'algorithme est rapide lors des premières étapes : le critère est convexe lorsque l'ensemble actif $\hat{\mathcal{A}}$ associé à l'estimateur MCP est petit (ceci s'explique par un bon choix de γ et par le fait que la convexité de la perte "domine" la non-convexité de la pénalité) ; il l'est de moins en moins à mesure que le nombre de variables actives augmente. L'auteur établit également des résultats de consistance (en *prédiction/estimation/sélection*) en grande dimension sous des hypothèses sur la matrice de Gram Ψ^n proche de la condition de Designs Incohérents donnée par (2.28).

A.4 Pénalités Bridge

Dans ce paragraphe, nous considérons les estimateurs (A.1), définis par le biais de la pénalité *Bridge* :

$$\text{pen}(\beta, n) = \lambda_n \sum_{j=1}^{p} |\beta_j|^\gamma,$$

où γ appartient à l'intervalle $]0, 2]$. Cette pénalité, présentée par Frank et Friedman [66] et Fu [69], admet pour cas particulier la pénalité ridge lorsque $\gamma = 2$ et la pénalité Lasso pour $\gamma = 1$. Il est clair que pour $\gamma \geq 1$ la pénalité est convexe alors que pour $\gamma \leq 1$, la pénalité offre à l'estimateur la propriété de sparsité (puisque dans ce cas elle devient non différentiable en 0). Ainsi, l'estimateur défini avec la valeur $\gamma = 1$ est un candidat privilégié puisqu'il répond aux deux propriétés à la fois.

Les propriétés théoriques de l'estimateur Bridge ont fait l'objet d'études, dont nous faisons un recensement ici. Lorsque $\gamma \leq 1$, les propriétés de consistance en *estimation* et *sélection* de variables ont été prouvées dans le cas $p \leq n$ par Huang, Horowitz, et Ma [82]. Les auteurs montrent également que si les variables pertinentes sont peu corrélées avec les variables non pertinentes, et sous certaines conditions sur les paramètres n, p et γ, l'estimateur Bridge reste consistant en *sélection* de variables même lorsque $p \geq n$ (il est alors défini en deux étapes).

Notons également le travail de Koltchinskii [93] qui considère les performances théoriques

de l'estimateur défini avec la pénalité Bridge où $1 < \gamma \leq 1 + \gamma_{\max}$ avec $\gamma_{\max} = \frac{1}{\log(n)}$. Une motivation pour ce choix de γ_n est que celui-ci permet de garder une équivalence entre les normes ℓ_1 (définissant le Lasso) et ℓ_γ des pénalités. On constate qu'en perturbant légèrement le problème de minimisation ℓ_1, on arrive à garder une équivalence avec celui-ci. Koltchinskii [93] a essentiellement considéré l'erreur de prédiction. Il a établi des IS pour cet objectif, sous des hypothèses sur la matrice de Gram proches de celles utilisées pour le Lasso.

Annexe B

Aspects algorithmiques du Lasso

Dans cette annexe, nous présentons l'algorithme LARS et discutons le choix du paramètre de régularisation optimal λ. Commençons par quelques rappels. L'estimateur Lasso a été défini comme :

$$\hat{\beta}^L(\lambda_n) \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|Y - X\beta\|_n^2 + \lambda_n \|\beta\|_1, \tag{B.1}$$

et les conditions KKT associées :

$$\begin{cases} 2X_j' \left(Y - X\hat{\beta}^L(\lambda_n)\right) = \lambda_n \operatorname{sgn}(\hat{\beta}_j^L) & \forall j \in \hat{\mathcal{A}}_L, \\ 2\left|X_j' \left(Y - X\hat{\beta}^L(\lambda_n)\right)\right| < \lambda_n & \forall j \notin \hat{\mathcal{A}}_L, \end{cases} \tag{B.2}$$

où $\hat{\mathcal{A}}_L$ est l'ensemble actif associé à l'estimateur $\hat{\beta}^L$. Dans le chapitre 2, nous avons observé que la solution Lasso est linéaire par morceaux. De plus, nous avons établi que cette linéarité ne change que pour un nombre fini K de points. Ces points sont obtenus grâce aux conditions KKT ci-dessus.

Définition B.1.
Soient $\lambda^{(1)}, \ldots, \lambda^{(K)}$ les K valeurs du paramètre de régularisation λ_n telles que la linéarité de l'estimateur Lasso $\hat{\beta}^L(\lambda_n)$ change. L'ensemble $\{\lambda^{(1)}, \ldots, \lambda^{(K)}\}$ est alors appelé l'ensemble des points de transitions.

Les avancées récentes en terme d'optimisation (Turlach [141] et Efron, Hastie, Johnstone, et Tibshirani [61]) autorisent la manipulation d'estimateurs de la forme (B.1). Ainsi, il existe plusieures méthodes d'approximation de la solution Lasso.

Certaines reposent sur une approximation à λ_n fixé (ou de manière équivalente à t fixé si on considère la forme du Lasso définie sous contrainte (2.14)). C'est par exemple le cas des algorithmes de descente *coordonnée par coordonnée*. Ces algorithmes approchent la solution en ne modifiant qu'une seule composante du vecteur $\hat{\beta}^L$ à la fois. Cette stratégie est d'abord décrite pour le Lasso par Fu [69] sous le nom de *Shooting algorithm*, puis par Daubechies, Defrise, et De Mol [49] et finalement par Friedman, Hastie, Höfling, et Tibshirani [67]. Nous pouvons également citer l'algorithme de Osborne, Presnell, et Turlach [118] qui fournit la solution Lasso à λ_n fixé, ou celui de points intérieurs proposé par Chen, Donoho, et Saunders [43].

