Machine Learning and Big Data for outlier detection, and applications

Alain Virouleau

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2020IPPAX028

Thèse de doctorat

Apprentissage statistique pour la détection de données aberrantes et application en santé

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Ecole polytechnique

École doctorale n°574 École doctorale de mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées

Thèse présentée et soutenue à Visio-conference, le 18 Juin 2020, par

ALAIN VIROULEAU

Composition du Jury :

Erwan Le Pennec Professeur associé, École polytechnique (CMAP)	Président
Nathalie Vialaneix Directrice de Recherche, INRAE Toulouse (MIAT)	Rapporteur
Yohann De Castro Professeur des Universités, École Centrale de Lyon (ICJ)	Rapporteur
Stéphane Gaïffas Professeur des universités, Université Paris Diderot (LPSM)	Directeur de thèse
Agathe Guilloux Professeur des universités, Université d'Évry Val d'Essonne (LaMME)	Co-directrice de thèse

Remerciements

Je tiens avant tout à remercier chaleureusement mes directrice et directeur de thèse, Agathe Guilloux et Stéphane Gaïffas, pour m'avoir guidé dans mon travail et partagé leurs larges connaissances et expériences dans le domaine des statistiques. Au-delà des aspects statistiques théoriques, j'ai également beaucoup appris auprès d'eux en terme d'implémentation algorithmique de méthodes d'optimisation. Enfin, je veux leur exprimer toute ma reconnaissance pour leur patience et leur soutien durant la (longue !) période entre mes derniers jours à temps plein sur ma thèse et ma soutenance. Je mesure la chance que j'ai d'avoir pu saisir une opportunité professionnelle qui avait peu de chance de se représenter tout en poursuivant mon travail de fin de thèse, qui n'aurait pu aboutir sans leur soutien. Je tiens à remercier également Emmanuel Bacry, qui en plus de m'avoir encadré sur une courte période, m'a permis de faire des rencontres inattendues. Je retiendrai entre autres les pâtisseries de Philippe Conticini et le rapide tour en Tesla.

Je remercie les membres du jury de ma thèse, notamment Nathalie Vialaneix et Yohann De Castro pour l'intérêt qu'ils ont porté à mon travail en acceptant de rapporter ma thèse, dans une période si particulière. Leurs commentaires m'ont permis d'améliorer tant sur le fond que sur la forme certains aspects de mon travail. Je suis honoré de la présence d'Erwan Le Pennec dans mon jury en tant qu'examinateur et le remercie d'avoir accepté ce rôle.

Je suis également reconnaissant envers les chercheuses et chercheurs avec qui j'ai pu collaborer, en particulier Małgorzata Bogdan qui a contribué à l'avancée de mon travail à plusieurs reprises et avec qui cela a été un plaisir d'échanger. Je remercie également l'équipe d'Alex Duval, du Centre de Recherche Saint-Antoine, qui a permis de donner une dimension pratique aux aspects théoriques de ma thèse.

Je remercie également l'équipe administrative du CMAP, notamment Nasséra, Alexandra, Manoëlla, Vincent et Wilfried, pour leur soutien et leur efficacité (surtout devant les buts pour les deux derniers cités).

J'ai une pensée particulière pour les (post-)doctorant.e.s, stagiaires et ingénieur.e.s de la Data Science Initiative avec qui j'ai pu échanger, scientifiquement ou non : Martin, Maryan, Prosper, Philip, Sathiya, Mioty, Ling, Marcello, Peng, Phong, Yiyang. Youcef, merci d'avoir partager tes connaissances tant historiques que footballistiques, qui permettaient de se changer les idées lors des déjeuners. Daniel et Massil, merci d'avoir en plus contribué à écrire en lettres d'argent l'histoire de l'équipe de foot du CMAP. Je n'oublie pas les autres membres du CMAP, que j'ai plus souvent croisés sur le terrain ou en salle café, et dont certains ne savent pas/ont sans doute oublié que je n'ai pas encore soutenu : Cormac, Aymeric, Hélène, Simona, Gustaw, Lucas (le talisman), Lucas et Luca, Othmane, Belhal, Geneviève, Rémi, Martin, Hadrien, Romain, Kevish, et j'en oublie certainement !

Je remercie chaudement mes amis plus anciens, de Cachan et d'avant, qui ont contribué à qui je suis devenu et sans qui j'aurais dévié bien plus tôt vers le monde de l'enseignement, en particulier : Claire, Matthieu, Baptiste, Ludo, Lia, Benji, Jess, Pierre, Loïc, Lilian, Romain, Laurent. Docteur.e.s de cette liste, j'espère que vous saurez apprécier mon effort de vous avoir laissés soutenir avant moi !

Je remercie enfin l'équipe rouge qui m'a accueilli, en espérant qu'elle soit maintenant convaincue par ce transfert, et notamment le Cap', le sanglier des Ardennes et l'ours des Pyrénées, qui gravitent dans les hautes sphères universitaires et dont les rappels à l'ordre hebdomadaires m'ont mené vers la soutenance.

Mes derniers remerciements vont aux personnes qui m'ont soutenu (et me soutiennent encore !) quotidiennement et qui se demandaient quand j'allais enfin terminer cette thèse, Marine, mes parents, mes grands-parents, Bruno, Régine, mes frères et sœurs, Manou. Merci pour vos encouragements et votre soutien indéfectible.

Enfin, quelques personnes que je ne remercie pas : Jeanne, Tony et toutes les personnes liées à l'institut, dont les étudiant.e.s... C'est tellement génial de bosser avec vous que l'on a rarement le temps et l'envie de faire autre chose !

Résumé

Cette thèse porte sur l'étude statistique en grande dimension de données susceptibles de contenir des aberrations.

Le problème de la détection de données aberrantes et celui de régression robuste dans un contexte de grande dimension est fondamental en statistiques et a de nombreuses applications. Dans la lignée de récents travaux proposant de traiter conjointement ces deux problèmes de régression et de détection, nous considérons dans la première partie de ce travail un modèle linéaire gaussien en grande dimension avec ajout d'un paramètre individuel pour chaque observation. Nous proposons une nouvelle procédure pour simultanément estimer les coefficients de la régression linéaire et les paramètres individuels, en utilisant deux pénalités différentes basées toutes les deux sur une pénalisation ℓ_1 ordonnée, nommée SLOPE [11]. Nous faisons l'analyse théorique de ce problème: nous obtenons dans un premier temps une borne supérieure pour l'erreur d'estimation à la fois pour le vecteur des paramètres individuels et pour le vecteur des coefficients de régression. Puis nous obtenons un résultat asymptotique sur le contrôle du taux de fausse découverte et sur la puissance concernant la détection du support du vecteur des paramètres individuels. Nous comparons numériquement notre procédure avec les alternatives les plus récentes, à la fois sur des données simulées et sur des données réelles.

La seconde partie de ce travail est motivée par un problème issu de la génétique. Des séquences particulières d'ADN, appelées *multi-satellites*, sont des indicateurs du développement d'un type de cancer colorectal. Le but est de trouver parmi ces séquences celles qui ont un taux de mutation bien plus élevé (resp. bien moindre) qu'attendu selon les biologistes. Autrement dit, nous voulons aider à identifier deux sortes de séquences aberrantes, respectivement nommées *transformateurs* et *survivants* par les biologistes [52]. Ce problème mène à une modélisation probabiliste non-linéaire et n'entre ainsi pas dans le cadre abordé dans la première partie de cette thèse. Nous traitons ainsi dans cette partie le cas de modèles linéaires généralisés, avec de nouveau des paramètres individuels en plus du prédicteur linéaire, et analysons les propriétés statistiques d'une nouvelle procédure estimant simultanément les coefficients de régression et les paramètres individuels. Nous utilisons de nouveau

la pénalisation SLOPE mais nous nous restreignons au cas de la petite dimension. La performance de l'estimateur est mesuré comme dans la première partie en terme d'erreur d'estimation des paramètres et de taux de fausse découverte concernant la recherche du support du vecteur des paramètres individuels.

Toutes les expériences numériques de ce travail reposent sur l'utilisation d'une librairie open-source écrite en Python et C++ [6].

Abstract

This thesis is devoted to the statistical study of large datasets that contain outlying sample points.

The problems of outliers detection and robust regression in a high-dimensional setting are fundamental in statistics, and have numerous applications. Following a recent set of works providing methods for simultaneous robust regression and outliers detection, we consider in a first part a model of linear regression with individual intercepts, in a high-dimensional setting. We introduce a new procedure for simultaneous estimation of the linear regression coefficients and intercepts, using two dedicated sorted-ℓ_1 penalizations, also called SLOPE [11]. We develop a complete theory for this problem: first, we provide sharp upper bounds on the statistical estimation error of both the vector of individual intercepts and regression coefficients. Second, we give an asymptotic control on the False Discovery Rate (FDR) and statistical power for support selection of the individual intercepts. Numerical illustrations, with a comparison to recent alternative approaches, are provided on both simulated and several real-world datasets.

Our second part is motivated by a genetic problem. Among some particular DNA sequences called *multi-satellites*, which are indicators of the development or colorectal cancer tumors, we want to find the sequences that have a much higher (resp. much lower) rate of mutation than expected by biologist experts. That is, our goal is to help to identify those two kinds of outliers, called *transformators* (resp. *surivors*) by experts [52]. This problem leads to a non-linear probabilistic model and thus goes beyond the scope of the first part. In this second part we thus consider some generalized linear models with individual intercepts added to the linear predictor, and explore the statistical properties of a new procedure for simultaneous estimation of the regression coefficients and intercepts, using again the sorted-ℓ_1 penalization. We focus in this part only on the low-dimensional case and are again interested in the performance of our procedure in terms of statistical estimation error and FDR.

Experiments are conducted using an open-source software written in Python and C++ [6].

Contents

Introduction

The guiding principle of this thesis is to show how the recent convex optimization methods can help solving new robust estimation and outlier detection problem in regression models. While the classical framework of robust estimation problems [3] treat the regression coefficient as the quantity of interest with no modelization of the outlying data, recent healthcare applications [24, 79] point out the fact that outliers could be the thing of interest and thus require specific parametric modelization and guarantees to detect them without making too many mistakes. This mere statement motivates the use of a particular mathematical model called *Mean-Shift outliers (MSO) model* [22] introduced in the $80's$ but that gained some interest very recently thanks to new convex optimization techniques. Let us begin by presenting and motivating the questions on which we want to shed some light in this thesis.

Motivations

Outliers are a fundamental problem in statistical data analysis. Roughly speaking, an outlier is an observation point that differs from the data's "global picture" [36]. A rule of thumb is that a typical dataset may contain between 1% and 10% of outliers [35], or much more than that in specific applications such as web data, because of the inherent complex nature and highly uncertain pattern of users' web browsing [31]. This outliers problem was already considered in the early 50's [23, 30] and it motivated in the 70's the development of a new field called robust statistics [41, 42].

In the linear regression setting, classical estimators, such as the least-squares, are known to fail in presence of outliers [41]. In order to conduct regression analysis in the presence of outliers, roughly two approaches are well-known. The first is based on detection and removal of the outliers to fit least-squares on "clean" data [87]. Pop-

ular methods rely on leave-one-out methods (sometimes called case-deletion), first described in [22] with the use of residuals in linear regression. The main issue about these methods is that they are theoretically well-designed for the situations where only one given observation is an outlier. Repeating the process across all locations can lead to well-known masking and swamping effects [34]. An interesting recent method that does not rely on a leave-one-out technique is the so-called IPOD [73], a penalized least squares method with the choice of tuning parameter relying on a BIC criterion. This method relies on the so-called Mean-shift outliers model. A second approach is based on robust regression, that considers loss functions that are less sensitive to outliers [42]. This relies on the M-estimation framework, that leads to good estimators of regression coefficients in the presence of outliers, thanks to the introduction of robust losses replacing the least-squares. However, the computation of M-estimates is substantially more involving than that of the least-squares estimates, which to some extend counter-balance the apparent computational gain over previous methods. Moreover, robust regression focuses only on the estimation of the regression coefficients, and does not allow directly to localize the outliers, see also for instance [91] for a recent review.

We might thus ask ourselves the following questions.

Question 1. *What benefits are brought by the Mean-shift outliers model compared to the classical robust regression techniques and how is it linked with the variable selection problem ?*

Alternative approaches have been proposed to perform simultaneously outlier detection and robust regression. Such methods involve median of squares [74], S-estimation [69] and more recently robust weighted least-squares [28], among many others, see also [33] for a recent review on such methods. The MSO model and related techniques such as IPOD [73] also perform both outlier detection and robust estimation. However, many high-dimensional datasets, with hundreds or thousands of covariates, do suffer from the presence of outliers. Robust regression and detection of outliers in a high-dimensional setting is therefore important. Increased dimensionality and complexity of the data may amplify the chances of an observation being an outlier, and this can have a strong negative impact on the statistical analysis. In such settings, many of the aforementioned outlier detection methods do not work well. A

new technique for outlier detection in a high-dimensional setting is proposed in [1], which tries to find the outliers by studying the behavior of projections from the data set. A small set of other attempts to deal with this problem have been proposed in literature [86, 67, 32, 73, 26], and are described below with more details.

The MSO model seems well-suited to perform both estimation and outlier detection because it relies on parameters for regression and also for outlyingness of each observation. High-dimensional datasets can be handled because all the parameters can be penalized. In this setting, we can either consider the MSO model as a ultra-high dimensional linear regression model and apply one penalization on a concatenated version of the parameters as suggested in IPOD, or use two different penalizations for regression parameters and outlyingness parameters, as suggested in Robust Lasso [32].

Question 2. *How recent convex sparsity-inducing penalization can help to show strong guarantees in the MSO model for both outlier detection problem and robust estimation in the classical Gaussian Linear Model (LM) ?*

Sparse inference techniques, in particular applied to high-dimensional linear regression, are of importance in statistics, and have been an area of major developments over the past two decades, with deep results in the field of compressed sensing, and more generally convex relaxation techniques [80, 15, 16, 19, 18]. These led to powerful inference algorithms working under a sparsity assumption, thanks to fast and scalable convex optimization algorithms [4]. The most popular method allowing to deal with sparsity and variable selection is the LASSO [81], which is ℓ_1-penalized least-squares, with improvements such as the Adaptive LASSO [94], among a large set of other sparsity-inducing penalizations [13, 5].

Within the past few years, a large amount of theoretical results have been established to understand regularization methods for the sparse linear regression model, using so-called oracle inequalities for the prediction and estimation errors [43, 44, 53], see also [13, 29] for nice surveys on this topic. Another line of works focuses on variable selection, trying to recover the support of the regression coefficients with a high probability [49, 44, 21]. Other types of loss functions [85] or penalizations [25, 11] have also been considered. Very recently, the sorted-ℓ_1 norm penalization has been introduced [11, 12, 76] and very strong statistical properties have been shown. In

particular, when covariates are orthogonal, SLOPE allows to recover the support of the regression coefficients with a control on the False Discovery Rate [11]. For i.i.d covariates with a multivariate Gaussian distribution, oracle inequalities with optimal minimax rates have been shown, together with a control on a quantity which is very close to the FDR [76]. For more general covariate distributions, oracle inequalities with an optimal convergence rate are obtained in [14].

Question 3. *How can we generalize results in the context of Question 2 to other regression problem, particularly to Generalized Linear Models (GLM) ?*

Real datasets, particularly in biological applications, often differ from a linear regression model. For binary data, a Bernoulli or binomial model should be used. Therefore, studying to what extent our results can be adapted in the GLM setting is interesting.

Outline

Each question presented above corresponds to a chapter of the thesis. Let us now rapidly review the main contents and results of this thesis.

1 Summary of Chapter I

In Chapter I, we basically answer to Question 1. We review some statistical tools for outlier detection in the context of Linear Regression. We particularly focus on problems that have convex objectives, this restriction being at the core of much of modern optimization theory. The primary reasons for targeting convex problems are their widespread use in applications and their relative ease of solving them.

1.1 Convex optimization tools

First, many supervised machine learning problems can be cast into the minimization of an expected loss over a data distribution, possibly penalizing some parameter, writing

$$\min_{\theta \in \mathbb{R}^d} F(\theta) = f(\theta) + h(\theta), \tag{1}$$

where f is a goodness-of-fit measure depending implicitly on some observed data and h is a regularization term that imposes some structure to the solutions. Typically, f is a differentiable function with a Lipschitz gradient, whereas h might be non-smooth. Most machine learning optimization problems involve a data fitting loss function f averaged over sample points because of the empirical risk minimization principle [63]. Namely, the function f writes

$$f(\theta) = \frac{1}{n} \sum_{i=1}^{n} f_i(\theta),$$

where n is the number of observations, and f_i is the loss associated to the i^{th} observation. In regression context, n independent and identically distributed observations $(x_i, y_i)_{i=1,\dots,n}$ are given, where $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$ respectively stand for the vector of covariates and label of sample i. Each loss f_i is then of the form $\ell(y_i, x_i^\top \theta)$ where typical examples of functions ℓ are:

- $\ell(y, y') = (y - y')^2$ (Least Square loss),
- $\ell(y, y') = \log\left(1 + e^{-yy'}\right)$ (logistic loss with labels in $\{-1, 1\}$),

Typical examples of function h include sparsity inducing penalization - such as the ℓ_1 penalization where $h(\theta) = \sum_{i=1}^{d} |\theta_i|$. Such methods have been applied in the popular goal of variable selection in various applications [80, 54]. Thus we recall some convex optimization principle that will be used throughout the thesis.

Gradient Descent (GD) is the building block of the main first-order optimization algorithms. Starting at some initial point θ^0, this algorithm minimizes a differentiable function f by iterating the following equation

$$\theta^{t+1} = \theta^t - \eta_t \nabla f(\theta^t), \tag{2}$$

where $\nabla f(\theta)$ stand for the gradient of f evaluated at θ, and $(\eta_t)_t$ is a sequence of step-sizes.

As emphasized by Equation (1), the function to be optimized can be non-smooth in many situations because of the presence of a regularizing term. GD algorithm

can then be extended to cases where the function h is convex and non-differentiable whose *proximal operator* is easy to compute.

Definition .1. *Given a convex function h, we define its proximal operator as*

$$\text{prox}_h(x) = \underset{y}{\text{argmin}} \left[h(y) + \frac{1}{2}\|x - y\|^2 \right],$$

which is uniquely defined because of the strong convexity of the Euclidean norm.

The proximal operator can be seen as a generalization of the projection. Indeed, if $h = 0$ on a convex set $\mathscr{C}$ and $h = \infty$ on $\mathscr{C}^{\complement}$, prox_h is exactly the projection over $\mathscr{C}$. The computation of the proximal operator is also an optimization problem, but when the function h is simple enough, the proximal operator has a closed form solution [5]. Using these proximal operators, the iteration of Equation (2) is then replaced by the following iteration:

$$\theta^{t+1} = \text{prox}_{\eta_t h}\left(\theta^t - \eta_t \nabla f(\theta^t)\right). \tag{3}$$

1.2 Background on robust linear regression

We consider a linear model given by:

$$y_i = x_i^\top \beta + \varepsilon_i, \tag{4}$$

for $i = 1, \ldots, n$, where n is the sample size, $\beta \in \mathbb{R}^p$, $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$ and $\varepsilon_i \in \mathbb{R}$ respectively stand for the linear regression coefficients, vector of covariates, label and noise of sample i. The idea of robust regression is to change the usual least squares goodness of fit to another goodness of fit that will reduce the influence of outliers. These methods gain popularity since the $80's$ and are variants of M-estimation [40], which computes the following estimate of β:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \sum_{i=1}^{n} \ell(y_i, x_i^\top \beta), \tag{5}$$

with a loss function $\ell(y_i, \cdot)$ which is often non-convex.
This kind of method focuses on the estimation of the regression parameter and does not consider outlier detection as a problem of interest. In the context of fraud

detection, anomaly detection, outliers are quantities of interest and thus should be included in the model.

1.3 The Mean-Shift outlier model

The *mean-shift single outlier* model [22] is the first model introducing a parameter to decide whether a particular observation is an outlier or not. The model writes:

$$y_i = x_i^\top \beta + \mu_{i_0} + \varepsilon_i, \tag{6}$$

for $i = 1, \dots, n$, where n is the sample size, $\beta \in \mathbb{R}^p$, $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$ and $\varepsilon_i \in \mathbb{R}$ respectively stand for the linear regression coefficients, vector of covariates, label and noise of sample i, and where i_0 is the index of the observation we suspect to be an outlier. If μ_{i_0} is non zero then observation i_0 is an outlier. With the development of convex optimization method and sparsity inducing penalization, this model has been extended to the following *mean-shift outlier* model:

$$y_i = x_i^\top \beta + \mu_i + \varepsilon_i, \tag{7}$$

with $\mu = (\mu_1, \mu_2, \dots, \mu_n)^\top \in \mathbb{R}^n$ in which a non-zero coordinate indicates that the corresponding observation is an outlier. This model rewrites as a high-dimensional linear regression model: with $Y = (y_1, y_2, \dots, y_n)^\top$, with $X \in \mathbb{R}^{n \times p}$ with i^{th} row given by $x_i^\top$, with an extended features matrix $Z = [X\ I]$ being the concatenation of X and the identity matrix, and an unknown regression vector $\gamma = (\beta^\top, \mu^\top)^\top$, the model given in Equation (7) rewrites

$$Y = Z\gamma + \varepsilon,$$

where $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)^\top$.

The outlier detection problem is then a variable selection problem, therefore we recall some classical results about estimation and variable selection, focusing on the approach of minimization of the penalized negative log-likelihood using popular sparsity-inducing penalizations. In this context, the optimization problem takes the following form

$$\min_\gamma \frac{1}{2n} \| Y - Z\gamma \|_2^2 + \mathrm{pen}(\gamma),$$

where pen is usually a sparsity-inducing and convex function.

We end this chapter by a discussion motivating the fact that we should apply two different penalizations instead of considering the concatenated problem above. This has already been done successfully with the Lasso penalization in [32], in which the minimization problem is of the following form

$$\min_{\beta,\mu} \frac{1}{2n}\|Y - X\beta - \mu\|_2^2 + \lambda_\beta\|\beta\|_1 + \lambda_\mu\|\mu\|_1,$$

where $\|x\|_1 = \sum_{i=1}^n |x_i|$ for any $x \in \mathbb{R}^n$. A critical step is the choice of the tuning parameters $\lambda_\beta, \lambda_\mu$ in the penalizations. Traditionally, it is achieved by a cross-validation technique. We discuss in Section 4.3 of Chapter I this technique since in our context data is non-stationary.

1.4 SLOPE

SLOPE is the acronym for Sorted L-One norm PEnalization [11], which is defined as follows.