D'autres méthodes se regroupent dans la famille d'algorithmes d'homotopie ; elles considèrent λ_n ou t_n en tant que paramètres d'homotopie. Dans cette famille, on peut considérer les algorithmes de Osborne, Presnell, et Turlach [119], Turlach [141] ou encore le LARS de Efron, Hastie, Johnstone, et Tibshirani [61].

B.1 Description de l'algorithme LARS

Nous présentons dans ce paragraphe l'algorithme *LARS (Least Angle Regression)*. Cet algorithme, très employé, est implémenté sur différents logiciels. Le lecteur intéressé pourra consulter le site de Tibshirani : http ://www-stat.stanford.edu/ ˜ tibs/lasso.html, pour des versions sur Splus ou R ; ou bien le site de Karl Skoglund pour une version Matlab du LARS.

L'algorithme LARS est un algorithme d'homotopie itératif *forward*, introduit par Efron, Hastie, Johnstone, et Tibshirani [61]. Il fournit rapidement l'ensemble de *toutes* les solutions Lasso :

$$\{\hat{\beta}^L(\lambda_n),\ \lambda_n \geq 0\}. \tag{B.3}$$

A chaque étape, la construction s'appuie sur le maximum de corrélation entre les variables X_j et le résidu $Y - X\hat{\beta}^L(\lambda_n)$. L'ensemble (B.3) définit le *chemin de régularisation* du Lasso. L'algorithme LARS ne calcule qu'un petit nombre K de solutions Lasso appartenant au chemin de régularisation : il s'agit des solutions Lasso évaluées aux K points de transition (cf. Définition B.1). L'ensemble de toutes les solutions (B.3) est ensuite produit par interpolation

linéaire. Afin d'alléger les notations dans cette section, nous notons $\hat{\beta}$ l'estimateur Lasso et $\hat{\mathcal{A}}$ l'ensemble actif associé.

Remarque B.1.
L'algorithme LARS construit l'estimateur Lasso (B.1) *à l'aide des Conditions KKT* (B.2). *Le LARS base en effet sa construction sur les corrélations* $X_j'(Y - X\hat{\beta}(\lambda_n))$, $j = 1, \ldots, p$ *: il fournit des estimations de* β^* *lorsque* λ_n *correspond à un des* K *points de transitions introduits dans la Définition* (B.1). *Comme la linéarité de ce chemin de régularisation ne change qu'au niveau des points de transition* $\{\lambda^{(1)}, \ldots, \lambda^{(K)}\}$, *il suffit d'interpoler linéairement entre ces points pour reconstruire le reste du chemin. L'algorithme LARS s'arrête donc en* K *étapes.*

Nous décrivons ci-dessous le calcul des K estimateurs Lasso par l'algorithme LARS. Introduisons au préalable quelques notations supplémentaires. Soit K, le nombre total d'étapes réalisées par l'algorithme LARS et k une étape quelconque. A chaque étape k, on associe le paramètre de régularisation $\lambda^{(k)}$ (*l'étape k est donc un paramètre de régularisation au même titre que* λ_n) et l'ensemble actif $\hat{\mathcal{A}}^k$. Pour tout $a \in \mathbb{R}^p$, on note $a_{\hat{\mathcal{A}}^k}$ le vecteur de taille $|\hat{\mathcal{A}}^k|$ correspondant à la restriction du vecteur a aux composantes d'indice $j \in \hat{\mathcal{A}}^k$. De même, pour une matrice Z de taille $n \times p$, on note $Z_{\hat{\mathcal{A}}^k}$, la matrice de taille $n \times |\hat{\mathcal{A}}^k|$, constituée des colonnes de Z dont les indices appartiennent à $\hat{\mathcal{A}}^k$. Enfin, on note $\hat{\beta}^{[k]}_{\hat{\mathcal{A}}^k}$ l'estimateur Lasso construit par l'algorithme LARS à l'étape k.

Étape 1 (correspondant à $\lambda^{(0)} = \infty$).
À cette première étape, $\hat{\mathcal{A}}^0 = \emptyset$ et $\hat{\beta}^{[0]}_{\hat{\mathcal{A}}^0} = 0$. Par conséquent, le résidu vaut Y. On identifie la variable X_j ayant la plus grande corrélation avec Y, notons la $X_{(1)}$.
On actualise alors l'ensemble actif : $\hat{\mathcal{A}}^1 = \hat{\mathcal{A}}^0 \cup \{(1)\}$. Le vecteur $X_{(1)}$ correspond à la direction de déplacement à la première étape. La longeur d'avancement γ sur l'axe $X_{(1)}$ correspond au plus petit réel positif tel que la deuxième condition KKT (B.2) ne soit plus satisfaite pour une variable X_j, avec $j \notin \hat{\mathcal{A}}^1$. Notons cette variable $X_{(2)}$. C'est le "premier instant" où une variable inactive devient active.
A ce niveau de la procédure, la première Condition KKT (B.2) fait que les variables $X_{(1)}$ et $X_{(2)}$ ont la même corrélation (en valeur absolue) avec le résidu $Y - X_{\hat{\mathcal{A}}^1}\hat{\beta}^{[1]}_{\hat{\mathcal{A}}^1}$. A la fin de cette première étape, le vecteur $\hat{\beta}^{[1]}_{\hat{\mathcal{A}}^1}$ est un réel et vaut γ, on a à présent $\hat{\mathcal{A}}^2 = \hat{\mathcal{A}}^1 \cup \{(2)\}$.

Étape 2.
On considère à présent les variables $X_{(1)}$ et $X_{(2)}$ et $\hat{\mathcal{A}}^2 = \{(1), (2)\}$. L'algorithme calcule le vecteur équiangulaire entre les vecteurs $X_{(1)}$ et $X_{(2)}$ (i.e. le vecteur bissectrice entre les

deux vecteurs $X_{(1)}$ et $X_{(2)}$). Ce vecteur correspond à la direction de déplacement choisie, dont la longueur d'avancement est γ, avec $\gamma = \min\{\overline{\gamma}, \tilde{\gamma}\}$, où $\overline{\gamma}$ et $\tilde{\gamma}$ désignent respectivement les plus petits réels positifs tels que, la première et deuxième Condition KKT (B.2) ne sont plus satisfaites. Deux cas sont alors possibles :

- si $\gamma = \overline{\gamma}$, la variable pour laquelle le minimum est atteint, disons $X_{(l)}$ avec $(l) \in \hat{\mathcal{A}}^2$, devient inactive, i.e. $\hat{\mathcal{A}}^3 = \hat{\mathcal{A}}^2 \setminus \{(l)\}$. Dans ce cas, l'ensemble actif ne contient plus qu'une variable et l'algorithme retourne à l'Étape 1 ;
- si $\gamma = \tilde{\gamma}$, la variable pour laquelle le minimum est atteint, disons $X_{(3)}$ avec $(3) \in (\hat{\mathcal{A}}^2)^C$, devient active, i.e. le résidu $Y - X_{\hat{\mathcal{A}}^2}\hat{\beta}^{[2]}_{\hat{\mathcal{A}}^2}$ admet une corrélation égale (en valeur absolue) avec chacune des variables $X_{(1)}$, $X_{(2)}$ et $X_{(3)}$. On a donc $\hat{\mathcal{A}}^3 = \hat{\mathcal{A}}^2 \cup \{(3)\}$.

Étape k.

On considère les variables actives $X_{(1)}, \ldots, X_{(m)}$ sélectionnées par l'algorithme à l'Étape $k-1$, i.e. $\hat{\mathcal{A}}^k = \{(1), (2), \ldots, (m)\}$. L'algorithme calcule le vecteur équiangulaire entre les vecteurs décrits par les variables actives. Ce vecteur correspond à la direction d'avancement, dont la longueur est déterminée à partir des Conditions KKT : $\gamma = \min\{\overline{\gamma}, \tilde{\gamma}\}$. De la même manière que précédemment, le choix de γ se fait selon que la première ou la deuxième Condition KKT n'est plus satisfaite en premier. Dans le premier cas, une variable est écartée de $\hat{\mathcal{A}}^k$ et l'algorithme opère à l'étape suivante comme à l'Étape $k-1$. Dans le second cas, une variable de $(\hat{\mathcal{A}}^k)^C$ est ajoutée à $\hat{\mathcal{A}}^k$.

Étape K : la dernière étape.

À cette dernière étape, deux cas sont possibles :

- si $p \leq n$, l'algorithme s'arrête en donnant comme solution, l'estimateur des moindres carrés, après avoir estimé l'ensemble du chemin de régularisation (B.3), et arrêté pour une valeur du paramètre de régularisation $\lambda^{(K)} = 0$.
- si $p > n$, l'algorithme LARS ne peut pas atteindre de petites valeurs pour le paramètre de régularisation. Ainsi $\lambda^{(K)} > 0$ et l'algorithme ne sélectionne pas l'ensemble de toutes les variables, car il est restreint par le nombre d'observations n, inférieur à p.

Le lecteur intéressé pourra se référer à Efron, Hastie, Johnstone, et Tibshirani [61, page 8] pour plus de détails sur le calcul explicite des longueurs d'avancement à chaque étape.

Remarque B.2.

1. *La version originale de l'algorithme LARS ne prend pas en compte la première Condition KKT* (B.2). *La version présentée ici est une adaptation du LARS à l'estimateur Lasso.*

2. *Entre deux étapes de cet algorithme, les corrélations entre toutes les variables actives et le résidu sont égales en valeur absolue. D'un point de vue géométrique, ceci s'explique par le fait qu'à chaque étape de l'algorithme LARS, la direction d'avancement est équiangulaire entre les variables sélectionnées. De ce fait, la contribution de toutes les variables sur le résidu évolue de manière identique.*

Des généralisations de cet algorithme à d'autres problèmes ou à d'autres estimateurs ont fait l'objet de quelques travaux intéressants. Parmi ceux-ci, nous pouvons citer celui de Rosset et Zhu [126] qui propose des conditions nécessaires et suffisantes sur la perte et la pénalité utilisées dans la définition de l'estimateur, pour que ce dernier admette un chemin de régularisation linéaire par morceaux. Le Lasso (B.1) défini avec la perte ℓ_2 et la pénalité ℓ_1 satisfait ces conditions. La propriété de linéarité par morceaux est fortement recherchée pour réduire le temps de calcul de l'ensemble des solutions d'un estimateur. Les auteurs fournissent également un algorithme de type LARS pour résoudre ces problèmes. D'autres extensions de cet algorithme sont également considérées par Zou et Hastie [167], Zou [166], Yuan et Lin [160], James, Radchenko, et Lv [84] parmi d'autres.

B.2 Choix du paramètre λ_n optimal

L'algorithme LARS fournit une solution approchée du chemin de régularisation du Lasso (B.3). Il permet en particulier de trouver K solutions correspondant aux points de transition $\{\lambda^{(1)}, \ldots, \lambda^{(K)}\}$. L'estimateur final est alors sélectionné parmi les K solutions en fonction de l'objectif de l'étude (*prédiction/estimation/sélection*).

Le critère de sélection de l'estimateur final doit prendre en compte l'objectif de l'étude (Leng, Lin, et Wahba [98]). Nous pouvons par exemple utiliser, parmi les méthodes de ré-échantillonnage (Efron [57, 59]), la validation croisée et ses multiples variantes (Shao [128], Allen [3] et Stone [133]), mais également les critères de type ℓ_0 comme ceux considérés dans le chapitre 2.