Definition .2 ([12]). *Let $x \in \mathbb{R}^n$. The sorted ℓ_1 norm associated to the positive non-increasing sequence $\lambda = (\lambda_1, \dots, \lambda_n)$ is defined as:*

$$J_\lambda(x) = \sum_{i=1}^n \lambda_i |x|_{(i)}, \tag{8}$$

where $|x|_{(i)}$ is the ith largest element of $|x| = (|x_1|, \dots, |x_n|)$.

This norm has demonstrated interesting properties in terms of estimation and support recovery in linear regression with specific design matrix. In Section 1.3 of Chapter I we recall these results that motivate its future use.

2 Summary of Chapter II

In Chapter II, we deeply study the *Mean-Shift outlier* model of Equation (7). This model gained interest recently with new developments in convex optimization [46, 73, 90]. Recall that Equation (7) also writes:

$$Y = X\beta + \mu + \varepsilon, \tag{9}$$

where $Y = (y_1, y_2, \ldots, y_n)^\top$, $X \in \mathbb{R}^{n \times p}$ with i^{th} row given by $x_i^\top$, $\mu = (\mu_1, \mu_2, \ldots, \mu_n)^\top$ and $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^\top$.

We answer to Question 2 by studying the properties of $\hat{\mu}$ and $\hat{\beta}$ being solution to the following minimization problem:

$$(\hat{\beta}, \hat{\mu}) \in \operatorname*{argmin}_{\beta,\mu} \frac{1}{2n} \| Y - X\beta - \mu \|_2^2 + J_\lambda(\beta) + J_{\tilde{\lambda}}(\mu), \tag{10}$$

where J_λ and $J_{\tilde{\lambda}}$ for two sequences $\lambda, \tilde{\lambda}$ of weights denotes the Slope norm given in Definition .2.

We study the properties of $\hat{\beta}$ and $\hat{\mu}$ both in terms of estimation and variable selection/outlier detection. Assuming the true parameters β and μ are respectively $k-$sparse and $s-$sparse, our goal is to estimate properly not only the true parameters but also the true support of μ, as it describes the outlying sample points.

2.1 Estimation results

The metrics used to measure the performance in term of estimation is the ℓ_2 norm. In Section 3 of Chapter II we obtain the following error bound in Theorem II.4:

$$\| \hat{\beta} - \beta \|_2^2 + \| \hat{\mu} - \mu \|_2^2 = O\left(k \log\left(\frac{2ep}{k} \right) + s \log\left(\frac{2en}{s} \right) \right),$$

where $(\hat{\beta}, \hat{\mu})$ is given by Equation (10). This convergence rate is typical of what can be found in the literature about parametric regression problems [44, 66]. This result is based on a RE-type condition that we establish in Theorem II.1 (Section 3, Chapter II), and that is known to be mandatory in order to derive fast rates of convergence for penalizations based on the convex-relaxation principle [92]. We establish similar convergence rates for other combinations of penalties in Theorem II.2 and Theorem II.3 (Section 3 of Chapter II).

2.2 Outlier detection results

The previous result makes sense regarding estimation, but not outlier detection since it does not guarantee that we will recover the non-zero coefficients of μ. For the purpose of outlier detection we introduce the *support* and the *sign* of a vector.

Definition .3. *The support of $x \in \mathbb{R}^n$ is*

$$\mathrm{supp}(x) = \#\big\{i \in \{1,\dots,n\} \mid x_i \neq 0\big\},$$

where # stands for the cardinality, while the signed support of $x \in \mathbb{R}^n$ is

$$\mathrm{sgn}(x) = \big(\mathrm{sgn}(x_1), \mathrm{sgn}(x_2), \dots, \mathrm{sgn}(x_n)\big),$$

where for any $t \in \mathbb{R}$, $\mathrm{sgn}(t) = 1$ if $t > 0$, $\mathrm{sgn}(t) = -1$ if $t < 0$ and $\mathrm{sgn}(0) = 0$.

The purpose of outlier detection is to find an estimate $\hat{\mu}$ such that $\mathrm{supp}(\hat{\mu}) = \mathrm{supp}(\mu)$. We use the False Discovery Rate (FDR, [8]) as a metric, which is the expectation of false discoveries among all the discoveries, defined by

$$\mathrm{FDR}(\hat{\mu}) = \mathbb{E}\left[\frac{\#\{i \mid \mu_i = 0 \text{ and } \hat{\mu}_i \neq 0\}}{\#\{i \mid \hat{\mu}_i \neq 0\}}\right]. \tag{11}$$

Note that in our context, a discovery is a non-zero coefficient in μ since it corresponds to finding an outlying sample point. We also define the True Positive Rate, which is the expected proportion of outliers found, defined as:

$$\mathrm{TPR}(\hat{\mu}) = \mathbb{E}\left[\frac{\#\{i \in \{1,\dots,n\} \mid \hat{\mu}_i \neq 0 \text{ and } \mu_i \neq 0\}}{\#\{i \in \{1,\dots,n\}, \mu_i \neq 0\}}\right]. \tag{12}$$

The main result of Section 4 of Chapter II is given in Theorem II.5 in which we establish that for any target level q, we can set the sequence of weights λ and $\tilde{\lambda}$ in minimization (10) to obtain

$$\mathrm{TPR}(\hat{\mu}) \to 1, \qquad \limsup \mathrm{FDR}(\hat{\mu}) \leq q.$$

These theoretical results are supported by intensive numerical experiments. In particular, two interesting questions are raised.

- How much can we lower the outlier magnitude and still be able to find them ? The answer is basically that it can be lowered as much as it does not confound with the Gaussian noise.
- To what extent is this result really asymptotic ? Numerical experiments suggest that in a low-dimensional setting (few covariates) a small sample size is enough

to control the FDR. However, the higher the dimension, the higher the sample size.

2.3 Noise variance

The choice of the weights λ and $\tilde{\lambda}$ in the penalization (10) relies on the knowledge of the noise variance σ^2 [32, 11]. While σ^2 is typically unknown in practice, cross-validation can overcome this issue, but we explain in Section 4.3 that it is not an option in our context because of non-stationarity .

In low-dimensional settings, the estimation of the noise variance is not an issue [40, 71]. However the task of outlier detection in high-dimensional settings goes beyond the traditional robust analysis which requires a large number of observations relative to the dimensionality [73, 42]. In Section 5.7 of Chapter II we propose Algorithm 4, a new algorithm based on successive steps of estimating the model parameters and updating the noise variance, in the spirit of [11]. Although no theoretical results are shown for this algorithm, its performance is again measured through intensive numerical simulations.

3 Summary of Chapter III

Chapter III answers to Question 3. The linear regression model is the first model we investigated on the outlier detection problem because it is the most familiar model in the regression context and it is widely used in many fields of science. However it remains a particular case, while other problems, such as binary classification, are also of paramount importance.

One example is when from a binary outcome $Y \in \{0, 1\}$ one wants to estimate the probability $\mathbb{P}(Y = 1)$. This is motivated by a problem in biology: we consider whole exome sequencing data for 47 primary colorectal cancer tumors, characterized by a global genomic instability affecting repetitive DNA sequences (also known as microsatellite unstable tumors, see [24]). In details, micro-satellites are portions of DNA sequence that are composed of a base motif (one or several nucleotides) repeated several times (generally 5 to 50). For example, $AAAAA$ is a micro-satellite with the base motif A (Adenine) repeated five times. Such portions of DNA have higher mutation rate than other DNA sequences, leading to genetic diversity (instability).

Here, the binary outcome Y describes whether a micro-satellite has been mutated or not. The purpose is to estimate the mutation rate and find outliers, which are micro-satellites that are mutated more or less than expected. This requires the use of a logistic model [24, 52], which is a particular case of the Generalized Linear Model we study in Chapter III.

3.1 Outliers in Generalized Linear Models

The Generalized Linear Model (GLM) [61] generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value. The corresponding log-density is given by:

$$\log f(y; x, \beta) = y x^\top \beta - b(x^\top \beta) + c(y), \tag{13}$$

where $y \in \mathbb{R}$ is the label, $x \in \mathbb{R}^p$ the vector of covariates, $\beta \in \mathbb{R}^p$ the regression coefficients, b a twice continuously differentiable function with derivative b' being a one-to-one function, and c a normalization function. Typical examples are:

- Gaussian linear model with variance σ^2:

$$b(\eta) = \frac{\eta^2}{2\sigma^2}, \qquad c(y) = -\log\left(\sqrt{2\pi\sigma^2}\right) - \frac{y^2}{2\sigma^2}.$$

- Logistic regression:

$$b(\eta) = \log(1 + e^\eta), \quad c(y) = 1.$$

Following recent work on modelling outliers in GLM [88], we include parameters to take into account the presence of outliers in the following way: we have labels $y = (y_1, ..., y_n)^\top$ whose elements are observations of independent random variables from a distribution with log-density

$$\log f(y_i; x_i, \beta^\star, \mu_i^\star) = y_i(x_i^\top \beta + \mu_i) - b(x_i^\top \beta + \mu_i) + c(y_i), \tag{14}$$

for $i = 1, \ldots, n$, where n is the sample size. A non-zero μ_i means that observation i is an outlier, and $\beta \in \mathbb{R}^p$, $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$ respectively stand for the regression coefficients,

vector of covariates, label of sample i. We assume that $\mu^\star \in \mathbb{R}^n$ is sparse with support S and that $|S| = s \ll n$.

3.2 Outlier detection results

We focus in this part on the outlier detection property of the following penalized negative log-likelihood estimator

$$(\hat{\beta}, \hat{\mu}) \in \underset{\beta \in \mathbb{R}^p, \mu \in \mathbb{R}^n}{\operatorname{argmin}} -\frac{1}{n} \sum_{i=1}^{n} \left(y_i (x_i^\top \beta + \mu_i) - b(x_i^\top \beta + \mu_i) \right) + J_\lambda(\mu), \tag{15}$$

where J_λ is the Slope penalization given in Definition .2 for a positive non-increasing sequence λ. We focus on the low-dimensional case therefore we do not apply any penalization on the regression coefficients.

Under assumptions on the distribution of the labels given in Section 2.3 of Chapter III, we establish in Theorem III.1 the following result: for any fixed target level α, we have

$$\mathrm{TPR}(\hat{\mu}) \to 1, \qquad \limsup \mathrm{FDR}(\hat{\mu}) \leq \alpha, \tag{16}$$

where $(\hat{\beta}, \hat{\mu})$ is given by Equation (15), with λ depending on α. This basically means that in appropriate settings we can recover the true support of $\mu^\star$ while keeping the false discovery rate under a desired level.

3.3 The Binomial model and biological context

One particular model that satisfies the required assumptions is the binomial model, where observations come from independent Binomial distributions. The log-density is then given by Equation (14) with:

$$b(\eta) = -n_s \log(1 - \sigma(\eta)), \quad \sigma(\eta) = \frac{1}{1 + e^{-\eta}}, \quad c(y) = \log \binom{n_s}{y},$$

where n_s is the number of trials. In Section 5 of Chapter III, we perform intensive simulations to illustrate the theoretical results, in particular regarding FDR control.

The interest in this model comes from a particular biological context. In the study of colorectal cancer, one is interested in a global genomic instability affecting repetitive DNA sequences. Hence, one observation consists in computing over n_s

patients the number of times a particular DNA sequence has been mutated, thus leading to a binomial random variable. There are then as many observations as DNA sequences of interest.

The particular dataset we study contains repetitive sequences (of different lengths) of the single nucleotide A, observed in $n_s = 47$ colorectal cancer tumors. Based on the observed mutation rate of these sequences across the 47 tumors, the goal is to detect sequences that are much more (or less) mutated than they should. The results are the following: over more than 45 thousands DNA sequences, we identify 151 outliers, see Figure A, which must be subject to further biological analysis. However, the plot indicates some kind of overdispersion in the data, a phenomenon which is not included in this model. The inclusion of overdispersion such as in [59] is beyond the scope of this thesis and is to be developed in future works.

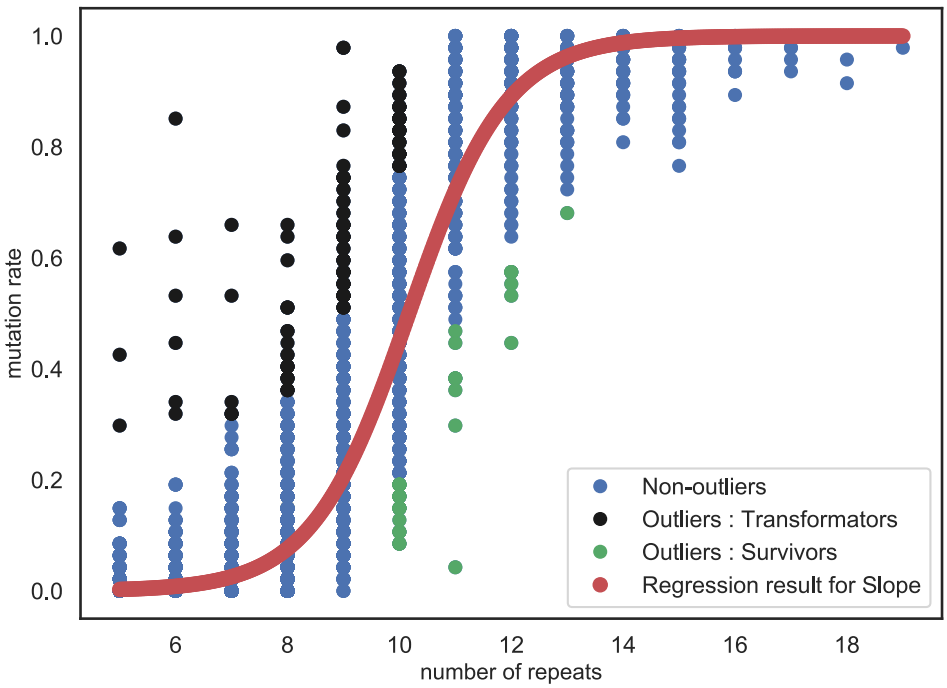

Figure A: Multi-satellites with base motif A: identification of 151 outliers for a target FDR $q = 0.05$. DNA sequences composed of repetitions of nucleotide A are observed in 47 tumors. Mutation rates are plotted against the length of the DNA sequence. Red curve is the regression given by $\hat{\beta}$ given by Equation (15). Blue dots are non-outlier sample points. Green dots correspond to DNA sequences that mutated less than expected. Black dots correspond to DNA sequences that mutated more than expected. The last two categories are different kind of outliers.

"Y en a le même nombre que les autres. Parce que la farce est la même pour tous les saucissons. Avec une moyenne de trente-deux à trente-quatre noisettes par pièce. Seulement, avec le hasard de la coupe, vous êtes tombé sur une tranche où les éclats étaient mal répartis."

— Perceval, *Kaamelott, Livre VI*

CHAPTER I

A review of Statistical Tools for Outliers Detection

1 Notations and Technical tools

In this Section we present tools that are used as the building blocks of this thesis.

1.1 Notations

Unless it is explicitly mentioned, n, m, p denote positive integers, Y denotes an observation vector in $\mathbb{R}^n$, X a design matrix in $\mathbb{R}^{n\times p}$ whose columns correspond to covariates. We denote $\|x\|_q = (\sum_{i=1}^n |x_i|_q)^{1/q}$ the ℓ_q norm of any $x \in \mathbb{R}^n$ for $q \in \mathbb{N}^\star$, and $|x|_0$ the ℓ_0 "norm", namely the cardinality of $\{i \mid x_i \neq 0\}$. The Euclidean inner product in $\mathbb{R}^n$ is denoted by $\langle u, v\rangle$ or $u^\top v$ for any $u, v \in \mathbb{R}^n$.

1.2 Convex optimization tools

Convex optimization is important for training statistical learning model [5]. Convex relaxations techniques, such as Lasso, allowed to solve model selection problem through convex optimization [5]. In statistics, model selection is the task of selecting a subset of covariates that perform best according to a given criteria. A common procedure is to introduce a minimization problem with an objective function defined as the weighted sum of two components: a term responsible for the goodness-of-fit (typically the negative log-likelihood of the model), and the second term which is

a sparsity-inducing penalization on parameters of the model so that a null coefficient corresponds to an irrelevant variable. For least-square regression, the objective function takes the following form:

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2} \|Y - X\beta\|_2^2 + \text{pen}(\beta), \tag{I.1}$$

where pen is any penalization function.

The theory of convex optimization guarantees the existence of a global minimum if the penalization is convex. The most natural idea for model selection is the penalization $\text{pen}(\beta) = |\beta|_0$. However, this penalization is not convex and so a popular penalization is the convexified version of $|\beta|_0$, that is the ℓ_1 norm $\|\beta\|_1$, which leads to the popular Lasso minimization [81, 75, 43, 27]. Results about Lasso will be considered in Section 3.2, together with results about the Slope penalization introduced in Section 1.3.

Subdifferential. As enlighted by the example above, the penalization function can be non-smooth, therefore we recall here the definition and some properties of the subdifferential, that generalizes the notion of gradient to non-smooth function. The following definitions, properties and algorithms are based on [4].

Definition I.1. *Given a convex function $g : \mathbb{R}^p \to \mathbb{R}$ and a vector $x \in \mathbb{R}^p$, let us define the subdifferential of g at x as*

$$\partial g(x) := \{z \in \mathbb{R}^p \mid \forall x' \in \mathbb{R}^p \, g(x) + \langle z, x' - x \rangle \leq g(x')\}. \tag{I.2}$$

The elements of $\partial g(x)$ are called the subgradients of g at x.

The proposition below states that the subgradient allows to find the optimum of a (possibly non-smooth) convex function:

Proposition I.1. *For any convex function $g : \mathbb{R}^p \to \mathbb{R}$, a point $x \in \mathbb{R}^p$ is a global minimum of g if and only if the condition $0 \in \partial g(x)$ holds.*

Note that the concept of subgradient is mainly useful for nonsmooth functions. If g is differentiable at x, the set $\partial g(x)$ is indeed the singleton $\{\nabla g(x)\}$ and Proposition I.1 above reduces to the classical first-order optimality condition $\nabla g(x) = 0$.

As we explained above, the penalization term in Equation (I.1) is typically a sparsity-inducing norm. Therefore we must be able to compute the subgradient of such a norm. This leads to the notion of *dual norm* below:

Definition I.2. *The dual norm $J^\star$ of a norm J is defined for any vector $z \in \mathbb{R}^p$ by*

$$J^\star(z) := \max_{x \in \mathbb{R}^p, J(x) \leq 1} \langle z, x \rangle, \tag{I.3}$$

which allows to caracterize the subdifferential of a norm:

Proposition I.2. *The subdifferential of a norm J in any $x \in \mathbb{R}^p$ is given by*

$$\partial J(x) = \{z \in \mathbb{R}^p \mid J^\star(z) \leq 1, \langle x, z \rangle = J(x)\}. \tag{I.4}$$

Proximal gradient method. For smooth minimization problems, a fundamental method is gradient descent. For non-smooth problem, it generalizes to subgradient descent as long as an element of the subgradient can be found. However, faster methods called proximal methods [4] have been recently introduced and only rely on the computation of a *proximal operator* defined below:

Definition I.3. *The proximal operator of a convex function J, with parameter $\lambda > 0$, is given by*

$$\text{prox}_{\lambda J}(x) := \operatorname*{argmin}_{u \in \mathbb{R}^p} \frac{1}{2}\|x - u\|_2^2 + \lambda J(u), \tag{I.5}$$

where $x \in \mathbb{R}^p$.

Since the objective function is strictly convex, the proximal operator is uniquely defined. Now let us consider a minimization problem of the form:

$$\min_{\beta \in \mathbb{R}^p} f(\beta) + \lambda J(\beta), \tag{I.6}$$

with J a convex function for which the proximal operator can be computed and f a smooth convex function whose gradient ∇f is L-lipschitz, namely:

$$\|\nabla f(u) - \nabla f(v)\| \leq L\|u - v\|, \tag{I.7}$$

for any $u, v \in \mathbb{R}^p$. Note that the objective function Equation (I.1) is a particular case of the one in Equation (I.6).

This minimization problem can be solved iteratively. Having a guess β^t at step t, β^{t+1} is computed by minimizing a first-order Taylor expansion of f around β^t:

$$\beta^{t+1} = \min_{\beta \in \mathbb{R}^p} f(\beta^t) + \langle \nabla f(\beta^t), \beta - \beta^t \rangle + \frac{L}{2}\|\beta - \beta^t\|_2^2 + \lambda J(\beta),$$

which can be rewritten as:

$$\beta^{t+1} = \min_{\beta \in \mathbb{R}^p} \frac{1}{2}\left\|\beta - (\beta^t - \frac{1}{L}\nabla f(\beta^t))\right\|_2^2 + \frac{\lambda}{L} J(\beta). \tag{I.8}$$

Then Equation (I.8) leads to the use of the proximal operator in the following algorithm:

Algorithm 1 Proximal Gradient Descent

initialize β
while not converged **do**
  $\beta \leftarrow \text{prox}_{\frac{\lambda}{L}J}(\beta - \frac{1}{L}\nabla f(\beta))$
end while
return β

In some problem such as linear regression or logistic regression, L can be easily computed. When L is harder to find, one can use linesearch [4].

1.3 The Sorted L-One norm

This Section contains technical details about the penalization we will use, which has been introduced and developed in [12], and therefore will be used in the proof of the main results of Chapter II. In the rest of this Section, we consider $\lambda = (\lambda_1, \dots, \lambda_n)$ a non-increasing sequence of positive real numbers.

Definition I.4 ([12]). *Let $x \in \mathbb{R}^n$. The sorted ℓ_1 penalization (Slope) associated to the sequence λ is defined as:*

$$J_\lambda(x) = \sum_{i=1}^{n} \lambda_i |x|_{(i)}, \tag{I.9}$$

where $|x|_{(i)}$ is the ith largest absolute value of the elements of x.

Note that this includes ℓ_1 norm as a special case if λ is constant, but if not this means that the higher the coordinate (in absolute value), the higher the individual weight in the penalization.

Proposition I.3 ([12]). *J_λ is a norm.*

In particular, J_λ is a convex function.