Ces méthodes de sélection sont très répandues et leurs performances connues. Ainsi, les méthodes de validation croisée tendent à sélectionner des estimateurs utilisant un nombre excessif de variables (cf. Meinshausen et Bühlmann [112] et Leng, Lin, et Wahba [98]).

La divergence entre les différentes solutions obtenues par ces méthodes de sélection de l'estimateur final, est d'autant plus importante que certaines variables X_j, $j \in \{1, \ldots, p\}$ sont corrélées. Dans ce cas, en effet, l'algorithme LARS, comme la plupart des algorithmes utilisés pour résoudre le problème d'optimisation (B.1), échouent dans l'estimation du chemin

de régularisation (B.3). Le principal désagrément en présence de fortes corrélations se présente dans les performances en *sélection de variables*. En effet, les fortes corrélations entre variables perturbent l'ordre de sélection de celles-ci. Lorsqu'une variable est sélectionnée, les variables qui lui sont corrélées sont considérées comme étant non pertinentes. Elles sont écartées dans la plupart des cas de l'étude.

Bibliographie

[1] Adler R. J. *An introduction to continuity, extrema, and related topics for general Gaussian processes*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 12. Institute of Mathematical Statistics, Hayward, CA, 1990.

[2] Akaike H. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973.

[3] Allen D. M. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16 :125–127, 1974.

[4] Alquier P. LASSO, iterative feature selection and the correlation selector : oracle inequalities and numerical performances. *Electron. J. Stat.*, 2 :1129–1152, 2008.

[5] Alquier P. Pac-bayesian bounds for randomized empirical risk minimizers. *Math. Methods Statist.*, 17 (4) :279–304, 2008.

[6] Alquier P. Iterative feature selection in least square regression estimation. *Ann. Inst. Henri Poincaré Probab. Stat.*, 44(1) :47–88, 2008.

[7] Alquier P. et Hebiri M. Generalization of ℓ_1 constraint for high-dimensional regression problems. Preprint Laboratoire de Probabilités et Modèles Aléatoires, submitted, 2008.

[8] Alquier P. et Hebiri M. Transductive extensions of the LASSO and the Dantzig Selector. Manuscript, 2009.

[9] Audibert J. Y. Fast learning rates in statistical inference through aggregation. *Ann. Statist., to appear*, 2008.

[10] Bach F. Bolasso : model consistent lasso estimation through the bootstrap. *Proceedings of the Twenty-fifth International Conference on Machine Learning (ICML)*, 2008.

[11] Bach F. Model-consistent sparse estimation through the bootstrap. Manuscript, 2009.

[12] Bach F. Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.*, 9 :1179–1225, 2008.

[13] Baraud Y. Model selection for regression on a random design. *ESAIM Probab. Statist.*, 6 :127–146 (electronic), 2002.

[14] Barron A., Birgé L., et Massart P. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3) :301–413, 1999.

[15] Bauer P., Pötscher B. M., et Hackl P. Model selection by multiple test procedures. *Statistics*, 19(1) :39–44, 1988.

[16] Benjamini Y. et Hochberg Y. Controlling the false discovery rate : a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57(1) :289–300, 1995.

[17] Bickel P., Ritov Y., et Tsybakov A. Simultaneous analysis of lasso and dantzig selector. Submitted to Ann. Statist. Available at http ://www.proba.jussieu.fr/pageperso/tsybakov/, 2007.

[18] Bickel P. J. et Li B. Regularization in statistics. *TEST*, 15(2) :271–344, 2006.

[19] Birgé L. Model selection via testing : an alternative to (penalized) maximum likelihood estimators. *Ann. Inst. H. Poincaré Probab. Statist.*, 42(3) :273–325, 2006.

[20] Birgé L. et Massart P. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2) :33–73, 2007.

[21] Birgé L. et Massart P. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3) : 203–268, 2001.

[22] Breiman L. Heuristics of instability and stabilization in model selection. *Ann. Statist.*, 24(6) :2350–2383, 1996.

[23] Breiman L. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4) :373–384, 1995.

[24] Bühlmann P. et Hothorn T. Twin boosting : improved feature selection and prediction. Manuscript, 2009.

[25] Bühlmann P. et Yu B. Invited discussion on three papers on boosting by j.g, lugosi and vayatis, and zhang. 2004.

[26] Bunea F. Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization. *Electron. J. Stat.*, 2 :1153–1194, 2008.

[27] Bunea F. Consistent covariate selection and post model selection inference in semiparametric regression. *Ann. Statist.*, 32(3) :898–927, 2004.

[28] Bunea F. Consistent selection via the lasso for high dimensional approximating regression models. 2008. IMS Collections, B. Clarke and S. Ghosal Editors.

[29] Bunea F., Tsybakov A., et Wegkamp M. H. Aggregation for regression learning. manuscript, 2004.

[30] Bunea F., Tsybakov A. B., et Wegkamp M. H. Aggregation and sparsity via l_1 penalized least squares. In *Learning theory*, volume 4005 of *Lecture Notes in Comput. Sci.*, pages 379–391. 2006.

[31] Bunea F., Wegkamp M. H., et Auguste A. Consistent variable selection in high dimensional regression via multiple testing. *J. Statist. Plann. Inference*, 136(12) :4349–4364, 2006.

[32] Bunea F., Tsybakov A., et Wegkamp M. Aggregation for Gaussian regression. *Ann. Statist.*, 35(4) :1674–1697, 2007.

[33] Bunea F., Tsybakov A. B., et Wegkamp M. H. Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.*, 1 :169–194, 2007.