Definition I.5 ([76]). *A vector $a \in \mathbb{R}^n$ is said to majorize $b \in \mathbb{R}^n$ (denoted $b \preccurlyeq a$) if they satisfy for all $i \in \{1, \dots, n\}$:*

$$|a|_{(1)} + \cdots + |a|_{(i)} \geq |b|_{(1)} + \cdots + |b|_{(i)}. \tag{I.10}$$

Proposition I.4 ([11]). *The unit ball of the dual norm of J_λ is:*

$$\mathscr{C}_\lambda = \{v \in \mathbb{R}^n \mid v \preccurlyeq \lambda\}. \tag{I.11}$$

The property above is important as it allows to describe the subgradient of the J_λ norm, which is, as seen in the previous subsection, a crucial tool in convex optimization as it generalizes the gradient to non-differentiable functions. Hence, as a particular case of Proposition I.2, we conclude with the following property:

Proposition I.5. *The subdifferential of the J_λ norm at any $x \in \mathbb{R}^n$ is given by:*

$$\partial J_\lambda(x) = \{\omega \in \mathscr{C}_\lambda \mid \langle \omega, x \rangle = J_\lambda(x)\}. \tag{I.12}$$

As explained in Section 1.2 above, the proximal operator is a convenient tool to solve penalized optimization problems. A fast algorithm to compute the proximal operator of J_λ has been developed [12], making this optimization problem easy to solve. Note that this algorithm has been implemented in the open-source `tick` library [6], available at https://x-datainitiative.github.io/tick/ that we will use in our experiments.

1.4 Multiple testing

The concept of multiple testing arises naturally in model selection. Let us consider Equation (I.1) with a sparsity-inducing penalization on β, and let us measure the

efficiency of our estimator in terms of variable selection. Formally we want to perform p tests with hypothesis

$$H_{0,i} : \beta_i = 0 \qquad \text{versus} \qquad H_{1,i} : \beta_i \neq 0, \quad i = 1, \dots, p, \tag{I.13}$$

and decide to reject or not based on our estimator of β. This is a multiple testing problem and several measures of efficiency can be considered. We recall in the following the basics of multiple testing and some interesting error measures to be controlled.

Single test. Suppose that we test some null hypothesis H_0. Let R denotes the rejection of H_0. A *discovery* is a rejection of this hypothesis. This name is motivated by the fact that in applications, a null hypothesis corresponds to something expected, a global trend, and so a rejection can correspond to a new trend that needs to be explored. In the linear regression example above this would correspond to the discovery of one variable of influence. A *false discovery* arises when H_0 is rejected whereas it is true.

When doing a single statistical test, one typically wants to control the *Type I* error, namely the probability of making a false discovery, by a certain small level $\alpha > 0$ (which is typically of the order of 5%):

$$\mathbb{P}_{H_0}(R) \leq \alpha.$$

Note that most of the time a statistical test is described by its *p-value*.

Definition I.6. *The p-value p_{val} of a test is the smallest α that leads to the rejection of H_0, all other things being kept unchanged.*

The proposition above gives an example of a computable p-value:

Proposition I.6. *Consider a two-sided test $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$ with a test statistic $\mathcal{T}$. If the test statistic is distributed as $\mathcal{N}(0,1)$ under the null hypothesis, then the p-value is:*

$$p_{val} = 2(1 - \Phi(\mathcal{T}^{obs})),$$

where Φ is the c.d.f of $\mathcal{N}(0,1)$ and $\mathcal{T}^{obs}$ is the observed value of the statistic.

Note that α is always fixed *a priori*, that is before performing the test. The p-value is not always computable but when it is, one can equivalently decide to reject H_0 if $p_{val} \leq \alpha$. Therefore p-value is a convenient way to describe a test and the lower the p-value the stronger the rejection.

Multiple tests. Now suppose we perform m tests, each of level α, that is:

$$\mathbb{P}_{H_{0,i}}(R_i) \leq \alpha,$$

where $H_{0,i}$ and R_i respectively stand for the null hypothesis and the rejection of the i-th test. Then, under independence, the probability of making at least one false rejection is $1-(1-\alpha)^m$, meaning that we will make a false discovery with high probability. In such a framework, it is not clear what quantity is relevant to be controlled at some level α. Two popular possible quantities [38] are as follows.

Definition I.7. *The Family-wise Error Rate (FWER) is defined as:*

$$\text{FWER} = \mathbb{P}\Big(\bigcup_{i=1}^{m}(H_{0,i} \text{ is true } \cap R_i\Big). \tag{I.14}$$

This is the probability of making at least one false discovery. The False Discovery Rate (FDR) is defined as:

$$\text{FDR} = \mathbb{E}\Big[\frac{|\mathscr{H}_0 \cap \mathscr{R}|}{|\mathscr{R}|}\Big], \tag{I.15}$$

where $\mathscr{H}_0$ and $\mathscr{R}$ are respectively the set of true null hypotheses and the set of rejected hypotheses, namely

$$\mathscr{H}_0 = \{i \mid H_{0,i} \text{ is true}\}, \qquad \mathscr{R} = \{i \mid H_{0,i} \text{ is rejected}\}.$$

The FDR is the expected number of the proportions of false discoveries among all the discoveries.

A popular and still widely used procedure to control the FWER at level α is the *Bonferroni correction*[1], consisting in doing each of the m tests at level α/m, so that the

[1]Which is not due to the mathematician whose name has been given to this method (as often in Mathematics particularly when it comes to women), but to Olive Jean Dunn [50]

union bound ensures that FWER $\leq \alpha$. However, this is a very conservative procedure in the sense that it leads to a small number of rejections [68].

In many applications it is not a problem to allow some false discoveries as long as it allows to do much more true discoveries. That is why the control of the FDR can be more interesting since it is less restrictive, as it controls the fraction of false discoveries authorized instead of controlling the probability of making at least one discovery. This is the quantity we will focus on for the problem of outliers detection considered in Chapter 2 and Chapter 3.

A fundamental procedure is the following *Benjamini-Hochberg* procedure [9] that achieves FDR $\leq \alpha$ when the p-values $p_1, \ldots, p_m$ of the tests are independent. Then the algorithm is the following:

Algorithm 2 The Benjamini-Hochberg procedure for FDR control

Input $p_1, \ldots, p_m$ p-values and fixed level α.
Sort the p_i $p_{(1)} \leq p_{(2)} \leq \ldots, \leq p_{(m)}$.
Compute the largest k such that $p_{(k)} \leq \frac{k}{m}\alpha$.
Reject $H_{0,(i)}$ for $i = 1, \ldots, k$.

Note that indepence assumption can be weakened and even arbitrary dependence can be handled [10] (the *Benjamini-Hochberg-Yekutieli* procedure). The *Benjamini-Hochberg* procedure is the most used and has a particularly interesting connection with Slope [11] as explained in Section 3.3 of Chapter I.

2 Outliers in Linear Regression and Robust Estimation

In this section we review some of the tools that are used when dealing with outliers in regression problems. Mainly two approaches coexist: changing the goodness-of-fit to lower the influence of outliers, or adding parameters to take into account the presence of outliers.

2.1 Outliers in Linear Regression

Linear regression is an important tools in statistical data analysis. The model is described as follows:

$$y_i = x_i^\top \beta^\star + \varepsilon_i \tag{I.16}$$

for $i = 1, \dots, n$, where n is the sample size and $\beta^\star \in \mathbb{R}^p$, $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$ and $\varepsilon_i \in \mathbb{R}$ respectively stand for the linear regression coefficients, vector of covariates, label and noise of sample i. The noise is assumed to be independent and identically distributed as $\mathcal{N}(0, \sigma^2)$.

In its basic applications where β has to be estimated from the sample, the Ordinary Least Square (OLS) estimate minimizes the sum of squared residuals:

$$\beta^{OLS} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i^\top \beta)^2. \tag{I.17}$$

In a statistical context, an observation is an outlier if it "deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism" [37].

In a regression context, in particular in linear regression, there can be outliers in the covariates or in the observed vector y. Observations having outliers in the covariates are named *influencial* or *high-leverage* points.

Definition I.8. *Consider the model described by Equation* (I.16) *and let X be the matrix in which the i^{th} row is given by $x_i^\top$ (design matrix). Suppose that X has rank p and let $H = X(X^\top X)^{-1}X^\top$. Then, the leverage of observation i is defined as the i^{th} diagonal element of H and for all $i \in \{1, \dots, n\}$:*

- $0 \leq h_{ii} \leq 1$,
- $\operatorname{var}(y_i - x_i^\top \beta^{OLS}) = (1 - h_{ii})\sigma^2$.

This clearly shows that the higher the leverage, the higher the influence of the observation on the regression. Note that this definition cannot be used when $\operatorname{rank}(X) < p$, in particular in high dimension settings.

In this thesis, we focus on discovering outliers as unusual values of y, in low or high dimension. This problem can be correlated with the former one, in the sense

that observations that are both high leverage points and outliers in y are difficult to detect [22].

In the following of this Section, we review some classical techniques to overcome this issue of outliers in the observation vector y. Two approaches coexist and have different purposes: to get rid of the outliers influence to estimate the regression coefficients, or precisely identify the outliers besides the estimation of the regression coefficients.

2.2 Median of Means

In a general regression framework, (X, Y) is a pair of random variables, where $Y \in \mathbb{R}$ is some label depending of some inputs $X \in \mathbb{R}^p$. Then, one wishes to find a function f among some class $\mathscr{F}$, for which $f(X)$ is a good prediction of Y. In linear regression, f is to be found among the set of linear functions $\mathscr{F} = \{\langle t, \cdot \rangle \mid t \in \mathbb{R}^p\}$. Naturally, the best performance one may hope for is of the risk minimizer in the class, given by

$$t_0 = \operatorname*{argmin}_{t \in \mathbb{R}^p} \mathbb{E}(\langle t, X \rangle - Y)^2. \tag{I.18}$$

Assume that a sample $(X_i, Y_i)_{i=1,\dots,n}$ is given, assumed for now to be i.i.d. with the same distribution as (X, Y). The aim is to approximate t_0 with a small error (accuracy) and with high probability (confidence) using these random data only. The most natural way of choosing an estimate $\hat{t}$ is by Empirical Risk Minimization (ERM) [63], that is, by least squares regression:

$$\hat{t} = \operatorname*{argmin}_{t \in \mathbb{R}^p} \sum_{i=1}^{n} (\langle t, X_i \rangle - Y_i)^2.$$

Assume now that the dataset is made of $n - |\mathscr{O}|$ i.i.d. data $(X_i, Y_i)_{i \in \mathscr{I}}$ with the same distribution as (X, Y), and $|\mathscr{O}|$ outliers $(X_i, Y_i)_{i \in \mathscr{O}}$ that can be arbitrarily distributed, in particular with another distribution than (X, Y). ERM is known to be sensitive to outliers [58]. To overcome this issue, Median of Means (MoM) estimators [58, 57, 60] rely on two building blocks:

1. Estimate the difference of the risks for all pairs $t, u \in \mathbb{R}^p$, namely replace mini-

mization (I.18) by:

$$t_0 = \underset{t\in\mathbb{R}^p}{\operatorname{argmin}} \sup_{u\in\mathbb{R}^p} \mathbb{E}\big[(\langle t, X\rangle - Y)^2 - (\langle u, X\rangle - Y)^2\big].$$

2. Replace the calculation of the mean in ERM by the median of K means computed over a partition of K blocks of the sample $(X_i, Y_i)_{i=1,\dots,n}$.

Additional steps are then added depending on the context and the method [58, 57, 60]. Then with mild assumptions on $(X_i, Y_i)_{i\in\mathcal{O}}$, MoM estimators perform as good as ERM would on the clean subset [58, 60], namely, the calculated MoM estimator $\hat{t}^{MoM}$ verifies:

$$\|\hat{t}^{MoM} - t_0\|_2 \leq C\sigma\sqrt{\frac{p}{n}},$$

with large probability, where $C > 0$ is some numerical constant.

Note that MoM estimation also extents to the high-dimensional setting [57, 58], providing optimal results for the estimation problem. However, MoM estimators do not provide any guarantee about outliers detection, which is the problem we are interested in.

2.3 MM-Estimates

In this section we review some popular robust regression methods, which are methods based on the change of the loss function in the minimization (I.17), allowing to not look too much at the outliers. The first method, called M-estimation, has been proposed by Huber [40] in 1964 and has been much more developed in the 70's and 80's [69, 89, 71].

M-Estimation. The term M-estimation is for Maximum Likelihood type Estimator. A robust M-estimator minimises the sum of a less rapidly increasing objective function than the least squares estimator, thus down-weighting the larger residuals. Instead of minimizing (I.17), the estimator of the regression coefficients is now defined as:

$$\hat{\beta}^M = \underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \ell(y_i, x_i^\top \beta), \tag{I.19}$$

where ℓ is a loss function that can be chosen in multiple ways to reduce the influence of observations with large residuals, such as a loss less rapidly increasing at $-\infty$ and $+\infty$. For example, the Huber loss [40] is defined as:

$$\ell(y,z) = \ell_{\text{Huber}}(y,z) = \begin{cases} \frac{1}{2}(z-y)^2 & \text{if } |z-y| \leq \delta \\ \delta(|z-y| - \frac{1}{2}\delta) & \text{if } |z-y| > \delta \end{cases},$$

and thus allow not to take into account too much observations with large residuals. Figure I.1 below recall some other popular losses, such as Tukey's loss [3].

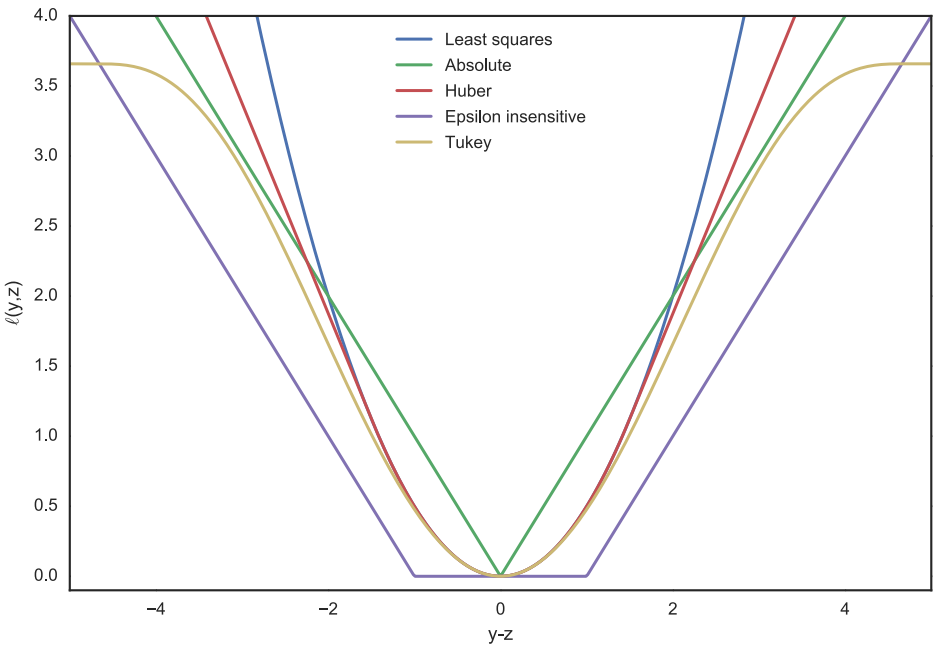

Figure I.1: Plot of losses as functions of the residuals

When $p/n \to 0$ and when observations have low leverages, M-estimators have nice properties such as consistency and asymptotic normality [42] when ℓ is convex and continuous.

Definition I.9. *An estimator* $\hat{e}(y)$ *calculated from the data* (x,y) *is said to be* scale invariant *if*

$$\hat{e}(cy) = c\hat{e}(y),$$

for any positive constant c.

It is interesting to consider estimators satisfying this property since they provide a coherent interpretation of the results. Indeed, if we change the scale of our measurements y by an arbitrary change of units, the selected variables are the same and the prediction changes accordingly.

To achieve scale invariance in M-estimation, it is convenient to use loss function in the form

$$\ell(y,z) = \phi\Big(\frac{y-z}{s}\Big),$$

where s is an estimate of the standard deviation σ of residuals [48]. The standard deviation of the sample of residuals cannot be used for s because it is strongly affected by outliers. Typical choice for s is the Median Absolute Deviation (MAD) scale estimate, adjusted by a factor for asymptotically normal consistency [42]:

$$MAD = \text{median}(|y - X\hat{\beta}|),$$

where $\hat{\beta}$ is an estimate of the regression coefficients and $|y - X\hat{\beta}|$ stands for the vector with coordinates $|y_i - x_i^\top \hat{\beta}|$. As $\hat{\beta}^M$ is not scale invariant, an iterative procedure that gradually converges to an estimate for both errors and scale will be required [3]. This approach is highly resistant to outlying observations as it is based on the median rather that the mean [3].

S-estimation. Rousseeuw and Yohai [71] first proposed these estimators, calling them S-estimators because they are based on estimates of scale.

For any sample of residuals $(y_i - x_i^\top \beta)_{i=1,\dots,n}$, we define the scale estimate $s(y_1 - x_1^\top \beta, \dots, y_n - x_n^\top \beta)$ as a positive solution of

$$\sum_{i=1}^{n} \phi\Big(\frac{y_i - x_i^\top \beta}{s}\Big) = K, \tag{I.20}$$

that is an M-estimate of scale, where ϕ is continuously differentiable, symmetric, increasing on $[0,a]$, constant on $[a,\infty[$ for some $a > 0$, $\phi(0) = 0$, where K is a constant depending on ϕ. Typical choice for ϕ is Tukey bisquare function plotted in Figure I.1, for which appropriate values of K are discussed in [71, 70]. The S-estimator $\hat{\beta}^{\text{scale}}$ is

then defined as a solution of

$$\min_{\beta \in \mathbb{R}^p} s(y_1 - x_1^\top \beta, ..., y_n - x_n^\top \beta), \tag{I.21}$$

As before, this leads to an iterative procedure that requires initial estimates. Compared to M-estimators, S-estimators are useful to handle cases where there are some high-leverage observations in the data [70].

MM-Estimation. MM-estimators are obtained in three stages, combining M-estimation and S-estimation to obtain a robust estimator that has the good properties of each one of these [89]:

1. Compute a S-estimator and the corresponding residuals,
2. Using these residuals, compute an M-estimate of scale following Equation (I.20) with objective function ϕ_0,
3. Compute an M-estimator with the scale obtained at the previous stage and objective function ϕ_1 with $\phi_1 \leq \phi_0$

The widely used *rlm* function from the R MASS package computes such an estimator, using the same Tukey bisquare objective function in both Stages 2 and 3.

Such techniques of robust estimation can be computationally heavy, particularly on large datasets and high dimension. Furthermore it focuses on the estimation problem and does not provide any guarantee about outliers detection, which is the problem we are interested in. What follows in this chapter is a summary of some previously developed techniques that focus on the detection of outliers.

2.4 Mean-shift and variance-shift single outlier model

The two models below, contrary to Robust Regression explained above, include some parameters to model the outliers.

Mean-shift single outlier model. Introduced in the 80's, the mean-shift model [22] has been the first technique considering outliers as the object of interest. It relies on studentized residuals to construct statistical testing to conclude whether or not an

observation is an outlier. It requires the problem to be low-dimensional (small p) as it uses the "hat matrix" $H = X(X^\top X)^{-1}X^\top$ introduced in Definition I.8.

Suppose we want to know if observation i is an outlier. First, one can slightly modify the original model (I.16) to include a new parameter μ_i of interest:

$$y_i = x_i^\top \beta^\star + \mu_i + \varepsilon_i, \tag{I.22}$$

$$y_j = x_j^\top \beta^\star + \varepsilon_j \quad j \neq i, \tag{I.23}$$

where a non-zero μ_i means that observation i is detected as an outlier.

Those new parameters allow to measure the performances of inference procedures in terms of the ability to discover outliers without making too many mistakes, which was not possible with the robust regression procedures considered in the previous paragraph. A natural way to do this is to construct a statistical test with the following hypotheses:

$$H_0 : \mu_i = 0, \qquad H_1 : \mu_i \neq 0. \tag{I.24}$$

Let $r_j, j = 1, \ldots, n$ be the residuals computed by OLS, namely:

$$r_j = y_j - x_j^\top \hat{\beta}^{OLS}.$$

Let $\hat{\sigma}$ be the residual mean square, namely:

$$\hat{\sigma} = \frac{1}{n-p} \sum_{j=1}^{n} r_j^2$$

and $\hat{\sigma}_{(i)}$ the residual mean square computed without the ith observation, that can be computed with the following expression [22]:

$$\hat{\sigma}^2_{(i)} = \frac{(n-p)\hat{\sigma}^2 - r_i^2/(1-h_{ii})}{n-p-1},$$

where h_{ii} is the leverage introduced in Definition I.8. Based on these quantities, the following result holds:

Theorem I.1 ([22]). *Define the quantity:*

$$t_i = \frac{r_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}} = r_i\sqrt{\frac{n-p-1}{n-p-r_i^2}}, \tag{I.25}$$

then under H_0, t_i has a student distribution with parameter $n-p-1$.

Note that this can also be extended to test whether a fixed set of observations is a set of outliers [22].

A main issue is that it only allows to test one or a group of observations. To test if each observation is an outlier, considering n times the Mean-shift single outlier model above ends with n tests of the form of Equation (I.24). It is widely known that performing many tests leads to many false rejections if we do not apply a multiple testing procedure as explained in Section 1.4. However, even with this additional computation, this framework leads to "masking" effects, such that an outlier is undetected because of the presence of other adjacent ones. This phenomenon arises even with low leverage observations [20].

Variance-shift single outlier model. The variance-shift model [64] also introduces a single parameter to include in the model the possibility that one fixed observation is an outlier. The difference with the mean-shift model above is that instead of the mean, this new parameter will influence the variance of the noise, as described below. Fix an observation i that is a possible outlier, then define the model as:

$$y_i = x_i^\top \beta^\star + \alpha_i \sigma_i \tag{I.26}$$

$$y_j = x_j^\top \beta^\star + \varepsilon_j, \quad j \neq i, \tag{I.27}$$

where the notations are the same as in Equation (I.16) except that the standard deviation of observation i is shifted by a factor $\alpha_i \geq 1$. A value $\alpha_i > 1$ would lead to the conclusion that observation i is an outlier. As in Theorem I.1, a test can be constructed and would lead to the same masking issue.

However, both of the above models have also been studied via maximum likelihood, through the problem of detecting the most likely outliers [22, 64]. It has been shown [64] in the mean-shift single outlier model that this is equivalent to select the observation i with maximum absolute value of t_i defined in Theorem I.1, whereas

in the variance-shift single outlier model the small sample distribution property of the Maximum Likelihood Estimator is untractable. In particular, the inference of the two models above are not equivalent in term of MLE, in the sense that it does not necessarily lead to the same conclusion for identifying outliers [64].

The mean-shift and variance-shift single outlier models are less popular than MM-estimation because (i) these models are more efficient to test whether or not a fixed (group of) observation(s) is outlying and (ii) the estimation problem in presence of outliers is historically more popular. However the new developments in the area of convex penalization and sparsity-inducing penalization allow to overcome the main issue of the mean-shift and variance-shift single outlier models, as explained in the next Section.