[34] Cai T. On block thresholding in wavelet regression : adaptivity, block size and threshold level. *Statist. Sinica*, 12(4) :1241–1273, 2002.

[35] Candès E. et Tao T. Rejoinder : "The Dantzig selector : statistical estimation when p is much larger than n" [Ann. Statist. **35** (2007), no. 6, 2313–2351 ; mr2382644]. *Ann. Statist.*, 35(6) :2392–2404, 2007.

[36] Casella G. et Berger R. L. *Statistical inference*. The Wadsworth & Brooks/Cole Statistics/Probability Series. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1990.

[37] Catoni O. A pac-bayesian approach to adaptive classification. Manuscript, 2003.

[38] Catoni O. *Statistical learning theory and stochastic optimization*, volume 1851 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2004. Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001.

[39] Catoni O. Pac-bayesian supervised classification : The thermodynamics of statistical learning. *IMS Lecture Notes Monograph Series*, 56, December 03 2007.

[40] Cavalier L. et Tsybakov A. B. Penalized blockwise Stein's method, monotone oracles and sharp adaptive estimation. *Math. Methods Statist.*, 10(3) :247–282, 2001. Meeting on Mathematical Statistics (Marseille, 2000).

[41] Chapelle O., Schölkopf B., et Zien A. *Semi-supervised learning*. MIT Press, Cambridge, MA, 2006.

[42] Chen S. S. et Donoho D. L. Atomic decomposition by basis pursuit. Technical Report, 1995.

[43] Chen S. S., Donoho D. L., et Saunders M. A. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1) :33–61 (electronic), 1998.

[44] Chesneau C. et Hebiri M. Some theoretical results on the grouped variables lasso. *Math. Methods Statist.*, 17(4) :317–326, 2008.

[45] Ciarlet P. G. *Introduction à l'analyse numérique matricielle et à l'optimisation*. Collection Mathématiques Appliquées pour la Maîtrise. [Collection of Applied Mathematics for the Master's Degree]. Masson, Paris, 1982.

[46] Dalalyan A. S. et Tsybakov A. B. Aggregation by exponential weighting and sharp oracle inequalities. In *Learning theory*, volume 4539 of *Lecture Notes in Comput. Sci.*, pages 97–111. Springer, Berlin, 2007.

[47] Dalalyan A. S. et Tsybakov A. B. Aggregation by exponential weighting, sharp pac-bayesian bounds and sparsity. *Mach. Learn.*, 72(1-2) :39–61, 2008.

[48] Dalalyan A. S. et Tsybakov A. B. Sparse regression learning by aggregation and langevin monte-carlo. Manuscript, 2009.

[49] Daubechies I., Defrise M., et De Mol C. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math.*, 57(11) :1413–1457, 2004.

[50] De Mol C., De Vito E., et Rosasco L. Elastic-net regularization in learning theory. Manuscript, 2008.

[51] Donoho D., Elad M., et Temlyakov V. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory*, 52(1) :6–18, 2006.

[52] Donoho D. L. For most large underdetermined systems of linear equations the minimal l_1-norm solution is also the sparsest solution. *Comm. Pure Appl. Math.*, 59(6) :797–829, 2006.

[53] Donoho D. L. et Johnstone I. M. Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, 90(432) :1200–1224, 1995.

[54] Donoho D. L. et Johnstone I. M. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3) :425–455, 1994.

[55] Donoho D. L. et Tanner J. Sparse nonnegative solution of underdetermined linear equations by linear programming. *Proc. Natl. Acad. Sci. USA*, 102(27) :9446–9451 (electronic), 2005.

[56] Donoho D. L., Johnstone I. M., Kerkyacharian G., et Picard D. Wavelet shrinkage : asymptopia ? *J. Roy. Statist. Soc. Ser. B*, 57(2) :301–369, 1995. With discussion and a reply by the authors.

[57] Efron B. Bootstrap methods : another look at the jackknife. *Ann. Statist.*, 7(1) :1–26, 1979.

[58] Efron B. How biased is the apparent error rate of a prediction rule ? *J. Amer. Statist. Assoc.*, 81(394) :461–470, 1986.

[59] Efron B. *The jackknife, the bootstrap and other resampling plans*, volume 38 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pa., 1982.

[60] Efron B. et Tibshirani R. J. *An introduction to the bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, New York, 1993.

[61] Efron B., Hastie T., Johnstone I., et Tibshirani R. Least angle regression. *Ann. Statist.*, 32(2) :407–499, 2004. With discussion, and a rejoinder by the authors.

[62] Fan J. Comments on Śwavelets in statistics : A reviewŠ by a. antoniadis. *J. Ital. Statist. Assoc.*, 6 :131–138, 1997.

[63] Fan J. et Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456) :1348–1360, 2001.

[64] Fan J. et Peng H. Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.*, 32(3) :928–961, 2004.

[65] Foster D. P. et George E. I. The risk inflation criterion for multiple regression. *Ann. Statist.*, 22(4) :1947–1975, 1994.

[66] Frank I. E. et Friedman J. H. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2) :109–135, 1993.

[67] Friedman J., Hastie T., Höfling H., et Tibshirani R. Pathwise coordinate optimization. *Ann. Appl. Stat.*, 1(2) :302–332, 2007.

[68] Friedman J. H. Greedy function approximation : a gradient boosting machine. *Ann. Statist.*, 29(5) :1189–1232, 2001.

[69] Fu W. J. Penalized regressions : the bridge versus the lasso. *J. Comput. Graph. Statist.*, 7(3) :397–416, 1998.

[70] Garrigues P. et El Ghaoui L. An homotopy algorithm for the lasso with online observations. *To appear in Neural Information Processing Systems (NIPS) 21*, 2008.

[71] Greenshtein E. et Ritov Y. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6) :971–988, 2004.

[72] Guyon X. et Yao J-F. On the underfitting and overfitting sets of models chosen by order selection criteria. *J. Multivariate Anal.*, 70(2) :221–249, 1999.

[73] Györfi L., Kohler M., Krzyżak A., et Walk H. *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer-Verlag, New York, 2002.

[74] Hastie T., Tibshirani R., et Friedman J. *The elements of statistical learning*. Springer Series in Statistics. Springer-Verlag, New York, 2001. Data mining, inference, and prediction.

[75] Hastie T. J. et Tibshirani R. J. *Generalized additive models*, volume 43 of *Monographs on Statistics and Applied Probability*. Chapman and Hall Ltd., London, 1990.

[76] Haughton D. M. A. On the choice of a model to fit data from an exponential family. *Ann. Statist.*, 16(1) :342–355, 1988.

[77] Hebiri M. Sparse conformal predictors. Preprint Laboratoire de Probabilités et Modèles Aléatoires, 2008.

[78] Hebiri M. Regularization with the smooth-lasso procedure. Preprint Laboratoire de Probabilités et Modèles Aléatoires, 2008.

[79] Hoerl Arthur E. et Kennard R. W. A note on a power generalization of ridge regression. *Technometrics*, 17 :269, 1975.

[80] Huang C., Cheang G. L. H., et Barron A. Risk of penalized least squares, greedy selection and l1 penalization for flexible function libraries. preprint, 2008.

[81] Huang J. et Zhang T. The benefit of group sparsity. manuscript, 2009.

[82] Huang J., Horowitz J. L., et Ma S. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.*, 36(2) :587–613, 2008.

[83] Hunter D. R. et Li R. Variable selection using MM algorithms. *Ann. Statist.*, 33(4) : 1617–1642, 2005.

[84] James G. M., Radchenko P., et Lv J. Dasso : connections between the dantzig selector and lasso. *J. Roy. Statist. Soc. Ser. B*, 71 :127–142, 2009.

[85] Jia J. et Yu B. On model selection consistency of the elastic net when p » n. *Technical Report*, 2008.

[86] Juditsky A. et Nemirovski A. Functional aggregation for nonparametric regression. *Ann. Statist.*, 28(3) :681–712, 2000.

[87] Juditsky A., Rigollet P., et Tsybakov A. B. Learning by mirror averaging. *Ann. Statist.*, 36(5) :2183–2206, 2008.

[88] Kim S. J., Koh K., Lustig M., Boyd S., et Gorinevsky D. An interior-point method for large-scale l1-regularized least squares. *IEEE Journal of Selected Topics in Signal Processing*, 1(4) :606–617, 2007.

[89] Kim Y., Kim J., et Kim Y. Blockwise sparse regression. *Statist. Sinica*, 16(2) :375–390, 2006.

[90] Knight K. et Fu W. Asymptotics for lasso-type estimators. *Ann. Statist.*, 28(5) : 1356–1378, 2000.

[91] Koltchinskii V. Dantzig selector and sparsity oracle inequalities. *Bernoulli*, to appear, 2008.

[92] Koltchinskii V. Sparse recovery in convex hulls via entropy penalization. *Ann. Statist.*, 37(3) :1332–1359, 2009.

[93] Koltchinskii V. Sparsity in penalized empirical risk minimization. *Ann. IHP. Probability and Statistics*, to appear, 2007.

[94] Land S. R. et Friedman J. H. Variable fusion : a new method of adaptive signal regression. *Manuscript*, 1996.

[95] Lauritzen S. L. *Graphical models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press Oxford University Press, New York, 1996. Oxford Science Publications.

[96] Lawson C. L. et Hanson R. J. *Solving least squares problems*, volume 15 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1995. Revised reprint of the 1974 original.

[97] Leeb H. et Pötscher B. M. Model selection and inference : facts and fiction. *Econometric Theory*, 21(1) :21–59, 2005.

[98] Leng C., Lin Y., et Wahba G. A note on the lasso and related procedures in model selection. *Statist. Sinica*, 16(4) :1273–1284, 2006.

[99] Leung G. et Barron A. R. Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory*, 52(8) :3396–3410, 2006.

[100] Li K-C. Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation : discrete index set. *Ann. Statist.*, 15(3) :958–975, 1987.

[101] Lounici K. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electron. J. Stat.*, 2 :90–102, 2008.

[102] Lounici K., Pontil M., Tsybakov A. B., et van de Geer S. Taking advantage of sparsity in multi-task learning. Manuscript, 2008.

[103] Lugosi G. et Vayatis N. On the Bayes-risk consistency of regularized boosting methods. *Ann. Statist.*, 32(1) :30–55, 2004.

[104] Mallat S. *A wavelet tour of signal processing*. Elsevier/Academic Press, Amsterdam, third edition, 2009. The sparse way, With contributions from Gabriel Peyré.

[105] Mallows C. L. Some comments on c_p. *Technometrics*, 15(4) :661–675, 1973.

[106] Massart P. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[107] McKay Curtis S. et Ghosal S. Approximate posterior model probabilities in additive models via the group lasso. Manuscript, 2008.

[108] McQuarrie A. D. R. et Tsai C-L. *Regression and time series model selection.* World Scientific Publishing Co. Inc., River Edge, NJ, 1998.

[109] Meier L., van de Geer S., et Bühlmann P. The group lasso for logistic regression. Research Report 131, ETH Zürich,, 2006.

[110] Meier L., van de Geer S., et Bühlmann P. High-dimensional additive modeling. Manuscript, 2008.

[111] Meinshausen N. Relaxed Lasso. *Comput. Statist. Data Anal.*, 52(1) :374–393, 2007.

[112] Meinshausen N. et Bühlmann P. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3) :1436–1462, 2006.