2.5 A variable selection problem in high-dimensional linear regression ?

Recent developments in convex optimization and sparsity inducing penalization such as Lasso [81] or Slope [12] allow to do inference not only on a single parameter such as in model (2.4) but on a whole vector which is assumed to be sparse.

The Mean-shift outlier model is the model we use for our problem and for which new theory will be developed in Chapter II. It is described as above:

$$y_i = x_i^\top \beta^\star + \mu_i^\star + \varepsilon_i, \tag{I.28}$$

for $i = 1, \dots, n$, where n is the sample size. A non-zero $\mu_i^\star$ means that observation i is an outlier, and $\beta^\star \in \mathbb{R}^p$, $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$ and $\varepsilon_i \in \mathbb{R}$ respectively stand for the linear regression coefficients, vector of covariates, label and noise of sample i. In what follows we assume that the noise is $\mathcal{N}(0, \sigma^2)$ and i.i.d. In matrix notation, this model rewrites as:

$$Y = X\beta^\star + \mu^\star + \varepsilon, \tag{I.29}$$

where $Y = (y_1, \dots, y_n)^\top$, $X \in \mathbb{R}^{n \times p}$ is the feature (or design) matrix for which line i is given by $x_i^\top$ for $i \in \{1, \dots, n\}$, $\mu^\star = (\mu_1^\star, \dots, \mu_n^\star)^\top$ and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top$.

In the linear model (no parameter $\mu^\star$), the model is identifiable if $\operatorname{rank}(X) = p$.

But model (I.29) is clearly an ill-posed problem because the application

$$\begin{pmatrix}\beta\\ \mu\end{pmatrix} \mapsto X\beta + \mu$$

is not injective without assumptions on β, μ. A discussion about the identifiability of high-dimensional linear model is given in [72]. Therefore we impose sparsity on the vector of individual intercepts $\mu^\star$. Note that if the features are in high dimension, $\beta^\star$ can also be assumed sparse.

Interestingly, Model (I.29) can also be written as a $(n+p)-$dimensional linear regression problem. Define $Z = [X \quad I_n] \in \mathbb{R}^{n\times(n+p)}$ as the concatenation of X and the $n-$dimensional identity matrix I_n. Define $\gamma^\star = ({\beta^\star}^\top, {\mu^\star}^\top)^\top \in \mathbb{R}^{n+p}$, then the model has the following concatenated form:

$$Y = Z\gamma^\star + \varepsilon. \tag{I.30}$$

In the following we assume that all design matrices $X \in \mathbb{R}^{n\times p}$ verify $\|X_i\| = 1, i = 1,\dots,p$, which is a common normalization assumption [76, 44].

False Discovery Rate. The False Discovery Rate of Equation (I.15) is a popular measure of performance when one wants to encourage discoveries (rejected hypotheses) without making too many mistakes [8]. In the context of parametric estimation, there is a natural way to compute FDR to measure the performance of support recovery of a true parameter θ with an estimator $\hat{\theta}$:

Definition I.10. *Given an estimator $\hat{\theta}$ of a true parameter θ, we define the False Discovery Rate for the support recovery of θ as*

$$\mathrm{FDR}(\hat{\theta};\theta) = \mathbb{E}\left[\frac{\#\{i \in \{1,\dots,n\}, \hat{\theta}_i \neq 0 \text{ and } \theta_i = 0\}}{\#\{i \in \{1,\dots,n\}, \hat{\theta}_i \neq 0\}}\right]. \tag{I.31}$$

This definition matches the one of Equation (I.15) when considering $H_{0,i} = \hat{\theta}_i = 0$ for all $i \in \{1,\dots,n\}$. Another distinct measure of performance of an estimator, which also comes from multiple testing, is the ability of the estimator not to miss too many discoveries.

Definition I.11. *The True Positive Rate of an estimator $\hat{\theta}$ of a true parameter θ is given by*

$$\mathrm{TPR}(\hat{\theta};\theta) = \mathbb{E}\left[\frac{\#\{i \in \{1,\dots,n\}, \hat{\theta}_i \neq 0 \text{ and } \theta_i \neq 0\}}{\#\{i \in \{1,\dots,n\}, \theta_i \neq 0\}}\right]. \tag{I.32}$$

There is a balance between FDR and TPR, as it is difficult to achieve both low FDR and high TPR. Indeed, a FDR equals to zero suggests that only few discoveries are made and therefore TPR may be low too,. At the opposite if TPR is 1 then a lot of discoveries are made, possibly leading to more false discoveries. Yet, in many applications the priority is to maintain the FDR below a certain level (typically 5%).

Recently, Slope has been used in the context of high-dimensional linear regression [12, 76] and has demonstrated interesting properties regarding FDR control in variable selection that we recall in the next Section. This motivated our use of this penalization in the context of outliers detection as explained in Chapter II.

In the next Section, we recall important results about estimation and variable selection in high-dimensional linear regression, and discuss the relevance of the concatenated model (I.30) above for our problem.

3 Convex penalization in high-dimensional linear regression: estimation, variable selection

Equation (I.30) is seducing because it is a high-dimensional regression model, which is well-known and has been well studied in the past few years in terms of estimation bound [44, 66, 14, 76] and support recovery [49, 44, 12, 76]. Established results in literature are presented in Section 3.2 and Section 3.3 below. They require hypotheses on the design matrix X, we present now the most classical ones in the following Section.

3.1 Classical hypotheses on the design matrix

The first definition is popular in compress sensing [47] as it allows to prove that basis pursuit [16] allows exact recovery of a sparse signal β from "compressed measurements" $y = X\beta$ [45].

Definition I.12 (Restricted Isometry Property, [47]). *A matrix $A \in \mathbb{R}^{n\times p}$ satisfies RIP of order s if there exists $\delta \in [0,1[$ such that for all $x \in \mathbb{R}^p$ with $|x|_0 \leq s$:*

$$(1-\delta)\left\|x\right\|_2^2 \leq \left\|Ax\right\|_2^2 \leq (1+\delta)\left\|x\right\|_2^2. \tag{I.33}$$

The RIP condition above is one of the strongest in the literature about compressed sensing, and implies various weaker assumptions (see [27] for a whole review and comparison of required assumptions to derive estimation bounds in high-dimensional linear model). Two other definitions that lead to weaker assumptions are given below:

Definition I.13 (Strong Restricted Eigenvalue condition, [14]). *A matrix $A \in \mathbb{R}^{n\times p}$ satisfies $SRE(s, c_0)$ if there exists $\kappa > 0$ such that:*

$$\min_{\substack{x\in\mathscr{C}_{SRE}(s,c_0)\\ x\neq 0}} \frac{\|Ax\|_2}{\|x\|_2} \geq \kappa, \tag{I.34}$$

where

$$C_{SRE}(s,c_0) = \left\{\beta \in \mathbb{R}^p \mid \|\beta\|_1 \leq (1+c_0)\sqrt{s}\|\beta\|_2\right\}. \tag{I.35}$$

Note that this implies the left inequality of Equation (I.33). Different versions of RE conditions exist [43], differing in the subset on which the minimum is taken in Equation (I.34). The version above is strong in the sense that the subset can be less stringent. In the same spirit, [14] defines a *Weighted Restricted Eigenvalue* (WRE) condition as:

Definition I.14 ([14]). *Let $c_0 > 0$, $s \in \{1,\dots,p\}$ and $\lambda = (\lambda_1,\dots,\lambda_p)$ a non-increasing sequence of positive numbers. Then a matrix $A \in \mathbb{R}^{n\times p}$ satisfies the $WRE(s, c_0)$ condition if:*

$$\kappa(s,c_0) := \min_{\substack{x\in\mathscr{C}_{WRE}(s,c_0)\\ x\neq 0}} \frac{\|Ax\|_2}{\|x\|_2} > 0, \tag{I.36}$$

where

$$C_{WRE}(s,c_0) = \left\{\beta \in \mathbb{R}^p, J_\lambda(\beta) \leq (1+c_0)\|\beta\|_2\sqrt{\sum_{i=1}^{s}\lambda_i^2}\right\}, \tag{I.37}$$

with J_λ the Slope norm introduced in Definition I.4.

However, all these RE-types conditions are roughly equivalent [14]. A wide class of random matrices satisfies Equation (I.34) and Equation (I.36), and such assump-

tions are known to be mandatory in order to derive fast rates of convergence for penalizations based on the convex-relaxation principle [92].

A last definition that is used in literature is the notion of coherence, that measures the correlation among the columns of a matrix.

Definition I.15 (Coherence Property, [44]). *A matrix A with unit-normed columns is said to satisfy the Coherence Property if:*

$$\sup_{1\leq i<j\leq p} |\langle A_i, A_j\rangle| \leq \frac{A_0}{\log p}, \tag{I.38}$$

where A_0 is some positive numerical constant.

Below we recall some estimation and variable selection results where hypotheses on the design matrix are linked to the definitions above. Results of Section 3.2 are obtained with the ℓ_1 norm as sparsity-inducing penalization (Lasso [81]) and results of Section 3.3 are obtained with Slope of Definition I.4. Results are presented in two different sections because they are really different in nature, particularly on the question of variable selection.

3.2 Results for Lasso

Let $X \in \mathbb{R}^{n\times p}$, $\beta \in \mathbb{R}^p$, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ and

$$Y = X\beta + \varepsilon. \tag{I.39}$$

Assume that β is k-sparse, that is $|\beta|_0 \leq k$. Define $\hat{\beta}$ as a solution of the following minimization problem for some $\lambda > 0$:

$$\min_{\beta\in\mathbb{R}^p} \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1. \tag{I.40}$$

Three error measures can be interesting to evaluate: the *prediction error* $\|X\beta - X\hat{\beta}\|_2$, the *estimation error* $\|\beta - \hat{\beta}\|_2$ and eventually, in terms of variable selection the support recovery of β that can be measured with the FDR and the TPR introduced in Definition I.10 and Definition I.11 of Section 2.5.

The following result is about the prediction error rate. Note that this type of inequality appears several times in literature [66, 43, 14] with variations of design hypothesis, probability and constants.

Theorem I.2 ([44]). *Suppose that X obeys the coherence property. Suppose that $k \leq c_0 p/(\|X\|^2 \log p)$ for some positive numerical constant c_0 and where $\|X\|$ denotes the largest singular value of X. Then the Lasso estimate* (I.40) *computed with $\lambda = 2\sigma\sqrt{2\log p}$ satisfies*

$$\|X\beta - X\hat{\beta}\|_2^2 \leq 2C\sigma^2 k \log p \tag{I.41}$$

with probability at least $1 - 6p^{-2\log 2} - p^{-1}(2\pi \log p)^{-1/2}$. The constant C may be taken as $8(1+\sqrt{2})^2$.

The second result shows that under additional assumptions, the lasso estimate can recover the support of the true regression parameter:

Theorem I.3 ([44]). *Let I be the support of β and suppose that:*

$$\min_{i \in I} |\beta_i| > 8\sigma\sqrt{2\log p},$$

then under the assumptions of Theorem I.2 the lasso estimate obeys:

$$\operatorname{sgn}(\hat{\beta}) = \operatorname{sgn}(\beta),$$

with probability at least $1 - 2p^{-1} - ((2\pi \log p)^{-1/2} + |I|p^{-1}) - O(p^{-2\log 2})$, with sgn *defined in Definition .3 of Section 2.*

This particularly implies that $\mathrm{FDR}(\hat{\beta};\beta) = 0$ and $\mathrm{TPR}(\hat{\beta},\beta) = 1$, namely perfect recovery.

To obtain support recovery the Coherence Property roughly is the only condition that appears in literature, along with a closely related condition named *Irrepresentable condition* [49], which is weaker [27] but more of theoretical than practical interest. Indeed, the only examples of matrices found in literature that verify the *Irrepresentable condition* are random matrices that do not have a high correlation [93], which are examples that also match the Coherence Property.

The last result presented here is about the prediction and estimation errors of the regression parameter. Once again this type of result appears widely in literature and

the transition from prediction error to estimation error always requires some *RE*-type condition on the design matrix.

Theorem I.4 ([14]). *Suppose that X obeys $SRE(k,7)$. Let $\hat{\beta}$ be the Lasso estimator with tuning parameter λ satisfying $\lambda \geq 2(4+\sqrt{2})\sigma\sqrt{\log(2ep/k)}$. Then, we have:*

$$\mathbb{P}\left(\|X\beta - X\hat{\beta}\|_2^2 \leq \frac{49k\lambda^2}{16\kappa^2(k,7)}\right) \geq 1 - \frac{1}{2}\left(\frac{k}{2ep}\right)^{k/\kappa^2(k,7)}, \tag{I.42}$$

and

$$\mathbb{P}\left(\|\beta - \hat{\beta}\|_2 \leq \frac{49\sqrt{k}\lambda}{8\kappa^2(k,7)}\right) \geq 1 - \frac{1}{2}\left(\frac{k}{2ep}\right)^{k/\kappa^2(k,7)}. \tag{I.43}$$

Note that the theorem above can lead to a better rate than Theorem I.2 ($k\log(p/k)$ instead of $k\log p$) but requires the sparsity k to be known since it is used in the tuning parameter λ. If k is not known, it is still possible to compute an aggregated Lasso estimator that adapt to the sparsity level and provides the rate $k\log(p/k)$ without the knowledge of k [14]. This is the minimax optimal rate of convergence for high-dimensional regression [66].

The next section focuses on Slope in the same context. Slope demonstrates the very attractive property that contrary to the Lasso it automatically adapts to the sparsity level. Moreover, while the Lasso theoretically provides a perfect support recovery (Theorem I.3), in practice it can demonstrate a low TPR and a tendency to shrink large coefficients too much yet insufficiently shrink small coefficients by applying the same penalty to every regression coefficient [94]. Slope allows FDR control for specific design matrix, which is also very interesting because focusing on FDR allows a more liberal procedure and a higher TPR. This is illustrated for example in [12].

3.3 Results for Slope

In this Section we focus on the same minimization as Equation (I.40) replacing the Lasso penalty by a Slope penalty:

$$\min_{\beta\in\mathbb{R}^p} \frac{1}{2}\|Y - X\beta\|_2^2 + J_\lambda(\beta), \tag{I.44}$$

where J_λ is given by Equation (I.9).

The first results have been obtained for orthogonal matrix design, namely $X^\top X = I_p$ [12]. Even if this design is very specific, it is very interesting because it closely links Slope with the Benjamini-Hochberg procedure and provides FDR control for the support recovery of β.

Orthogonal design and connection with the Benjamini-Hochberg procedure. When X is orthogonal, with $\tilde{Y} = X^\top Y$, the model given by Equation (I.39) reduces to:

$$\tilde{Y} = \beta + \tilde{\varepsilon} \tag{I.45}$$

after multiplication by $X^\top$, with $\tilde{\varepsilon} = X^\top \varepsilon$ distributed as $\mathcal{N}(0, \sigma^2 I_p)$.

In this context, a statistical testing for nullity of each regression coefficient is given by:

$$H_{0,i} : \beta_i = 0 \qquad H_{1,i} : \beta_i \neq 0, \quad i = 1, \dots, n.$$

Then under $H_{0,i}$, $\tilde{y}_i$ is distributed as $\mathcal{N}(0, \sigma^2)$. Assuming that σ is known, the test statistic is simply $\tilde{y}_i / \sigma$. Therefore according to Proposition I.6 of Section 1.4, Chapter I, the p-value of the i-th test is given by $p_i = 2(1 - \Phi(y_i/\sigma))$ and are independent, where Φ is the cumulative distribution function of $\mathcal{N}(0,1)$.

Applying now the Benjamini-Hochberg procedure (B-H) described in Algorithm 2 (Section 1.4, Chapter I) with a fixed level α leads to compare each $p_{[i]}$ with $\alpha i/n$ namely looking for the largest k such that:

$$y_{(k)} \leq \sigma \Phi^{-1}\big(1 - \frac{k\alpha}{2n}\big),$$

where $y_{(1)} \geq y_{(2)} \geq \cdots \geq y_{(n)}$, noting that the largest the value of y, the smallest the p-value. Therefore the Benjamini-Hochberg procedure compares each $y_{(i)}$ with

$$\lambda_i^{BH}(\alpha) := \sigma \Phi^{-1}\big(1 - \frac{i\alpha}{2n}\big).$$

Note that Slope works with the same idea of penalizing more the largest coefficients, that would correspond to the highest p-values. Another remarkable property that links B-H procedure and Slope is that with an orthogonal design, Slope applied with the weights $\lambda^{BH}(\alpha)$ leads to a FDR lower than α for the support estimation of β, as recalled in the following theorem:.

Theorem I.5 ([12]). *Let X be orthogonal. Fix $0 < \alpha < 1$ and let $\hat{\beta}^{\text{slope}}$ be a solution of Equation* (I.44) *with $\lambda = \lambda^{BH}(\alpha)$. Then:*

$$\text{FDR}(\hat{\beta}^{\text{slope}}) \leq \frac{\alpha k}{p}. \tag{I.46}$$

As mentioned in previous sections, we are also interested in the estimation and prediction errors, that are the same in the particular case of an orthogonal design. The following result holds:

Theorem I.6 ([76]). *Let X be orthogonal and assume that $p \to \infty$ with $k/p \to 0$. Fix $0 < q < 1$. Let $\hat{\beta}^{\text{slope}}$ be a solution of Equation* (I.44) *with $\lambda = \lambda^{BH}(\alpha)$. Then:*

$$\sup_{\|\beta\|_0 \leq k} \mathbb{E}\big[\|\hat{\beta}^{\text{slope}} - \beta\|_2^2\big] = (1 + o(1))2\sigma^2 k \log(p/k). \tag{I.47}$$

In Section 3.4 it is shown that this result is optimal in some sense, demonstrating once again the interesting properties of Slope.

However, an orthogonal design is very particular and stringent. The next two paragraphs contain results for a more general design, beginning with a design that is not too far from being orthogonal.

Near-Orthogonal design. The near-orthogonal situation is described by a Gaussian independent random design. Namely, we suppose that entries of $X \in \mathbb{R}^{n \times p}$ are independent and distributed as a $\mathcal{N}(0, 1/\sqrt{n})$. This is called a near-orthogonal situation since such tall Gaussian matrices verify the Restricted Isometry Property condition with the smallest known upper bound for random matrices [7].

FDR control is valid in orthogonal design, a natural question is to wonder if this still holds in the near-orthogonal design defined above. The answer is yes, partially, in the sense that the following asymptotic result holds:

Theorem I.7 ([76]). *Fix $0 < \alpha < 1$ and let $\hat{\beta}^{\text{slope}}$ be solution of Equation* (I.44) *with $\lambda = (1+\epsilon)\lambda_{BH}(\alpha)$ for some arbitrary constant $0 < \epsilon < 1$. Suppose $k/p \to 0$ and $(k \log p)/n \to 0$. Then:*

$$\text{FDR}(\hat{\beta}^{\text{slope}}) \leq (1 + o(1))\alpha. \tag{I.48}$$

Additionally, if the non-zero coefficients of the true parameter β have absolute values greater than $(1+\epsilon)\lambda_1^{BH}(\alpha)$, then:

$$\mathrm{TPR}(\hat{\beta}^{\mathrm{slope}}) \to 1. \tag{I.49}$$

Note that this theorem nicely encourages the use of SLOPE in the variable selection context. Once again, as variable selection is not the only interest, estimation and prediction errors are recalled in the following theorem.

Theorem I.8 ([76]). *Under the same assumptions as Theorem I.7 the following results hold:*

$$\sup_{\|\beta\|_0 \le k} \mathbb{P}\left(\frac{\|\hat{\beta}^{\mathrm{slope}} - \beta\|_2^2}{2\sigma^2 k \log(p/k)} > 1 + 3\epsilon \right) \to 0 \tag{I.50}$$

and

$$\sup_{\|\beta\|_0 \le k} \mathbb{P}\left(\frac{\|X\hat{\beta}^{\mathrm{slope}} - X\beta\|_2^2}{2\sigma^2 k \log(p/k)} > 1 + 3\epsilon \right) \to 0 \tag{I.51}$$

Note that the bounds on the prediction and estimation errors are the same, which is not surprising since X is nearly an isometry.

The bound (I.51) is sharp in the sense that no other estimator can do better in this framework, as shown by the following theorem:

Theorem I.9 ([76]). *Under the same assumptions as Theorem I.8, for any $\epsilon > 0$, we have:*

$$\inf_{\hat{\beta}} \sup_{\|\beta\|_0 \le k} \mathbb{P}\left(\frac{\|\hat{\beta} - \beta\|_2^2}{2\sigma^2 k \log(p/k)} > 1 - \epsilon \right) \to 0, \tag{I.52}$$

where the infimum is taken over all possible estimators.

Optimality in a more general design is discussed in Section 3.4 below.

General design. The previous paragraphs give estimation and model selection results for a very particular design. Though model selection should be more complicated for a general design, estimation results are more classical and generally obtained with some *RE* condition (see Definition I.13). Hence, the following estimation and prediction results holds:

Theorem I.10 ([14]). *Assume that $WRE(k,7)$ holds. Let $\hat{\beta}^{\text{slope}}$ be solution of Equation* (I.44) *with $\lambda_j = A\sigma\sqrt{\log(2p/j)}$, $j = 1,\dots,p$ with $A \geq 2(4/\sqrt{2})$. Then*

$$\mathbb{P}\left(\|\hat{\beta}^{\text{slope}} - \beta^\star\|_2 \leq \frac{9\sum_{i=1}^s \lambda_j^2}{4\kappa^4(k,3)}\right) \geq 1 - \frac{1}{2}\left(\frac{k}{2p}\right)^{k/\kappa^2(k,3)} \tag{I.53}$$

and

$$\mathbb{P}\left(\|X\hat{\beta}^{\text{slope}} - X\beta^\star\|_2 \leq \frac{49\sum_{i=1}^s \lambda_j^2}{16\kappa^2(k,7)}\right) \geq 1 - \frac{1}{2}\left(\frac{k}{2p}\right)^{k/\kappa^2(k,7)}. \tag{I.54}$$

This is actually a simplified version of the theorem presented in [14], which is slightly more general and also provides bounds in expectation instead of bounds with high probability. Note that the weights used in Theorem I.10 above are not very different from λ^{BH} since

$$\lambda_j^{BH}(\alpha) := \sigma\Phi^{-1}(1 - j\alpha/2p) = \sigma(1 + o(1))\sqrt{2\log(2p/j\alpha)},$$

see [14]. Interestingly, authors in [14] provide links between the WRE condition and other RE-type conditions, showing in particular that a wide class of random matrices satisfy those conditions, such as matrices with i.i.d. sub-Gaussian rows.

3.4 Summary table and discussion

A natural question about the prediction and estimation errors of the previous Section is the optimality of such rates. A very specific case is already handled in Theorem I.9. The following theorems, that require the right-hand side of the RIP condition of Definition I.33 (Section 3.1), provide the optimal rate in more general cases. The global situation is then summarized in Table I.1 below.

Theorem I.11 ([66]). *Assume that X satisfies the right-hand side inequality in Equation* (I.33) *of order $2k$, then:*

$$\min_{\hat{\beta}} \max_{|\beta|_0 \leq k} \mathbb{E}\left[\|\hat{\beta} - \beta\|_2^2\right] \geq \frac{c}{1+\delta} k\sigma^2 \log(p/k), \tag{I.55}$$

for some positive constant c, where the minimum is taken over all possible estimators.

Theorem I.12 ([66]). *Assume that X satisfies the right-hand side inequality in Equation* (I.33) *of order $2k$ with constant δ_1, and the left-hand side inequality in Equation* (I.33) *of order $2k$ with constant δ_2, then:*

$$\min_{\hat{\beta}} \max_{|\beta|_0 \leq k} \mathbb{E}\left[\|X\hat{\beta} - X\beta\|_2^2\right] \geq \frac{c(1+\delta_2}{1+\delta_1} k\sigma^2 \log(p/k), \tag{I.56}$$

for some positive constant c, where the minimum is taken over all possible estimators.

Penalization	Estimation/Prediction rate	Hypotheses on X
Lasso	$k \log p$	RE-type or Coherence property (prediction)
Slope	$k \log(p/k)$	RE-type
Optimal rate	$k \log(p/k)$	RIP-like (only right part for estimation)

Table I.1: Convergence rates, up to constants, associated to Lasso and Slope penalizations, together with hypotheses required on the design. The optimal rates are given on the last row.

In particular we see that contrary to the Lasso, Slope achieves the optimal estimation rate if X satisfies RIP.

Now, recall that we are interested in Slope and Lasso properties in the high-dimensional linear regression problem because our problem can be written in such a form (see model of Equation (I.30)). Therefore, we want to know if a matrix that has the concatenated shape $Z = [X\ I]$ verifies one of the classical hypotheses on the design matrix exposed in Section 3.1.

Clearly, the identity part of Z does not allow to use neither the orthogonal situation nor the near-orthogonal situation results of Slope. However, the theorem below highlights the fact that Z satisfies RIP under some (strong) condition:

Theorem I.13 ([62]). *Assume that X has independent and identically distributed $\mathcal{N}(0, 1/n)$ entries. Then, if $n > c_1(k+s)\log((n+p)/(k+s))$, $Z = [X\ I]$ satisfies RIP of order $k+s$ with probability exceeding $1 - 3e^{-c_2 n}$, where c_1 and c_2 are constants that depend only on the desired RIP constant δ.*

To apply the results of high-dimensional linear model to our problem, we need weak correlation between covariates (Theorem I.13), and high magnitude of the coefficients (Theorem I.3). However, this is not satisfying since we would like to identify

outliers even in the presence of high correlation in X and low magnitude in $\mu^\star$. Roughly speaking, the maximum absolute value of n i.i.d. $\mathcal{N}(0,\sigma^2)$ is of order $\sigma\sqrt{2\log n}$, so that we want to be able to identify outliers of magnitude $\sigma\sqrt{2\log n}$.

Moreover, no results about the variable selection properties of Slope are directly applicable to our problem of outlier detection (support recovery). However, results of Section 3.3 show the great interest of Slope, both for estimation and variable selection, therefore we should find a way to use this penalty. Interestingly, FDR control has never been investigated theoretically in the context of outliers detection, while it seems particularly relevant because it is widely known that Lasso can show a lack of power in the context of variable selection [94]. This is precisely the aim of Chapter II.

Moreover, it makes sense to penalize in different ways β and μ since we might not want to penalize β, for example in a low-dimensional setting. Recent approaches follow this spirit and are presented in the next Section.

4 Convex penalization in the Mean-shift outlier model: recent approaches

In this Section we go back to the mean-shift outlier model of Equation (I.29) which writes, in matrix notation,

$$Y = X\beta^\star + \mu^\star + \varepsilon, \tag{I.57}$$

with one regression parameter $\beta^\star$ and one vector of individual intercepts $\mu^\star$.

4.1 The "two penalizations" approach

In light of the previous section, a "two penalizations" approach is adopted, that is one penalization is applied on the regression coefficients and another penalization is applied on the vector of individual intercepts. Note that $\mu^\star$ has to be sparse but $\beta^\star$ does not necessarily has to, so the setting considered includes the case of no penalization on β. Then, the minimization problem is the following:

$$\min_{\beta,\mu} \frac{1}{2}\sum_{i=1}^{n}(y_i - x_i^\top\beta + \mu_i)^2 + \lambda_1 J_1(\beta) + \lambda_2 J_2(\mu). \tag{I.58}$$

A general iterative algorithm which alternates between estimation of β and estimation of μ is proposed in [90] and runs as follow:

Algorithm 3 Alternated minimizations

initialize β, μ
while not converged **do**
$\beta \leftarrow \text{argmin}_\beta \frac{1}{2}\sum_{i=1}^n (y_i - x_i^\top \beta + \mu_i)^2 + \lambda_1 J_1(\beta)$
$\mu \leftarrow \text{argmin}_\mu \frac{1}{2}\sum_{i=1}^n (y_i - x_i^\top \beta + \mu_i)^2 + \lambda_2 J_2(\mu)$.
end while
return β, μ

However when penalizations allow it, proximal methods are more efficient. The example of Equation (I.58) with $\lambda_1 = 0$ and $J_2 = \|\cdot\|_1$ is given as an example in [90] but no theoretical investigation is proposed.

4.2 Recent approaches

Recent approaches rely on the mean-shift outlier model with minimization given by Equation (I.58) with lasso penalization on μ [32, 73, 90]. Note that sometimes a lasso penalization is also applied on β [32, 90]. Thus, the minimization problem is the following:

$$\min_{\beta,\mu} \frac{1}{2}\sum_{i=1}^n (y_i - x_i^\top \beta + \mu_i)^2 + \lambda_1 \sum_{i=1}^p |\beta_i| + \lambda_2 \sum_{i=1}^n |\mu_i|. \tag{I.59}$$

In the low-dimensional setting ($\lambda_1 = 0$), the solution of this minimization is studied in [73] and called soft-IPOD (Iterative Procedure for Outlier Detection). In [73], no theoretical guarantee is demonstrated but some interesting properties are shown. Assuming that $n > p$ and that X has full rank p, authors first show that Equation I.57 is reduced to a new linear regression model with $n - p$ observation, and μ as the regression coefficients, allowing to rely on a Bayesian Information Criterion to select the tuning parameter λ_2. Indeed, let $H = UDU^\top$ be a spectral decomposition of the orthogonal projection $H = X(X^\top X)^{-1}X^\top$, where D is a diagonal matrix and U an orthogonal matrix. Let U_c be the columns of U corresponding to a basis of $\ker H$. Then $U_c \in \mathcal{M}_{n,n-p}$ because X, D and H have rank p. Since H is an orthogonal projection on $\Im X$, the columns of U_c are orthogonal to $\Im X$ and multiplying by $U_c^\top$

the relation of Equation I.57

$$Y = X\beta^\star + \mu^\star + \varepsilon$$

gives

$$Y' = U_c^\top \mu^\star + \varepsilon',$$

with $Y' = U_c^\top Y \in \mathbb{R}^{n-p}$ and $\varepsilon' \sim \mathcal{N}(0, I_{n-p})$. Then, authors show that their procedure is equivalent to a M-estimation for estimation of the regression coefficients and therefore is not appropriate for high-leverage outliers. To overcome this difficulty, they develop hard-IPOD, based on hard-thresholding instead of soft-thresholding, making it a non-convex problem. The algorithm proposed is in the same spirit as [90], alternatively estimating β and μ, however a simplified version is proposed, that allows to compute the estimate of β only once at the end of the procedure.

Note that the two main drawbacks of Soft and Hard-IPOD is that it is limited to low-dimensional settings and that the non-convex method shows no theoretical guarantee and rely on a careful choice of an initial estimate. It should be noticed that a high-dimensional version of IPOD is proposed in [73] but it is done through the concatenated model of Equation (I.30), that we want to avoid for reasons already exposed in Section 3.4, and the choice of the tuning parameter is ambiguous.

The only work in literature that also deals with high dimension and provides theoretical results is the *Robust Lasso* [32]. Consider the model of Equation (I.29) with T and S being the respective support of $\beta^\star$ and $s^\star$ and consider minimizing Equation (I.59) and define $\lambda = \lambda_2/\lambda_1$. An hypothesis similar to RE condition of Definition I.13 is introduced, called *Extended RE*:

Definition I.16. *Define the extended RE cone as:*

$$\mathscr{C} = \left\{(\beta, \mu) \in \mathbb{R}^p \times \mathbb{R}^n \mid \|\beta_{T^c}\|_1 + \lambda\|\mu_{S^c}\|_1 \leq 3\|\beta_T\|_1 + 3\lambda\|\mu_S\|_1\right\}. \tag{I.60}$$

We say that $X \in \mathbb{R}^{n\times p}$ satisfies the extended RE condition if there exists $\kappa > 0$ such that for all $(\beta, \mu) \in \mathscr{C}$:

$$\|X\beta + \mu\|_2 \geq \kappa(\|\beta\|_2 + \|\mu\|_2). \tag{I.61}$$

Under this hypothesis, two results are given. The first result is about estimation rate, while the second result shows guarantee for the support recovery of both $\beta^\star$ and $\mu^\star$ under additional assumptions.

Theorem I.14 ([32]). *Assume that X satisfies the Extended RE. Define $\hat{\beta}, \hat{\mu}$ as a solution of* (I.59) *with $\lambda_1 = 4\sigma\sqrt{\log p}$ and $\lambda_2 = 4\sigma\sqrt{\log n}$. Then we have:*

$$\|\hat{\beta} - \beta^\star\|_2 + \|\hat{\mu} - \mu^\star\|_2 \leq \frac{12\sigma}{\kappa^2}(\sqrt{k\log p} + \sqrt{s\log n}). \tag{I.62}$$

Note that a similar result holds for the prediction error, which leads to the same convergence rate.

Assume now that the rows of X are i.i.d. and distributed as $\mathcal{N}(0, \Sigma)$. Then under an *irrepresentable condition* [32] on Σ and if the magnitude of outliers is large enough, it is shown in [32] that we can choose λ_1 and λ_2 in (I.59) such that a solution $(\hat{\beta}, \hat{\mu})$ recovers the signed support of $\beta^\star$ and $\mu^\star$, namely:

$$\operatorname{sgn}(\hat{\beta}) = \operatorname{sgn}(\beta^\star), \quad \operatorname{sgn}(\hat{\mu}) = \operatorname{sgn}(\mu^\star).$$

The influence of the magnitude will be investigated in Chapter II when comparing Robust Lasso to our procedure. We will show that a high magnitude is necessary for Robust Lasso to be efficient, whereas our procedure will be efficient even with outliers of low magnitudes.

Note that both [32, 73] rely on a Lasso penalization, but the penalization are applied differently: in [32], the procedure named Extended-LASSO uses two different ℓ_1 penalties for β and μ, with tuning parameters that are fixed according to theoretical results, while the (soft-)IPOD procedure from [73] applies the same penalization to both vectors, with a regularization parameter tuned with a modified BIC criterion.

4.3 Tuning parameters

The choice of the tuning parameters λ_1 and λ_2 is crucial in model (I.59) since a large parameter shrinks to zero many parameters, while a small parameter would leads to a lack of power. It is known that the theoretical tunings proposed in Theorem I.3 or Theorem I.14 are not necessarily interesting in practice, being too conservative [73]. This is why the criterion developed for IPOD [73] is interesting.

A popular way of choosing tuning parameters is to do cross validation [51]. Basically, cross validation uses data splitting and works in few steps: first, define a grid of tuning parameter values, then for each value of the grid, use half of the data to do inference on the parameters of the model (here $\beta^\star$ and $\mu^\star$), and use the other half to

compute a score (typically the negative log- likelihood) that reflects the quality of this fit.

However, cross-validation is applied to an independent test sample from the joint distribution of X and Y [51]. Therefore cross-validation cannot be applied for model (I.29) since the parameter $\mu^\star$ induces non-stationarity in the data.

This is a new argument in favour of the Slope penalization, since the theoretical weights of Slope are the one used in practice [12]. In Chapter II, we explore the use of Slope as a penalization in the Equation (I.58) and compare it to the previous methods exposed in this Chapter.

CHAPTER II

SLOPE for Outliers Detection and Robust Estimation in Linear Model

This Chapter is composed of the submitted article *High-dimensional robust regression and outliers detection with SLOPE* available on arXiv:1712.02640.

Abstract

The problems of outliers detection and robust regression in a high-dimensional setting are fundamental in statistics, and have numerous applications. Following a recent set of works providing methods for simultaneous robust regression and outliers detection, we consider in this paper a model of linear regression with individual intercepts, in a high-dimensional setting. We introduce a new procedure for simultaneous estimation of the linear regression coefficients and intercepts, using two dedicated sorted-ℓ_1 penalizations, also called SLOPE [11]. We develop a complete theory for this problem: first, we provide sharp upper bounds on the statistical estimation error of both the vector of individual intercepts and regression coefficients. Second, we give an asymptotic control on the False Discovery Rate (FDR) and statistical power for support selection of the individual intercepts. As a consequence, this paper is the first to introduce a procedure with guaranteed FDR and statistical power control for outliers detection under the mean-shift model. Numerical illustrations, with a comparison to recent alternative approaches, are provided on both simulated and several real-world datasets. Experiments are conducted using an open-source software written in Python and C++.

1 Introduction

Outliers are a fundamental problem in statistical data analysis. Roughly speaking, an outlier is an observation point that differs from the data's "global picture" [36]. A rule of thumb is that a typical dataset may contain between 1% and 10% of outliers [35], or much more than that in specific applications such as web data, because of the inherent complex nature and highly uncertain pattern of users' web browsing [31]. This outliers problem was already considered in the early 50's [23, 30] and it motivated in the 70's the development of a new field called robust statistics [41, 42].

In this paper, we consider the problem of linear regression in the presence of outliers. In this setting, classical estimators, such as the least-squares, are known to fail [41]. In order to conduct regression analysis in the presence of outliers, roughly two approaches are well-known. The first is based on detection and removal of the outliers to fit least-squares on "clean" data [87]. Popular methods rely on leave-one-out methods (sometimes called case-deletion), first described in [22] with the use of residuals in linear regression. The main issue about these methods is that they are theoretically well-designed for the situations where only one given observation is an outlier. Repeating the process across all locations can lead to well-known masking and swamping effects [34]. An interesting recent method that does not rely on a leave-one-out technique is the so-called IPOD [73], a penalized least squares method with the choice of tuning parameter relying on a BIC criterion. A second approach is based on robust regression, that considers loss functions that are less sensitive to outliers [42]. This relies on the M-estimation framework, that leads to good estimators of regression coefficients in the presence of outliers, thanks to the introduction of robust losses replacing the least-squares. However, the computation of M-estimates is substantially more involving than that of the least-squares estimates, which to some extend counter-balances the apparent computational gain over previous methods. Moreover, robust regression focuses only on the estimation of the regression coefficients, and does not allow directly to localize the outliers, see also for instance [91] for a recent review.

Alternative approaches have been proposed to perform simultaneously outliers detection and robust regression. Such methods involve median of squares [74], S-estimation [69] and more recently robust weighted least-squares [28], among many others, see also [33] for a recent review on such methods. The development of robust

methods intersected with the development of sparse inference techniques recently. Such inference techniques, in particular applied to high-dimensional linear regression, are of importance in statistics, and have been an area of major developments over the past two decades, with deep results in the field of compressed sensing, and more generally convex relaxation techniques [80, 15, 16, 19, 18]. These led to powerful inference algorithms working under a sparsity assumption, thanks to fast and scalable convex optimization algorithms [4]. The most popular method allowing to deal with sparsity and variable selection is the LASSO [81], which is ℓ_1-penalized least-squares, with improvements such as the Adaptive LASSO [94], among a large set of other sparsity-inducing penalizations [13, 5].

Within the past few years, a large amount of theoretical results have been established to understand regularization methods for the sparse linear regression model, using so-called oracle inequalities for the prediction and estimation errors [43, 44, 53], see also [13, 29] for nice surveys on this topic. Another line of works focuses on variable selection, trying to recover the support of the regression coefficients with a high probability [49, 44, 21]. Other types of loss functions [85] or penalizations [25, 11] have also been considered. Very recently, the sorted-ℓ_1 norm penalization has been introduced [11, 12, 76] and very strong statistical properties have been shown. In particular, when covariates are orthogonal, SLOPE allows to recover the support of the regression coefficients with a control on the False Discovery Rate [11]. For i.i.d covariates with a multivariate Gaussian distribution, oracle inequalities with optimal minimax rates have been shown, together with a control on a quantity which is very close to the FDR [76]. For more general covariate distributions, oracle inequalities with an optimal convergence rate are obtained in [14].

However, many high-dimensional datasets, with hundreds or thousands of covariates, do suffer from the presence of outliers. Robust regression and detection of outliers in a high-dimensional setting is therefore important. Increased dimensionality and complexity of the data may amplify the chances of an observation being an outlier, and this can have a strong negative impact on the statistical analysis. In such settings, many of the aforementioned outlier detection methods do not work well. A new technique for outliers detection in a high-dimensional setting is proposed in [1], which tries to find the outliers by studying the behavior of projections from the data set. A small set of other attempts to deal with this problem have been proposed in literature [86, 67, 32, 73, 26], and are described below with more details.

2 Contributions of the paper

Our focus is on possibly high dimensional linear regression where observations can be contaminated by gross errors. This so-called mean-shifted outliers model can be described as follows:

$$y_i = x_i^\top \beta^\star + \mu_i^\star + \varepsilon_i \tag{II.1}$$

for $i = 1, \dots, n$, where n is the sample size. A non-zero $\mu_i^\star$ means that observation i is an outlier, and $\beta^\star \in \mathbb{R}^p$, $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$ and $\varepsilon_i \in \mathbb{R}$ respectively stand for the linear regression coefficients, vector of covariates, label and noise of sample i. For the sake of simplicity we assume throughout the paper that the noise is i.i.d centred Gaussian with known variance σ^2.

2.1 Related works

We already said much about the low-dimensional problem so we focus in this part on the high-dimensional one. The leave-one-out technique has been extended in [86] to high-dimension and general regression cases, but the masking and swamping problems remains. In other models, outliers detection in high-dimension also includes distance-based approaches [67] where the idea is to find the center of the data and then apply some thresholding rule. The model (II.1) considered here has been recently studied with LASSO penalizations [32] and hard-thresholding [73]. LASSO was used also in [26], but here outliers are modelled in the variance of the noise. In [32, 73], that are closer to our approach, the penalization is applied differently: in [32], the procedure named Extended-LASSO uses two different ℓ_1 penalties for β and μ, with tuning parameters that are fixed according to theoretical results, while the IPOD procedure from [73] applies the same penalization to both vectors, with a regularization parameter tuned with a modified BIC criterion. In [32], error bounds and a signed support recovery result are obtained for both the regression and intercepts coefficients. However, these results require that the magnitude of the coefficients is very large, which is one of the issues that we want to overcome with this paper.

It is worth mentioning that model (II.1) can be written in a concatenated form $y = Z\gamma^\star + \varepsilon$, with Z being the concatenation of the covariates matrix X (with lines given by the x_i's) and the identity matrix I_n in $\mathbb{R}^n$, and $\gamma^\star$ being the concatenation of $\beta^\star$ and $\mu^\star$. This leads to a regression problem with a very high dimension $n+p$ for the vector

$\gamma^\star$. Working with this formulation, and trying to estimate $\gamma^\star$ directly is actually a bad idea. This point is illustrated experimentally in [32], where it is shown that applying two different LASSO penalizations on β and μ leads to a procedure that outperforms the LASSO on the concatenated vector. The separate penalization is even more important in case of SLOPE, whose aim is FDR control for the support recovery of $\mu^\star$. Using SLOPE directly on $\gamma^\star$ would mix the entries of μ and β together, which would make FDR control practically impossible due to the correlations between covariates in the X matrix.

2.2 Main contributions

Given a vector $\lambda = [\lambda_1 \cdots \lambda_m] \in \mathbb{R}_+^m$ with non-negative and non-increasing entries, we define the sorted-ℓ_1 norm of a vector $x \in \mathbb{R}^m$ as

$$\forall x \in \mathbb{R}^m,\ J_\lambda(x) = \sum_{j=1}^{m} \lambda_j |x|_{(j)}, \tag{II.2}$$

where $|x|_{(1)} \geq |x|_{(2)} \geq \cdots \geq |x|_{(m)}$. In [11] and [12] the sorted-ℓ_1 norm was used as a penalization in the Sorted L-One Penalized Estimator (SLOPE) of coefficients in the multiple regression. Degenerate cases of SLOPE are ℓ_1-penalization whenever λ_j are all equal to a positive constant, and null-penalization if this constant is zero. We apply two different SLOPE penalizations on β and μ, by considering the following optimization problem:

$$\min_{\beta \in \mathbb{R}^p, \mu \in \mathbb{R}^n} \left\{ \|y - X\beta - \mu\|_2^2 + 2\rho_1 J_{\tilde{\lambda}}(\beta) + 2\rho_2 J_\lambda(\mu) \right\} \tag{II.3}$$

where ρ_1 and ρ_2 are positive parameters, X is the $n \times p$ covariates matrix with rows $x_1, \ldots, x_n$, $y = [y_1 \cdots y_n]^T$, $\mu = [\mu_1 \cdots \mu_n]^T$, $\|u\|_2$ is the Euclidean norm of a vector u and $\lambda = [\lambda_1 \cdots \lambda_n]$ and $\tilde{\lambda} = [\tilde{\lambda}_1 \cdots \tilde{\lambda}_p]$ are two vectors with non-increasing and non-negative entries.