[113] Meinshausen N. et Bühlmann P. Stability selection. Manuscript, 2008.

[114] Meinshausen N. et Yu B. Lasso-type recovery of sparse representations for high-dimensional data. Ann. Statist., to appear, 2006.

[115] Nardi Y. et Rinaldo A. On the asymptotic properties of the group lasso estimator for linear models. *Electron. J. Stat.*, 2 :605–633, 2008.

[116] Nemirovski A. Topics in non-parametric statistics. In *Lectures on probability theory and statistics (Saint-Flour, 1998)*, volume 1738 of *Lecture Notes in Math.*, pages 85–277. Springer, Berlin, 2000.

[117] Obozinski G., Wainwright M. J., et Jordan M. I. Union support recovery in high-dimensional multivariate regression. Manuscript, 2008.

[118] Osborne M., Presnell B., et Turlach B. On the LASSO and its dual. *J. Comput. Graph. Statist.*, 9(2) :319–337, 2000.

[119] Osborne M. R., Presnell B., et Turlach B. A. A new approach to variable selection in least squares problems. *IMA J. Numer. Anal.*, 20(3) :389–403, 2000.

[120] Penrose R. A generalized inverse for matrices. *Proc. Cambridge Philos. Soc.*, 51 : 406–413, 1955.

[121] Polyak B. T. et Tsybakov A. B. Asymptotic optimality of the C_p-test in the projection estimation of a regression. *Teor. Veroyatnost. i Primenen.*, 35(2) :305–317, 1990.

[122] Pötscher B. M. Order estimation in ARMA-models by Lagrangian multiplier tests. *Ann. Statist.*, 11(3) :872–885, 1983.

[123] Pötscher B. M. et Leeb H. On the distribution of penalized maximum likelihood estimators : the lasso, scad, and thresholding. Manuscript, 2007.

[124] Pötscher B. M. et Schneider U. On the distribution of the adaptive lasso estimator. Manuscript, 2008.

[125] Rinaldo A. Properties and refinements of the fused lasso. *Technical Report*, 2008.

[126] Rosset S. et Zhu J. Piecewise linear regularized solution paths. *Ann. Statist.*

[127] Schwartz G. Estimating the dimension of a model. *Ann. Statist.*, 6(2) :461–464, 1978.

[128] Shao J. Linear model selection by cross-validation. *J. Amer. Statist. Assoc.*, 88(422) : 486–494, 1993.

[129] Shao J. An asymptotic theory for linear model selection. *Statist. Sinica*, 7(2) :221–264, 1997. With comments and a rejoinder by the author.

[130] Shen X. et Ye J. Adaptive model selection. *J. Amer. Statist. Assoc.*, 97(457) :210–221, 2002.

[131] Shibata R. An optimal selection of regression variables. *Biometrika*, 68(1) :45–54, 1981.

[132] Shibata R. Selection of the number of regression variables ; a minimax choice of generalized FPE. *Ann. Inst. Statist. Math.*, 38(3) :459–474, 1986.

[133] Stone M. Corrigendum : "Cross-validatory choice and assessment of statistical predictions" (J. Roy. Statist. Soc. Ser. B **36** (1974), 111–147). *J. Roy. Statist. Soc. Ser. B*, 38(1) :102, 1976.

[134] Tarigan B. et van de Geer S. A. Classifiers of support vector machine type with l_1 complexity regularization. *Bernoulli*, 12(6) :1045–1076, 2006.

[135] Tibshirani R. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1) :267–288, 1996.

[136] Tibshirani R., Saunders M., Rosset S., Zhu J., et Knight K. Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(1) :91–108, 2005.

[137] Tropp J. A. Greed is good : algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory*, 50(10) :2231–2242, 2004.

[138] Tsybakov A. Cours de statistiques appliquée. Université Paris 6.

[139] Tsybakov A. Optimal rates of aggregation. *In Proceedings of 16th Annual Conference on Learning Theory (COLT) and 7th Annual Workshop on Kernel Machines. Lecture Notes in Artificial Intelligence*, 2777 :303–313, 2003. Springer-Verlag, Heidelberg.

[140] Tsybakov A. B. et van de Geer S. A. Square root penalty : adaptation to the margin in classification and in edge estimation. *Ann. Statist.*, 33(3) :1203–1224, 2005.

[141] Turlach B. A. On algorithms for solving least squares problems under an ℓ_1 penalty or an ℓ_1 constraint. *In 2004 Proceedings of the American Statistical Association. Statistical Computing Section [CD-ROM], American Statistical Association*, pages 2572–2577, 2005. Alexandria, VA.

[142] van de Geer S. The deterministic lasso. *Tech Report n.140, Seminar für Statistik ETH, Zürich*, 2007.

[143] van de Geer S. High-dimensional generalized linear models and the lasso. *Ann. Statist.*, 36(2) :614–645, 2008.

[144] van der Vaart A. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[145] Vapnik V. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1998.

[146] Vapnik V. *Statistical learning theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons Inc., New York, 1998. A Wiley-Interscience Publication.

[147] Vovk V. Asymptotic optimality of transductive confidence machine. In *Algorithmic learning theory*, volume 2533 of *Lecture Notes in Comput. Sci.*, pages 336–350. Springer, Berlin, 2002.

[148] Vovk V. On-line confidence machines are well-calibrated. *In : Proceedings of the Forty-Third Annual Symposium on Foundations of Computer Science*, pages 187–196, 2002.

[149] Vovk V., Gammerman A., et Saunders C. Machine-learning applications of algorithmic randomness. *In : Proceedings of the Sixteenth International Conference on Machine Learning*, pages 444–453, 1999.

[150] Vovk V., Gammerman A., et Shafer G. *Algorithmic learning in a random world*. Springer, New York, 2005.