In this artice we provide the set of sequences λ and $\tilde{\lambda}$ which allow to obtain better error bounds for estimation of $\mu^\star$ and $\beta^\star$ than previously known ones [32], see Section 3 below. Moreover, in Section 4 we provide specific sequences which, under some asymptotic regime, lead to a control of the FDR for the support selection of $\mu^\star$, and such that the power of the procedure (II.3) converges to one. Procedure (II.3)

is therefore, to the best of our knowledge, the first proposed in literature to *robustly estimate $\beta^\star$, estimate and detect outliers at the same time, with a control on the FDR for the multi-test problem of support selection of $\mu^\star$, and power consistency.*

We compare in Section 5 our procedure to the recent alternatives for this problem, that is the IPOD procedure [73] and the Extended-Lasso [32]. The numerical experiments given in Section 5 confirm the theoretical findings from Sections 3 and 4. As shown in our numerical experiments, the other procedures fail to guarantee FDR control or exhibit a lack of power when outliers are difficult to detect, namely when their magnitude is not far enough from the noise-level. It is particularly noticeable that our procedure overcomes this issue.

The theoretical results proposed in this paper are based on two popular assumptions in compressed sensing or other sparsity problems, similar to the ones from [32]: first, a Restricted Eigenvalues (RE) condition [43] on X, then a mutual incoherence assumption [55] between X and I_n, which is natural since it excludes settings where the column spaces of X and I_n are impossible to distinguish. Proofs of results stated in Sections 3 and 4 are given in Section 9 and 10, while preliminary results are given in Sections 7 and 8. Section 11 provides contains supplementary extra numerical results.

3 Upper bounds for the estimation of $\beta^\star$ and $\mu^\star$

Throughout the paper, n is the sample size whereas p is the number of covariables, so that $X \in \mathbb{R}^{n\times p}$. For any vector u, $|u|_0$, $\|u\|_1$ and $\|u\|_2$ denote respectively the number of non-zero coordinates of u (also called sparsity), the ℓ_1-norm and the Euclidean norm. We denote respectively by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ the smallest and largest eigenvalue of a symmetric matrix A. We work under the following assumption

Assumption II.1. *We assume the following sparsity assumption*:

$$\left|\beta^\star\right|_0 \le k \quad \text{and} \quad \left|\mu^\star\right|_0 \le s \tag{II.4}$$

for some positive integers k and s, and we assume that the columns of X are normalized, namely $\|Xe_i\|_2 = 1$ for $i = 1,\dots,n$, where e_i stands for the i-th element of the canonical basis.

For the results of this Section, we consider procedure (II.3) with the following choice of λ:

$$\lambda_i = \sigma\sqrt{\log\Big(\frac{2n}{i}\Big)}, \tag{II.5}$$

for $i = 1,\dots,n$, and we consider three possibilities for $\tilde{\lambda}$, corresponding to no penalization, ℓ_1 penalization and SLOPE penalization on β.

Table II.1 below gives a quick view of the convergence rates of the squared ℓ_2 estimation errors of $\beta^\star$ and $\mu^\star$ obtained in Theorems II.2, II.3 and II.4. We give also the convergence rate obtained in [32] for ℓ_1 penalization applied to β and μ. In particular, we see that using two SLOPE penalizations leads to a better convergence rate than the use of ℓ_1 penalizations. Condition II.1 below is a Restricted Eigenvalue

Table II.1: Convergence rates, up to constants, associated to several penalization techniques. NO means no-penalization, L1 stands for ℓ_1 penalization, while SL1 stands for SLOPE. We observe that SL1 + SL1 leads to a better convergence rate than L1 + L1.

Penalization (β / μ)	Convergence rates	Reference
NO/SL1	$p \vee s\log(n/s)$	Theorem II.2
L1/L1	$k\log p \vee s\log n$	[32]
L1/SL1	$k\log p \vee s\log(n/s)$	Theorem II.3
SL1/SL1	$k\log(p/k) \vee s\log(n/s)$	Theorem II.4

(RE) type of condition which is adapted to our problem. Such an assumption is known to be mandatory in order to derive fast rates of convergence for penalizations based on the convex-relaxation principle [92].

Condition II.1. *Consider two vectors $\lambda = (\lambda_i)_{i=1,\dots,n}$ and $\tilde{\lambda} = (\tilde{\lambda}_i)_{i=1,\dots,p}$ with non-increasing and positive entries, and consider positive integers k, s and $c_0 > 0$. We define the cone $\mathscr{C}(k,s,c_0)$ of all vectors $[\beta^\top, \mu^\top]^\top \in \mathbb{R}^{p+n}$ satisfying*

$$\sum_{j=1}^{p} \frac{\tilde{\lambda}_j}{\tilde{\lambda}_p} \big|\beta\big|_{(j)} + \sum_{j=1}^{n} \frac{\lambda_j}{\lambda_n} \big|\mu\big|_{(j)} \le (1+c_0)\big(\sqrt{k}\|\beta\|_2 + \sqrt{s}\|\mu\|_2\big). \tag{II.6}$$

We also define the cone $\mathscr{C}^p(s, c_0)$ *of all vectors* $[\beta^\top, \mu^\top]^\top \in \mathbb{R}^{p+n}$ *satisfying*

$$\sum_{j=1}^{n} \frac{\lambda_j}{\lambda_n} |\mu|_{(j)} \leq (1 + c_0)\big(\sqrt{p}\|\beta\|_2 + \sqrt{s}\|\mu\|_2\big). \tag{II.7}$$

We assume that there are constants $\kappa_1, \kappa_2 > 0$ *with* $\kappa_1 > 2\kappa_2$ *such that* X *satisfies the following, either for all* $[\beta^\top, \mu^\top]^\top \in \mathscr{C}(k, s, c_0)$ *or for all* $[\beta^\top, \mu^\top]^\top \in \mathscr{C}^p(s, c_0)$:

$$\|X\beta\|_2^2 + \|\mu\|_2^2 \geq \kappa_1\big(\|\beta\|_2^2 + \|\mu\|_2^2\big) \tag{II.8}$$

$$|\langle X\beta, \mu\rangle| \leq \kappa_2\big(\|\beta\|_2^2 + \|\mu\|_2^2\big). \tag{II.9}$$

Equation (II.7) corresponds to the particular case where we do not penalize the regression coefficient β, namely $\tilde{\lambda}_i = 0$ for all i. Note also that Condition II.1 entails

$$\|X\beta + \mu\|_2 \geq \sqrt{\kappa_1 - 2\kappa_2}\sqrt{\|\beta\|_2^2 + \|\mu\|_2^2},$$

which actually corresponds to a RE condition on $[X^\top I_n]^\top$ and that Equation (II.8) is satisfied if X satisfies a RE condition with constant $\kappa < 1$. Finally, note that Equation (II.9), called mutual incoherence in the literature of compressed sensing, requires in this context that for all β and μ from the respective cones the potential regression predictor $X\beta$ is sufficiently not-aligned with potential outliers μ. An extreme case of violation of this assumption occurs when $X = I_n$, where we cannot separate the regression coefficients from the outliers.

The Condition II.1 is rather mild.Specifically, Theorem II.1 below, shows that it holds with large probability whenever X has i.i.d $\mathcal{N}(0, \Sigma)$ rows, with $\lambda_{min}(\Sigma) > 0$, and the vectors β and μ are sufficiently sparse.

Theorem II.1. *Let* $X' \in \mathbb{R}^{n\times p}$ *be a random matrix with i.i.d* $\mathcal{N}(0, \Sigma)$ *rows and* $\lambda_{min}(\Sigma) > 0$. *Let* X *be the corresponding matrix with normalized columns. Given positive integers* k, s *and* $c_0 > 0$, *define* $r = s \vee k(1 + c_0)^2$. *If*

$$\sqrt{n} \geq C\sqrt{r} \quad \text{and} \quad \sqrt{n} \geq C'\sqrt{r\log(p \vee n)}$$

with

$$C \geq 30\frac{\sqrt{\lambda_{\max}(\Sigma)}}{\min_j \Sigma_{jj}}\Big(\frac{256\times 5\max_j \Sigma_{jj}}{\lambda_{\min}(\Sigma)} \vee 16\Big) \quad \textit{and} \quad C' \geq 72\sqrt{10}\frac{(\max_j \Sigma_{jj})^{3/2}}{\min_j \Sigma_{jj}\sqrt{\lambda_{min}(\Sigma)}},$$

then there are $c, c' > 0$ such that for any $[\beta^\top, \mu^\top]^\top \in \mathscr{C}(k, s, c_0)$, we have

$$\|X\beta\|_2^2 + \|\mu\|_2^2 \geq \min\Big\{\frac{\lambda_{\min}(\Sigma)}{128(\max_j \Sigma_{jj})^2}, \frac{1}{8}\Big\}\big(\|\beta\|_2^2 + \|\mu\|_2^2\big)$$
$$2|\langle X\beta, \mu\rangle| \leq \min\Big\{\frac{\lambda_{\min}(\Sigma)}{256\times 5(\max_j \Sigma_{jj})^2}, \frac{1}{16}\Big\}\big(\|\beta\|_2^2 + \|\mu\|_2^2\big)$$

with a probability greater than $1 - cn\exp(-c'n)$. These inequalities also hold for any $[\beta^\top, \mu^\top]^\top \in \mathscr{C}^p(s, c_0)$ when k is replaced by p in the above conditions.

The proof of Theorem II.1 is given in Appendix 9.1. It is based on recent bounds results for Gaussian random matrices [65]. The numerical constants in Theorem II.1 are far from optimal and chosen for simplicity so that $\kappa_1 > 2\kappa_2$ as required in Assumption II.1. A typical example for Σ is the Toeplitz matrix $[a^{|i-j|}]_{i,j}$ with $a \in [0, 1)$, for which $\lambda_{\min}(\Sigma)$ is equal to $1 - a$ [65]. The required lower bound on n is non-restrictive, since k and s correspond to the sparsity of $\beta^\star$ and $\mu^\star$, that are typically much smaller than n. Note also that $\mathscr{C}^p(s, c_0)$ will only be used in low dimension, and in this case p is again much smaller than n.

Let us define $\kappa = \sqrt{\kappa_1 - 2\kappa_2}$ for the whole Section, with κ_1 and κ_2 defined in Assumption II.1. The three theorems below and their proofs are very similar in nature, but differ in some details, therefore are stated and proved separately. We emphasize that the proofs give slightly more general versions of the theorems, allowing the same result with $\hat{\mu}$ having any given support containing $\mathrm{Supp}(\mu^\star)$. This is of great theoretical interest and is a key point for the support detection of $\mu^\star$ investigated in 4. The proof use a very recent bound on the inner product between a white Gaussian noise and any vector, involving the sorted ℓ_1 norm [14]. Our first result deals with linear regression with outliers and no sparsity assumption on $\beta^\star$. We consider procedure (II.3) with no penalization on β, namely

$$(\hat{\beta}, \hat{\mu}) \in \operatorname*{argmin}_{\beta\in\mathbb{R}^p, \mu\in\mathbb{R}^n} \Big\{\|y - X\beta - \mu\|_2^2 + 2\rho J_\lambda(\mu)\Big\}, \tag{II.10}$$

with J_λ given by (II.2) and weights λ given by (II.5), and with $\rho \geq 2(4+\sqrt{2})$. Theorem II.2, below, shows that a convergence rate for procedure (II.10) is indeed $p \vee s\log(n/s)$, as reported in Table II.1 above.

Theorem II.2. *Suppose that Assumption II.1 is met with $k = p$, and that X satisfies Assumption II.1 on the cone $\mathscr{C}(k_1, s_1, 4)$ with $k_1 = p/\log 2$ and $s_1 = s\log(2en/s)/\log 2$. Then, the estimators $(\hat{\beta}, \hat{\mu})$ given by* (II.10) *satisfy*

$$\|\hat{\beta} - \beta^\star\|_2^2 + \|\hat{\mu} - \mu^\star\|_2^2 \leq \frac{4\rho^2}{\kappa^4} \sum_{j=1}^{s} \lambda_j^2 + \frac{5\sigma^2}{\kappa^4} p \leq \frac{\sigma^2}{\kappa^4}\left(4\rho^2 s\log\left(\frac{2en}{s}\right) + 5p\right),$$

with a probability larger than $1 - (s/2n)^s/2 - e^{-p}$.

The proof of Theorem II.2 is given in Appendix 9.2. The second result involves a sparsity assumption on $\beta^\star$ and considers ℓ_1 penalization for β. We consider this time

$$(\hat{\beta}, \hat{\mu}) \in \operatorname*{argmin}_{\beta,\mu} \left\{ \|y - X\beta - \mu\|_2^2 + 2\nu\|\beta\|_1 + 2\rho J_\lambda(\mu) \right\}, \tag{II.11}$$

where $\nu = 4\sigma\sqrt{\log p}$ is the regularization level for ℓ_1 penalization, $\rho \geq 2(4+\sqrt{2})$ and J_λ is given by (II.2). Theorem II.3, below, shows that a convergence rate for procedure (II.11) is indeed $k\log p \vee s\log(n/s)$, as reported in Table II.1 above.

Theorem II.3. *Suppose that Assumption II.1 is met and that X satisfies Assumption II.1 on the cone $\mathscr{C}(k_1, s_1, 4)$ with $k_1 = 16k\log p/\log 2$ and $s_1 = s\log(2en/s)/\log 2$. Suppose also that $\sqrt{\log p} \geq \rho\log 2/4$. Then, the estimators $(\hat{\beta}, \hat{\mu})$ given by* (II.11) *satisfy*

$$\|\hat{\beta} - \beta^*\|_2^2 + \|\hat{\mu} - \mu^*\|_2^2 \leq \frac{36}{\kappa^4}\sigma^2 k\log p + \frac{4\rho^2}{\kappa^4} \sum_{j=1}^{s} \lambda_j^2 \leq \frac{4\sigma^2}{\kappa^4}\left(9k\log p + \rho^2 s\log\left(\frac{2en}{s}\right)\right),$$

with a probability larger than $1 - (s/2n)^s/2 - 1/p$.

The proof of Theorem II.3 is given in Appendix 9.4. The third result is obtained using SLOPE both on β and μ, namely

$$(\hat{\beta}, \hat{\mu}) \in \operatorname*{argmin}_{\beta,\mu} \left\{ \|y - X\beta - \mu\|_2^2 + 2\rho J_{\tilde{\lambda}}(\beta) + 2\rho J_\lambda(\mu) \right\} \tag{II.12}$$

where $\rho \geq 2(4+\sqrt{2})$, J_λ is given by (II.2), and where

$$\tilde{\lambda}_j = \sigma\sqrt{\log\Big(\frac{2p}{j}\Big)}$$

for $j = 1,\dots,p$. Theorem II.4, below, shows that the rate of convergence of estimators provided by (II.12) is indeed $k\log(p/k) \vee s\log(n/s)$, as presented in Table II.1.

Theorem II.4. *Suppose that Assumption II.1 is met and that X satisfies Assumption II.1 on the cone $\mathscr{C}(k_1, s_1, 4)$ with $k_1 = k\log(2ep/k)/\log 2$ and $s_1 = s\log(2en/s)/\log 2$. Then, the estimators $(\hat{\beta}, \hat{\mu})$ given by* (II.12) *satisfy*

$$\begin{aligned} \|\hat{\beta}-\beta^*\|_2^2 + \|\hat{\mu}-\mu^*\|_2^2 &\leq \frac{C'}{\kappa^4}\Big(\sum_{j=1}^{k}\tilde{\lambda}_j^2 + \sum_{j=1}^{s}\lambda_j^2\Big) \\ &\leq \frac{C'\sigma^2}{\kappa^4}\left(k\log\left(\frac{2ep}{k}\right) + s\log\left(\frac{2en}{s}\right)\right), \end{aligned} \tag{II.13}$$

with a probability greater than $1-(s/2n)^s/2-(k/2p)^k/2$, where $C' = 4\rho^2 \vee (3+C)^2/2$.

The proof of Theorem II.4 is given in Appendix 9.4. Note that according to Theorem II.1, the assumptions of Theorem II.4 are satisfied with probability converging to one when the rows of X are i.i.d from the multivariate Gaussian distribution with the positive definite covariance matrix, and when the signal is sparse such that $(k \vee s)\log(n \vee p) = o(n)$.

4 Asymptotic FDR control and power for the selection of the support of $\mu^\star$

We consider the multi-test problem with null-hypotheses

$$H_i : \mu_i^* = 0$$

for $i = 1,\dots,n$, and we consider the multi-test that rejects H_i whenever $\hat{\mu}_i \neq 0$, where $\hat{\mu}$ (and $\hat{\beta}$) are given either by (II.10), (II.11) or (II.12). When H_i is rejected, or "discovered", we consider that sample i is an outlier. Note however that in this case, the value of $\hat{\mu}_i$ gives extra information on how much sample i is oulying.

We use the FDR as a standard Type I error for this multi-test problem [8]. The FDR is the expectation of the proportion of falses discoveries among all discoveries. Letting V (resp. R) be the number of false rejections (resp. the number of rejections), the FDR is defined as

$$\mathrm{FDR}(\hat{\mu}) = \mathbb{E}\left[\frac{V}{R \vee 1}\right] = \mathbb{E}\left[\frac{\#\{i : \mu_i^\star = 0, \hat{\mu}_i \neq 0\}}{\#\{i : \hat{\mu}_i \neq 0\}}\right]. \tag{II.14}$$

We use the Power to measure the Type II error for this multi-test problem. The Power is the expectation of the proportion of true discoveries. It is defined as

$$\Pi(\hat{\mu}) = \mathbb{E}\left[\frac{\#\{i : \mu_i^\star \neq 0, \hat{\mu}_i \neq 0\}}{\#\{i : \mu_i^\star \neq 0\}}\right], \tag{II.15}$$

the Type II error is then given by $1 - \Pi(\hat{\mu})$.

For the linear regression model without outliers, a multi-test for the support selection of $\beta^\star$ with controlled FDR based on SLOPE is given in [11] and [12]. Specifically, it is shown that SLOPE with weights

$$\lambda_i^{\mathrm{BH}} = \sigma\Phi^{-1}\Big(1 - \frac{iq}{2n}\Big) \tag{II.16}$$

for $i = 1, \ldots, n$, where Φ is the cumulative distribution function of $\mathcal{N}(0,1)$ and $q \in (0,1)$, controls FDR at the level q in the multiple regression problem with orthogonal design matrix $X^T X = I$. It is also observed that when the columns of X are not orthogonal but independent the weights have to be substantially increased to guarantee FDR control. This effect results from the random correlations between columns of X and the shrinkage of true nonzero coefficients, and in context of LASSO have been thoroughly discussed in [75].

In this paper we substantially extend current results on FDR controlling properties of SLOPE. Specifically, Theorem II.5 below gives asymptotic controls of $\mathrm{FDR}(\hat{\mu})$ and $\Pi(\hat{\mu})$ for the procedures (II.10), (II.11) and (II.12), namely different penalizations on β and SLOPE applied on μ, with slightly increased weights

$$\lambda = (1 + \epsilon)\lambda^{\mathrm{BH}}, \tag{II.17}$$

where $\epsilon > 0$, see also [76]. This choice of λ also yields optimal convergence rates,

however considering it in Section 3 would lead to some extra technical difficulties. Under appropriate assumptions on p, n, the signal sparsity and the magnitude of outliers, Theorem II.5 not only gives FDR control, but also proves that the Power is actually going to 1.

Note that all asymptotics considered here are with respect to the sample size n, namely the statement $d \to +\infty$ means that $d = d_n \to +\infty$ with $n \to +\infty$.

Theorem II.5. *Suppose that there is a constant M such that the entries of X satisfy $|x_{i,j}|\sqrt{n} \leq M$ for all $i, j \in \{1, \dots n\}$, and suppose that*

$$|\mu_i^\star| \geq (1 + \rho(1 + 2\epsilon))2\sigma\sqrt{\log n}$$

for any $i = 1, \dots, n$ such that $\mu_i^\star \neq 0$. Suppose also that $s \to +\infty$. Then, consider $(\hat{\beta}, \hat{\mu})$ given either by (II.10), (II.11) *and* (II.12), *with λ given by* (II.17). *For Procedure* (II.10), *assume the same as in Theorem II.2, and that*

$$\frac{p(s\log(n/s) \vee p)}{n} \to 0.$$

For Procedure (II.11), *assume the same as in Theorem II.3, and that*

$$\frac{(s\log(n/s) \vee k\log p)^2}{n} \to 0.$$

For Procedure (II.12), *assume the same as in Theorem II.4, and that*

$$\frac{\left(s\log(n/s) \vee k\log(p/k)\right)^2}{n} \to 0.$$

Then, the following properties hold:

$$\mathrm{TPR}(\hat{\mu}) \to 1, \qquad \limsup \mathrm{FDR}(\hat{\mu}) \leq q. \tag{II.18}$$

The proof of Theorem II.5 is given in Appendix 10. It relies on a careful look at the KKT conditions, also known as the dual-certificate method [49] or resolvent solution [76]. The assumptions of Theorem II.5 are natural. The boundedness assumption on the entries of X are typically satisfied with a large probability when X has random uniform (and uniformly bounded) entries, with columns normalized to

1. Note that we could allow Gaussian distribution for the rows of X by assuming $|x_{i,j}| \leq \frac{M \log(p \vee n)}{\sqrt{n}}$. Then, we should simply add the logarithmic factor $\log(p \vee n)$ in the last two asymptotic conditions for our results to remain valid. When $n \to +\infty$, it is also natural to assume that $s \to +\infty$ (let us recall that s stands for the sparsity of the sample outliers $\mu \in \mathbb{R}^n$). The asymptotic assumptions roughly ask for the rates in Table II.1 to converge to zero. Finally, the assumption on the magnitude of the non-zero entries of μ^* is somehow unavoidable, since it allows to distinguish outliers from the Gaussian noise. We emphasize that good numerical performances are actually obtained with lower magnitudes, as illustrated in Section 5.1.

5 Numerical experiments

In this section, we illustrate the performance of procedure (II.10) and procedure (II.12) both on simulated and real-world datasets, and compare them to several state-of-the art baselines described below. Experiments are done using the open-source `tick` library, available at https://x-datainitiative.github.io/tick/, notebooks allowing to reproduce our experiments are available on demand to the authors.

5.1 Simulation settings

The matrix X is simulated as a matrix with i.i.d rows distributed as $\mathcal{N}(0, \Sigma)$, with Toeplitz covariance $\Sigma_{i,j} = \rho^{|i-j|}$ for $1 \leq i, j \leq p$, with moderate correlation $\rho = 0.4$. Some results with higher correlation $\rho = 0.8$ are given in Section 11. The columns of X are normalized to 1. We simulate n observations according to model (II.1) with $\sigma = 1$ and $\beta_i^* = \sqrt{2 \log p}\sqrt{n}$. Two levels of magnitude are considered for $\mu^\star$: *low-magnitude*, where $\mu_i^\star = \sqrt{2 \log n}$ and *large-magnitude*, where $\mu_i^\star = 5\sqrt{2 \log n}$. In all reported results based on simulated datasets we display the average FDR, TPR and MSE over 100 replications.

Setting 1 (low-dimension) This is the setting described above with $n = 1000$ and $p = 20$. Here $\beta_1^\star = \ldots = \beta_{20}^\star = \sqrt{2 \log 20}\sqrt{n}$. Moreover, the sparsity of $\mu^\star$ varies between 1% to 50%

Setting 2 (high-dimension) This is the setting described above with $n = 1500$, $p = 2000$ and a sparse $\beta^\star$ with sparsity $k = 50$, with non-zero entries chosen uniformly at random. Moreover, the sparsity of $\mu^\star$ varies between 1% to 20%

5.2 Considered procedures

We consider the following baselines, featuring the best methods available in literature for the joint problem of outlier detection and estimation of the regression coefficients, together with the methods introduced in this paper.

E-SLOPE It is procedure (II.12). The weights used in SLOPE penalization of μ are given by

$$\lambda_i^{\mathrm{BH}}(q;n) = \sigma\Phi^{-1}\big(1-\frac{iq}{2n}\big), \quad i = 1,\dots,n, \tag{II.19}$$

with $q = 5\%$ (target FDR). In high-dimensional setting, the weights used in SLOPE penalization of β are given by $(\lambda_j^{\mathrm{BH}}(1;n))_{j=1,\dots,p}$. Similar results for $q = 10\%$ and $q = 20\%$ are provided in Section 11.

E-LASSO This is Extended LASSO from [32], that uses two dedicated ℓ_1-penalizations for β and μ with respective tuning parameters $\lambda_\beta = 2\sigma\sqrt{\log p}$ and $\lambda_\mu = 2\sigma\sqrt{\log n}$.

Soft-IPOD This is (soft-)IPOD from [73]. The Soft-IPOD considers Lasso penalization on μ, and a BIC criterion is used to choose the tuning parameter of ℓ_1-penalization. Note that this procedure, which involves a QR decomposition of X, makes sense only for p significantly smaller than n, so that we do not report the performances of IPOD on simulations with a large p.

Hard-IPOD This is (hard-)IPOD from [73]. The hard-IPOD considers Hard-thresholding on μ, which leads to a non convex procedure. Tuning parameter is chosen in the same way as Soft-IPOD, and the same remark holds about the high-dimensional cases.

SLOPE It is SLOPE applied to the concatenated problem, namely $y = Z\gamma^\star + \varepsilon$, where Z is the concatenation of X and I_n and $\gamma^\star$ is the concatenation of $\beta^\star$ and $\mu^\star$. We use a single SLOPE penalization on γ, with weights given by $(\lambda_j^{\mathrm{BH}}(q;n+p))_{j=1,\dots,n+p}$ We report the performances of this procedure both in low-dimensional

and high-dimensional experiments, but as it always penalizes β it would appear more relevant in the high-dimensional cases. This is considered mostly to illustrate the fact that working on the concatenated problem is indeed a bad idea, and that two distinct penalizations must be used on β and μ.

Note that the difference between IPOD and E-LASSO is that, as explained in [32], the weights used for E-LASSO to penalize μ (and β in high-dimension) are fixed, while the weights in IPOD are data-dependent. Another difference is that IPOD does not extend well to a high-dimensional setting, since its natural extension (considered in [73]) is a thresholding rule on the the concatenated problem, which is poorly performing, as explained before and as illustrated in our numerical experiments. Another problem is that there is no clear extension of the modified BIC criterion proposed in [73] for high-dimensional problems.

The tuning of the SLOPE or ℓ_1 penalizations in the procedure described above requires the knowledge of the noise level. We overcome this simply by plugging wherever it is necessary a robust estimation of the variance: we first fit a Huber regression model, and apply a robust estimation of the variance of its residuals. All procedures considered in our experiments use this same variance estimate (needed only for real datasets).

Remark II.1. *The noise level can be estimated directly by the Huber regression since in our simulations $p < n$. When p is comparable to or larger than n and the signal (both $\beta^\star$ and $\mu^\star$) is sufficiently sparse one can jointly estimate the noise level and other model parameters in the spirit of* scaled *LASSO [77]. The corresponding iterative procedure for SLOPE was proposed and investigated in [11] in the context of high-dimensional regression with independent regressors.*

5.3 Metrics

In our experiments, we report the "MSE coefficients", namely $\frac{1}{n}\|\hat{\beta} - \beta^\star\|_2^2$ and the "MSE intercepts", namely $\frac{1}{n}\|\hat{\mu} - \mu^\star\|_2^2$. We report also the FDR (II.14) and the Power (II.15) to assess the procedures for the problem of outliers detection, where the expectations are approximated by averages over 100 simulations.

5.4 Results and conclusions on simulated datasets

We comment the displays provided in Figures II.1, II.2 and II.3 below. On Simulation Setting 2 we only display results for the low magnitude case, since it is the most challenging one.

- In the low dimensional setting, our procedure E-SLOPE allows for almost perfect FDR control. Note that in this setting the MSE is plotted after debiasing the estimators, performing ordinary least squares on the selected support. The only case where some procedures outperform E-SLOPE is Setting 1 (see Figure II.1, where outliers are easy to detect.
- In the sparse (on β) high dimensional setting with correlated regressors, E-SLOPE allows to keep FDR below the nominal level even when the outliers consist 50% of the total data points. It also allows to maintain a small MSE and high power.
- E-SLOPE provides a massive gain of power compared to previous state-of-the-art procedures (power is increased by more than 30%) in settings where outliers are difficult to detect.

5.5 PGA/LPGA dataset

This dataset (available at http://users.stat.ufl.edu/~winner/datasets.html) contains Distance and Accuracy of shots, for Professional Golf Association (PGA) and Ladies Professional Golf Association (LPGA) players in 2008. This toy example, where the output Y is the Distance of shot and X is the Accuracy (in this context $p = 1$), will allow us to visually compare the performance of IPOD, E-LASSO and E-SLOPE. Our data contain 197 points corresponding to PGA (men) players, to which we add 8 points corresponding to LPGA (women) players, injecting outliers. We apply SLOPE and LASSO on μ with several levels of penalization. This leads to the "regularization paths" given in the top plots of Figure II.4, that shows the value of the 205 sample intercepts $\hat{\mu}$ as a function of the penalization level used in SLOPE and LASSO. Vertical lines indicate the choice of the parameter according the corresponding method (E-SLOPE, E-LASSO, IPOD). We observe that E-SLOPE correctly discovers the confirmed outliers (women data), together with two men observations

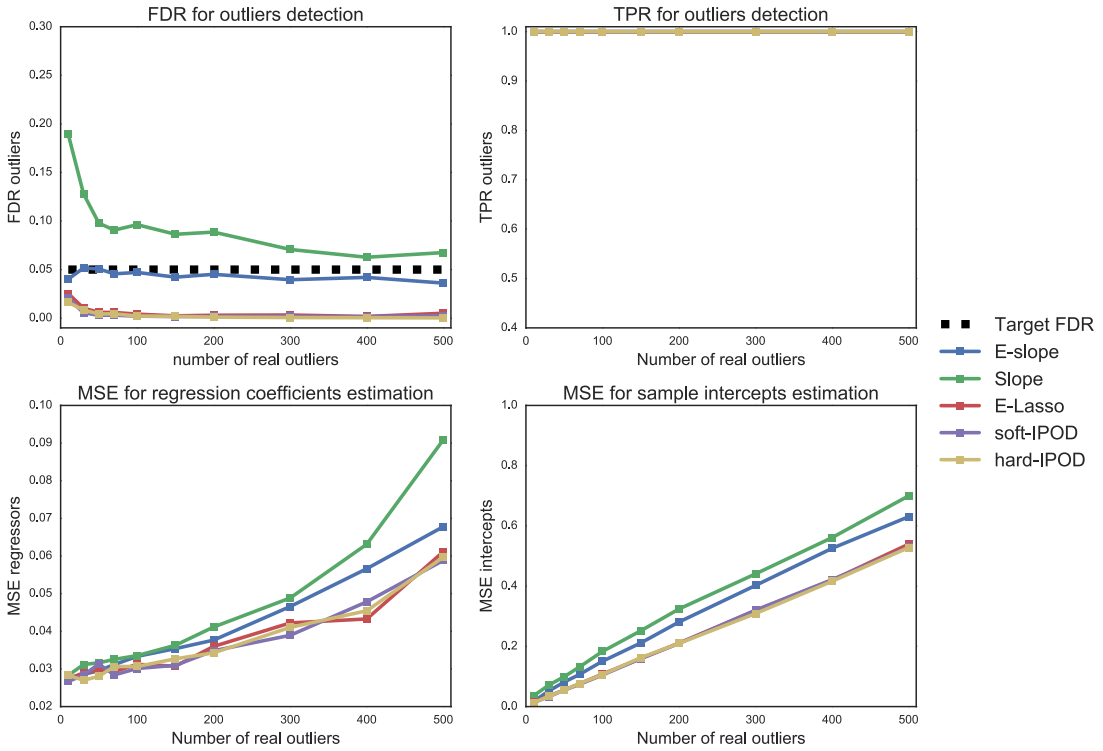

Figure II.1: Results for Simulation Setting 1 with high-magnitude outliers. The first row gives the FDR (left) and power (right) of each considered procedure for outliers discoveries. The second row gives the MSE for regressors (left) and intercepts (right).

that could be investigated as outliers in view of the scatter plot. IPOD procedure does quite good, with no false discovery, but misses some real outliers (women data) and the suspicious points detected by E-SLOPE. E-LASSO does not make any false discovery but clearly reveals a lack of power, with only one discovery.

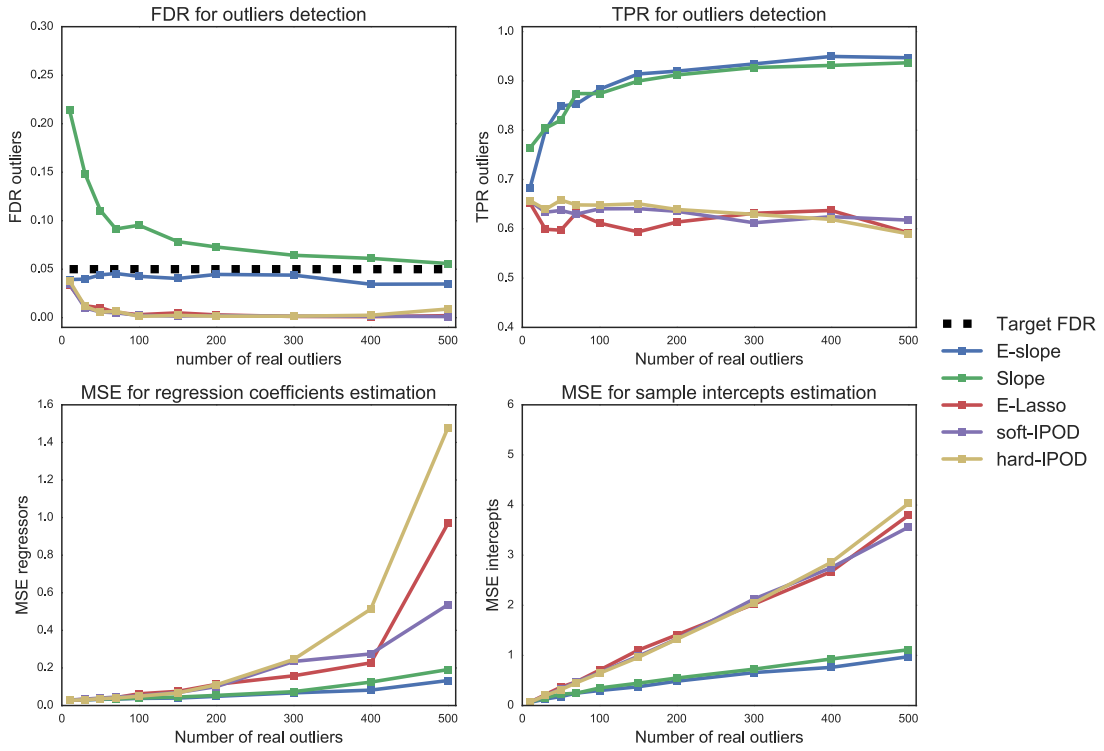

Figure II.2: Results for Simulation Setting 1 with low-magnitude outliers. The first row gives the FDR (left) and power (right) of each considered procedure for outliers discoveries. The second row gives the MSE for regressors (left) and intercepts (right). E-SLOPE gives the best power and provides the best MSEs while keeping the FDR below the desired level.

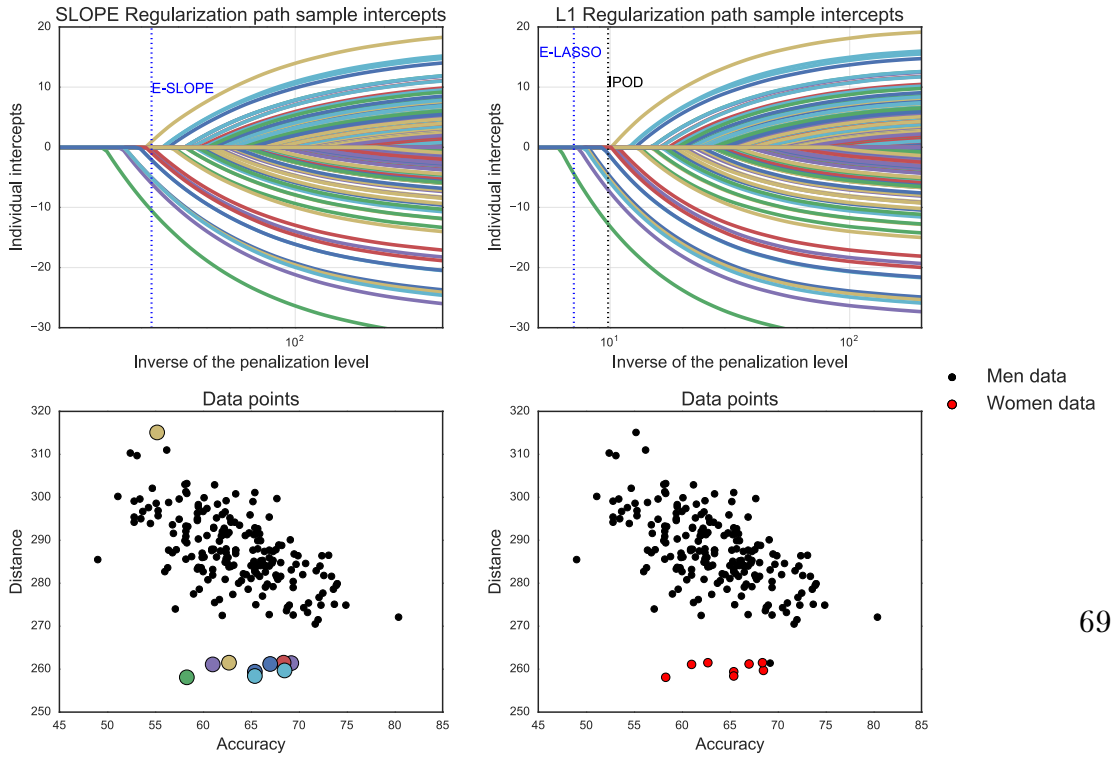

Figure II.4: PGA/LPGA dataset: top plots show the regularization paths for both types of penalization, bottom-left plot is a scatter plot of the data, with colored points corresponding to the discoveries made by E-SLOPE, bottom-right plot show the original data and the true outliers.

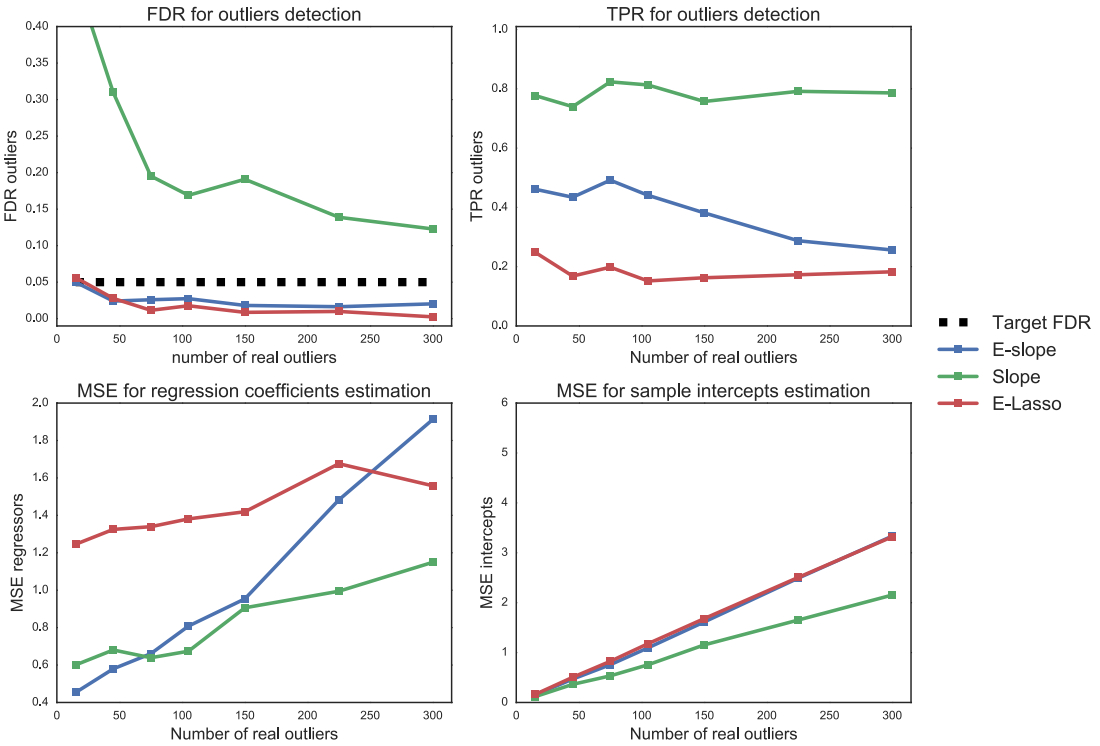

Figure II.3: Results for Simulation Setting 2 with low-magnitude outliers. The first row gives the FDR (left) and power (right) of each considered procedure for outliers discoveries. The second row gives the MSE for regressors (left) and intercepts (right). Once again E-SLOPE provides the best power among procedure that keep the FDR below the target level, and is competitive for estimating regressor coefficients. All procedures have a poor MSE when the number of outliers is large, since the simulation setting considered in this experiment is hard: low-magnitude outliers and high-dimension.

5.6 Retail Sales Data

This dataset is from the U.S. census Bureau, for year 1992. The informations contained in it are the per capita retail sales of 845 US counties (in \$1000s). It also contains five covariates: the per capita retail establishments, the per capita income (in \$1000s), per capita federal expenditures (in \$1000s), and the number of males per 100 females. No outliers are known, so we artificially create outliers by adding a small amount (magnitude 8, random sign) to the retail sales of counties chosen uniformly at random. We consider various scenarii (from 1% to 20% of outliers) and compute the false discovery proportion and the power. Figure II.5 below summarizes the results

for the three procedures.

The results are in line with the fact that E-SLOPE is able to discover more outliers than its competitors. E-SLOPE has the highest power, and the FDP remains under the target level.

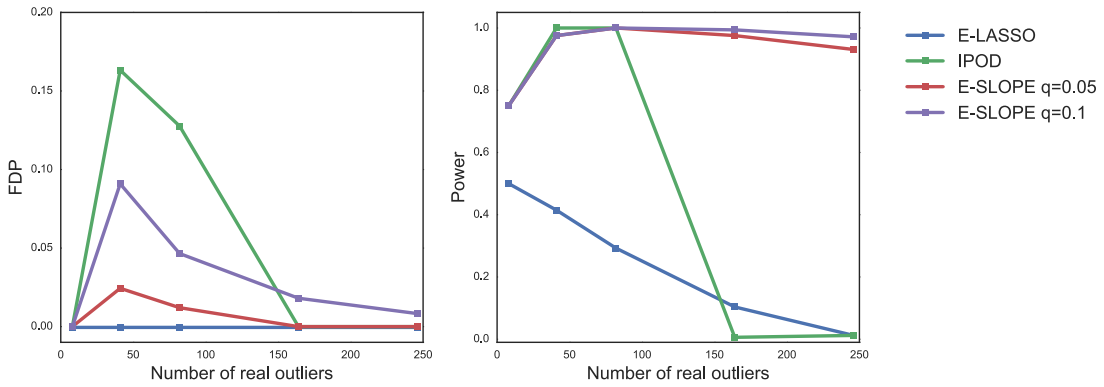

Figure II.5: *Left*: False Discovery proportion, E-SLOPE remains under the target level; *Right*: power, E-SLOPE performs better than the competitors.

5.7 Dealing with unknown variance

In all the previous simulations and real datasets analysises, the noise variance is either known or can be estimated by robust regression techniques because of the low dimension. However, dealing with unknown variance in a high-dimensional context is still an open problem. In this section we present numerical results for Setting 2 of Section 5.1 with optimization performed without the knowledge of σ. To this extent we adapt the idea of *scaled Lasso* [78] and *scaled Slope* [11] in Algorithm 4 below. The idea is to run E-Slope with a conservative estimation of σ, then to iteratively compute a new estimate of σ and run E-Slope with this new estimate. Note that there is no guarantee of convexity or consistency, or even convergence of this procedure. The stopping criteria is when the difference of two succesive estimates of σ is small enough.

Algorithm 4 Scaled E-Slope

initialize $\hat{\sigma}$
while not converged **do**
  Let $\hat{\mu}, \hat{\beta}$ E-Slope estimates for weights computed with $\hat{\sigma}$
  Set $T = \text{supp}(\hat{\mu})^{\complement}$ the subset of non-outlying points.
  Set $\hat{\sigma} = \text{MAD}(y_T - X_{T,\cdot}\hat{\beta})$
end while
return β, μ, σ

Figures below illustrate the performance of Algorithm 4 in Setting 2 (high-dimensional) described in Section 5.1. Figure II.6 illustrates the performance of Scaled E-Slope when outliers are of high magnitude and $q = 5\%$ is the target FDR level. Scaled E-Slope achieves perfect TPR, maintaining FDR below the target level. Figure II.7 illustrates the performance of Scaled E-Slope when outliers are of low magnitude and $q = 5\%$. While still controlling FDR, TPR drops when there is more than 10% outliers. Figure II.8 illustrate the performance of Scaled E-Slope when outliers are of low magnitude and $q = 5\%$. The procedure again controls FDR, TPR is maintained quite high. Note that in each of the above condition, Scaled E-Slope still outperforms E-Lasso (with known σ).