[151] Vovk V., Nouretdinov Ilia G., et Gammerman A. On-line predictive linear regression. *Technical Report*, 2007.

[152] Wainwright M. Sharp thresholds for noisy and high-dimensional recovery of sparsity using l1-constrained quadratic programming. Manuscript, 2006.

[153] Wang H., Li G., et Jiang G. Robust regression shrinkage and consistent variable selection through the LAD-Lasso. *J. Bus. Econom. Statist.*, 25(3) :347–355, 2007.

[154] Wegkamp M. Model selection in nonparametric regression. *Ann. Statist.*, 31(1) : 252–273, 2003.

[155] Witten D. M. et Tibshirani R. Covariance-regularized regression and classification for high-dimensional problems. Manuscript, 2009.

[156] Yang Y. Can the strengths of AIC and BIC be shared ? A conflict between model indentification and regression estimation. *Biometrika*, 92(4) :937–950, 2005.

[157] Yang Y. Regression with multiple candidate models : selecting or mixing ? *Statist. Sinica*, 13(3) :783–809, 2003.

[158] Yang Y. Aggregating regression procedures to improve performance. *Bernoulli*, 10 (1) :25–47, 2004.

[159] Yuan M. et Lin Y. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(1) :49–67, 2006.

[160] Yuan M. et Lin Y. On the non-negative garrote estimator. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 69(2) :143–161, 2007.

[161] Yuditskiĭ A. B., Nazin A. V., Tsybakov A. B., et Vayatis N. Recursive aggregation of estimators by the mirror descent method with averaging. *Problemy Peredachi Informatsii*, 41(4) :78–96, 2005.

[162] Zhang C-H. Nearly unbiased variable selection under minimax concave penality. Manuscript, 2008.

[163] Zhang C-H. et Huang J. The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.*, 36(4) :1567–1594, 2008.

[164] Zhao P. et Yu B. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7 : 2541–2563, 2006.

[165] Zhou S., van de Geer S., et Bühlmann P. Adaptive lasso for high dimensional regression and gaussian graphical modeling. Manuscript, 2009.

[166] Zou H. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101 (476) :1418–1429, 2006.

[167] Zou H. et Hastie T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(2) :301–320, 2005.

[168] Zou H. et Li R. One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.*, 36(4) :1509–1533, 2008.

[169] Zou H., Hastie T., et Tibshirani R. On the "degrees of freedom" of the lasso. *Ann. Statist.*, 35(5) :2173–2192, 2007.

Résumé : Le problème général étudié dans cette thèse est celui de la régression linéaire en *grande dimension*. On s'intéresse particulièrement aux méthodes d'estimation qui capturent la *sparsité* du paramètre cible, même dans le cas où la dimension est supérieure au nombre d'observations. Une méthode populaire pour estimer le paramètre inconnu de la régression dans ce contexte est l'estimateur des moindres carrés pénalisés par la norme ℓ_1 des coefficients, connu sous le nom de LASSO. Les contributions de la thèse portent sur l'étude de variantes de l'estimateur LASSO pour prendre en compte soit des informations supplémentaires sur les variables d'entrée, soit des modes semi-supervisés d'acquisition des données. Plus précisément, les questions abordées dans ce travail sont : i) l'estimation du paramètre inconnu lorsque l'espace des variables explicatives a une structure bien déterminée (présence de corrélations, structure d'ordre sur les variables ou regroupements entre variables) ; ii) la construction d'estimateurs adaptés au cadre transductif, pour lequel les nouvelles observations non étiquetées sont prises en considération. Ces adaptations sont en partie déduites par une modification de la pénalité dans la définition de l'estimateur LASSO. Les procédures introduites sont essentiellement analysées d'un point de vue non-asymptotique ; nous prouvons notamment que les estimateurs construits vérifient des *Inégalités de Sparsité Oracles*. Ces inégalités ont pour particularité de dépendre du nombre de composantes non-nulles du paramètre cible. Un contrôle sur la probabilité d'erreur d'estimation du support du paramètre de régression est également établi. Les performances pratiques des méthodes étudiées sont par ailleurs illustrées à travers des résultats de simulation.

Mots-clés : Régression linéaire, sélection de variables, pénalisation, LASSO, Group Lasso, Inégalité de Sparsité Oracle, transduction, prédiction conforme.

Discipline : **Mathématiques**

Abstract : In this thesis, we consider the linear regression model in the *high dimensional* setup. In particular, estimation methods which exploit the *sparsity* of the model are studied even when the dimension is larger than the sample size. The ℓ_1 penalized least square estimator, also known as the LASSO, is a popular method in such a framework which succeeds in providing sparse estimators. The contributions of the thesis concern extensions of the LASSO which take into account either additional information on the entries, or a semi-supervised data acquisition mode. More precisely, the questions considered in this work are : i) the estimation of the regression parameter when correlation or other structures may exist between the variables (presence of correlations, order structure on the variables or grouping of variables) ; ii) the construction of estimators adapted to the transductive setting. These extensions are derived from a modification of the penalty term in the definition of the LASSO. The performance of the methods is theoretically explored from a non-asymptotic point of view ; we prove that the estimators satisfy *Sparsity Oracle Inequalities*. Moreover variable selection consistency is also established. Furthermore, the practical performance of these procedures is illustrated through numerical experiments on simulated datasets.

Key words : Linear regression, variable selection, penalization, LASSO, Group Lasso, Sparsity Oracle Inequality, transduction, conformal prediction.

Laboratoire de Probabilités et Modèles Aléatoires,
CNRS-UMR 7599, UFR de Mathématiques, case 7012
Université Paris 7, Denis Diderot
2, place Jussieu, 75251 Paris Cedex 05.