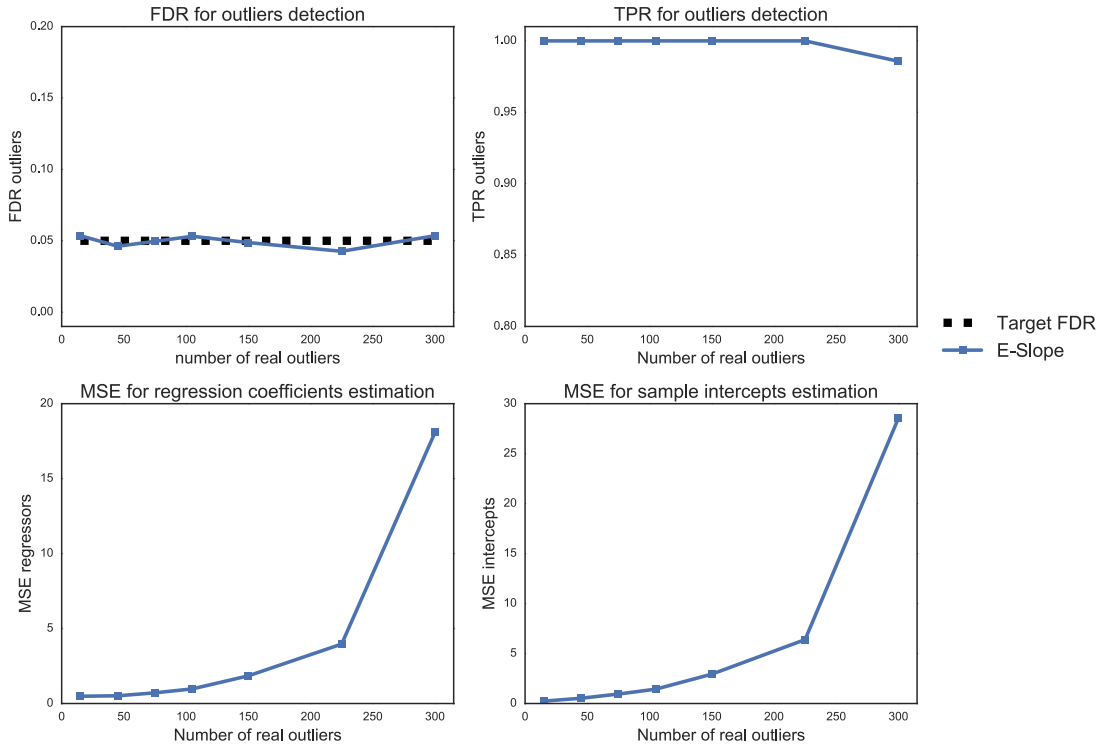

Figure II.6: Results for simulation Setting 2 with high-magnitude outliers. First row gives the FDR (left) and TPR (right) for Scaled E-SLOPE with target FDR $q = 5\%$.

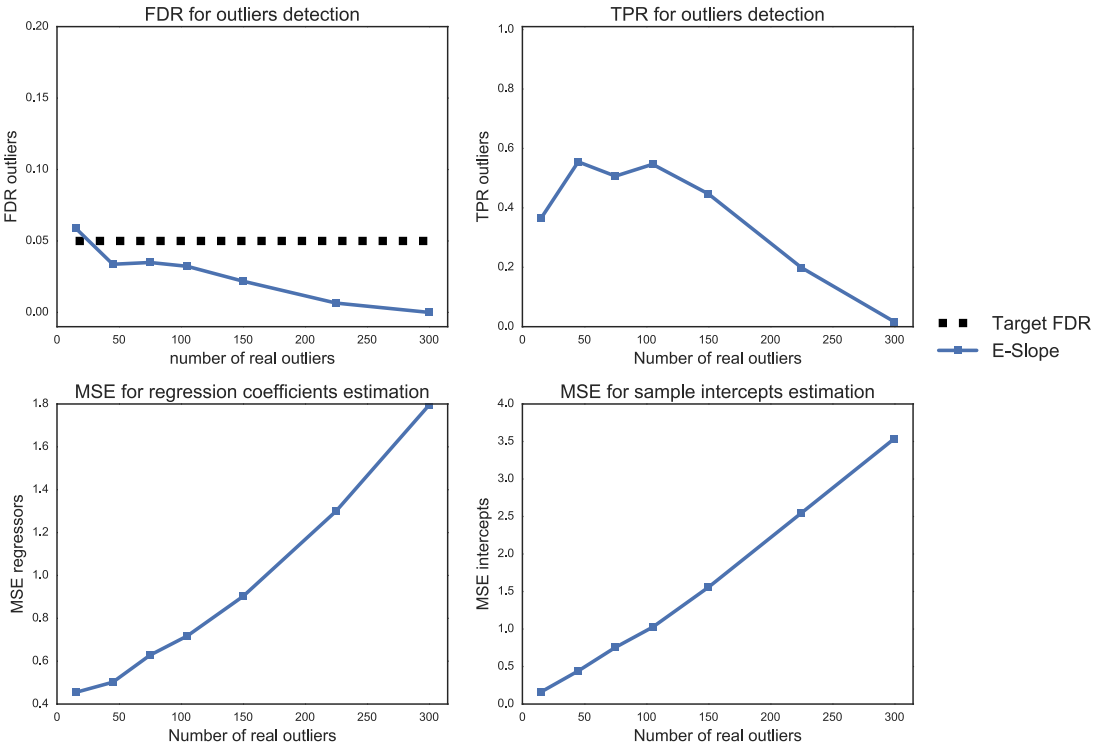

Figure II.7: Results for simulation Setting 2 with low-magnitude outliers. First row gives the FDR (left) and TPR (right) for Scaled E-SLOPE with target FDR $q = 5\%$.

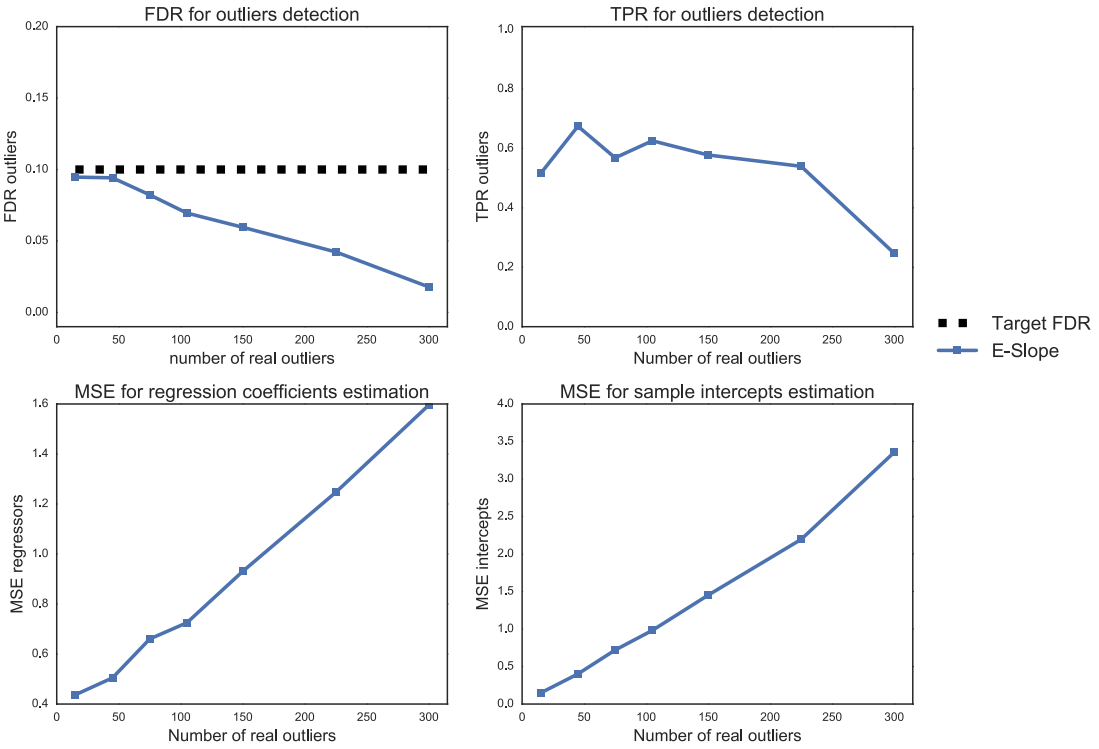

Figure II.8: Results for simulation Setting 2 with low-magnitude outliers. First row gives the FDR (left) and TPR (right) for Scaled E-SLOPE with target FDR $q = 10\%$.

6 Conclusion

In this chapter we introduce a novel approach for simultaneous robust estimation and outliers detection in the linear regression model. Three main results are provided: optimal bounds for the estimation problem in Section 3, that improve in particular previous results obtained with LASSO penalization [32], and asymptotic FDR control and power consistency for the outlier detection problem in Section 4. To the best of our knowledge, this is the first result involving FDR control in this context.

Our theoretical findings are confirmed on intensive experiments both on real and synthetic datasets, showing that our procedure outperforms existing procedure in terms of power, while maintaining a control on the FDR, even in challenging situations such as low-magnitude outliers, a high-dimensional setting and highly correlated features.

Finally, this work extends the understanding of the deep connection between the

SLOPE penalization and FDR control, previously studied in linear regression with orthogonal [11] or i.i.d gaussian [76] features, which distinguishes SLOPE from other popular convex penalization methods.

7 Technical inequalities

The following technical inequalities are borrowed from [14], where proofs can be found. Let m, n, p be positive integers. In the following lemma, an inequality for the sorted ℓ_1-norm J_λ (defined in equation II.2) is stated.

Lemma II.1. *For any two $x, y \in \mathbb{R}^m$, and any $s \in 1, \dots, m$ such that $|x|_0 \leq s$ we have*

$$J_\lambda(x) - J_\lambda(y) \leq \Lambda(s)\|x - y\|_2 - \sum_{j=s+1}^{m} \lambda_j \left|x - y\right|_{(j)},$$

where

$$\Lambda(s) = \sqrt{\sum_{j=1}^{s} \lambda_j^2}.$$

The following lemma gives a preliminary bound for the prediction error in our context, that are the starting point of our proof.

Lemma II.2. *Let $h : \mathbb{R}^p \to \mathbb{R}$ be a convex function. Consider a $n \times p$ design matrix X, a vector $\varepsilon \in \mathbb{R}^n$ and define $y = X\beta^\star + \varepsilon$ where $\beta^\star \in \mathbb{R}^p$. If $\hat{\beta}$ is a solution of the minimization problem $\min_{\beta \in \mathbb{R}^p}(\|y - X\beta\|_2^2 + 2h(\beta))$, then $\hat{\beta}$ satisfies:*

$$\|X\hat{\beta} - X\beta^\star\|_2^2 \leq \varepsilon^\top X(\hat{\beta} - \beta^\star) + h(\beta^\star) - h(\hat{\beta}).$$

Proof. Because the proof in [14] is more general, we give a proof adapted to our context. Optimality of $\hat{\beta}$ allows to choose v in the subdifferential of h sucht that

$$0 = X^\top(X\hat{\beta} - y) + v = X^\top(X\hat{\beta} - X\beta^\star - \varepsilon) + v.$$

Therefore,

$$\begin{aligned}\|X\hat{\beta} - X\beta^\star\|_2^2 &= (\hat{\beta} - \beta^\star)^\top X^\top X(\hat{\beta} - \beta^\star) \\ &= (\hat{\beta} - \beta^\star)^\top (X^\top \varepsilon - v) \\ &= \varepsilon^\top X(\hat{\beta} - \beta^\star) + \langle v, \beta^\star - \hat{\beta}\rangle.\end{aligned}$$

Now, by definition of subdifferential, $h(\beta^\star) \geq h(\hat{\beta}) + \langle v, \beta^\star - \hat{\beta}\rangle$. Combining this inequality with the previous equality leads to the conclusion. ■

The following lemma allows to bound the inner product between a white Gaussian noise and any vector. The resulting bound involved the sorted ℓ_1 norm.

Lemma II.3. *Let $\delta_0 \in (0,1)$ and let $X \in \mathbb{R}^{n\times p}$ with columns normed to 1. If ε is $\mathcal{N}(0, I_n)$ distributed, then the event*

$$\left\{\forall u \in \mathbb{R}^p, \varepsilon^\top X u \leq \max(H(u), G(u))\right\}$$

is of probability at least $1 - \delta_0/2$, where

$$H(u) = (4+\sqrt{2}) \sum_{j=1}^{p} |u|_{(j)} \sigma \sqrt{\log(2p/j)}$$

and

$$G(u) = (4+\sqrt{2})\sigma\sqrt{\log(1/\delta_0)}\|u\|_2.$$

8 Results related to Gaussian matrices

Inequalities for Gaussian random matrices are needed in this Chapter. They are stated here for the sake of clarity and we refer the reader to [29] for proofs (except for bounds II.23 and II.24 that are taken from Lemma 1 in [56]). Again, n and p denote positive integers.

Lemma II.4. *Let $X \in \mathbb{R}^{n\times p}$ with i.i.d $\mathcal{N}(0, I_p)$ rows. Denote by σ_{max} the largest singular*

value of X. Then, for all $\tau \geq 0$,

$$\mathbb{P}\left(\frac{\sigma_{max}}{\sqrt{n}} \geq 1 + \sqrt{\frac{p}{n}} + \tau\right) \leq \exp\left(-\frac{n\tau^2}{2}\right). \tag{II.20}$$

Lemma II.5. *Concentration inequalities:*

- *Let Z be $\mathcal{N}(0,1)$ distributed. Then for all $q \geq 0$:*

$$\mathbb{P}\left(|Z| \geq q\right) \leq \exp\left(-\frac{q^2}{2}\right). \tag{II.21}$$

- *Let $Z_1, Z_2, \ldots, Z_p$ be independent and $\mathcal{N}(0,\sigma^2)$ distributed. Then for all $L > 0$:*

$$\mathbb{P}\left(\max_{i=1,\ldots,p} |Z_i| > \sigma\sqrt{2\log p + 2L}\right) \leq e^{-L}. \tag{II.22}$$

- *Let X be $\chi^2(n)$ distributed. Then, for all $x > 0$:*

$$\mathbb{P}\left(X - n \geq 2\sqrt{nx} + 2x\right) \leq \exp(-x). \tag{II.23}$$

$$\mathbb{P}\left(n - X \geq 2\sqrt{nx}\right) \leq \exp(-x). \tag{II.24}$$

The following recent result ([65], Theorem 1) will also be useful.

Lemma II.6. *Let $X \in \mathbb{R}^{n\times p}$ with i.i.d $\mathcal{N}(0,\Sigma)$ rows. There exists positive constants c and c' such that with probability greater than $1 - c'\exp(-cn)$, we have for all $z \in \mathbb{R}^p$:*

$$\frac{\|Xz\|_2}{\sqrt{n}} \geq \frac{1}{4}\sqrt{\lambda_{min}(\Sigma)}\|z\|_2 - 9\sqrt{\max_j \Sigma_{jj}\frac{\log p}{n}}\|z\|_1, \tag{II.25}$$

where $\lambda_{min}(\Sigma)$ is the lowest eigenvalue of Σ.

9 Proof of Section 3

This section is devoted to the proof of our main results, stated in Section 3.

9.1 Proof of Theorem II.1

Define D the diagonal matrix such that $X = X'D$ (D is the diagonal matrix formed by the inverse of the norm of each column of X'). Applying now Lemma II.6 for X' and Dz we obtain for all $z \in \mathbb{R}^p$

$$\begin{aligned}\|Xz\|_2 &\geq \frac{1}{4}\sqrt{\lambda_{min}(\Sigma)}\|\sqrt{n}Dz\|_2 - 9\sqrt{\max_j \Sigma_{jj}\frac{\log p}{n}}\|\sqrt{n}Dz\|_1 \\ &\geq \frac{\sqrt{n}\sqrt{\lambda_{min}(\Sigma)}}{4M}\|z\|_2 - 9\frac{\sqrt{\max_j \Sigma_{jj}\log p}}{m}\|z\|_1,\end{aligned}$$

with probability greater than $1 - c'\exp(-cn)$, where M and m denote respectively the maximum and minimum of the norms of the columns of X'. Note that for all $1 \leq i \leq p$, the squared norm of the i^{th} column of X' is $\sigma_i^2\chi_2(n)$ distributed, so using the bounds II.23 and II.24 of Lemma II.5 (respectively with $x = n$ and $x = n/16$), together with a union bound we obtain that with probability greater than $1 - ne^{-n} - ne^{-n/16}$

$$M \leq (\max_j \Sigma_{jj})\sqrt{5n}, \qquad m \geq (\min_j \Sigma_{jj})\sqrt{\frac{n}{2}},$$

and we eventually obtain

$$\|Xz\|_2 \geq \frac{\sqrt{\lambda_{min}(\Sigma)}}{4\sqrt{5}\max_j \Sigma_{jj}}\|z\|_2 - \frac{9}{\min_j \Sigma_{jj}}\sqrt{\frac{2\log p \max_j \Sigma_{jj}}{n}}\|z\|_1. \tag{II.26}$$

Let us denote $v = [\beta^\top, \mu^\top]^\top \in \mathbb{R}^{p+n} \in \mathscr{C}(k, s, c_0)$ (see Definition II.6). Then,

$$\|\beta\|_1 \leq \sum_{j=1}^{p} \frac{\tilde{\lambda}_j}{\tilde{\lambda}_p}\left|\beta\right|_{(j)} \leq (1 + c_0)\left(\sqrt{k}\|\beta\|_2 + \sqrt{s}\|\mu\|_2\right). \tag{II.27}$$

Thus we obtain

$$\|\beta\|_1 \leq (1 + c_0)\left(\sqrt{k}\|\beta\|_2 + \sqrt{s}\|\mu\|_2\right). \tag{II.28}$$

Injecting (II.28) in (II.26) applied to the vector β now leads to

$$\|X\beta\|_2 + \|\mu\|_2 \geq \|\beta\|_2\Big(\frac{\sqrt{\lambda_{min}(\Sigma)}}{4\sqrt{5}\max_j \Sigma_{jj}} - \frac{9}{\min_j \Sigma_{jj}}(1+c_0)\sqrt{\frac{2k(\log p)\max_j \Sigma_{jj}}{n}}\Big) + \|\mu\|_2\Big(1 - \frac{9}{\min_j \Sigma_{jj}}(1+c_0)\sqrt{\frac{2s(\log p)\max_j \Sigma_{jj}}{n}}\Big). \quad \text{(II.29)}$$

For n large enough as explicited in the assumption of Theorem II.1, Equation (II.29) turns to

$$\|X\beta\|_2 + \|\mu\|_2 \geq \frac{\sqrt{\lambda_{min}(\Sigma)}}{8\sqrt{5}\max_j \Sigma_{jj}}\|\beta\|_2 + \frac{1}{2}\|\mu\|_2,$$

and thus, using the fact that $2(a^2+b^2) \geq (a+b)^2$,

$$\|X\beta\|_2^2 + \|\mu\|_2^2 \geq \min\left\{\frac{\lambda_{min}(\Sigma)}{128\times 5(\max_j \Sigma_{jj})^2}, \frac{1}{8}\right\}\|v\|_2^2. \quad \text{(II.30)}$$

Now if $v = [\beta^\top, \mu^\top]^\top \in \mathbb{R}^{p+n} \in \mathscr{C}^p(s, c_0)$, Equation (II.26) together with the inequality $\|\beta\|_1 \leq \sqrt{p}\|\beta\|_2$ lead to

$$\|X\beta\|_2 + \|\mu\|_2 \geq \|\beta\|_2\Big(\frac{\sqrt{\lambda_{min}(\Sigma)}}{4\sqrt{5}\max_j \Sigma_{jj}} - \frac{9}{\min_j \Sigma_{jj}}\sqrt{\frac{2p\log p\max_j \Sigma_{jj}}{n}}\Big) + \|\mu\|_2,$$

and we conclude as above. Thus the first part of the theorem is satisfied.

Now, we must lower bound the scalar product $\langle X\beta, \mu\rangle$.
Divide $\{1,\dots,p\} = T_1 \cup T_2 \cup \cdots \cup T_t$ with T_i $(1 \leq i \leq t-1)$ of cardinality k' containing the support of the k' largest absolute values of $\beta_{\left(\cup_{j=1}^{i-1} T_j\right)^c}$ and T_t of cardinality $k'' \leq k'$ the support of the remaining values. Divide in the same way $\{1,\dots,n\} = S_1 \cup S_2 \cup \cdots \cup S_q$ (of cardinalitys $s', \dots, s', s'' \leq s'$) with respect to the largest absolute values of μ (k' and s' to be chosen later). We use this to lower bound the scalar product:

$$|\langle X\beta, \mu\rangle| = |\langle X'D\beta, \mu\rangle| \leq \sum_{i=1}^{q}\sum_{j=1}^{t} |\langle X'_{S_i,T_j}(D\beta)_{T_j}, \mu_{S_i}\rangle|,$$

so

$$\big|\langle X\beta, \mu\rangle\big| \leq \max_{i,j}\|X'_{S_i,T_j}\|_2 \frac{1}{m}\sum_{j=1}^{t}\|\beta_{T_j}\|_2 \sum_{i=1}^{q}\|\mu_{S_i}\|_2, \quad \text{(II.31)}$$

where we recall that m is the minimal value of the column norms of X'. According to Lemma II.4, conditionnally on S_i and T_j, we have with probability greater than $1-\exp(-n\tau^2/2)$,

$$\|X'_{S_i,T_j}\|_2 \leq \|\Sigma^{1/2}_{T_j,T_j}\|\big(\sqrt{k'}+\sqrt{s'}+\sqrt{s'}\tau\big) \leq \sqrt{\lambda_{max}(\Sigma)}\big(\sqrt{k'}+\sqrt{s'}+\sqrt{s'}\tau\big).$$

Considering all possibilities for S_i and T_j, we have with probability greater than $1-\binom{p}{k'}\binom{n}{s'}e^{-n\tau^2/2}$,

$$\max_{i,j}\|X'_{S_i,T_j}\|_2 \leq \sqrt{\lambda_{max}(\Sigma)}\big(\sqrt{k'}+(1+\tau)\sqrt{s'}\big). \tag{II.32}$$

Moreover, thanks to the decreasing value along the subset T_j we can use the trick of [44], writing for all $j \in \{1,\dots,t-1\}$ and all $x \in \{1,\dots,|T_{j+1}|\}$:

$$\Big|\big(\beta_{T_{j+1}}\big)_x\Big| \leq \frac{\|\beta_{T_j}\|_1}{|T_j|}.$$

Squaring this inequality and summing over x gives:

$$\|\beta_{T_{j+1}}\|_2^2 \leq \frac{\|\beta_{T_j}\|_1^2}{|T_j|}\frac{|T_{j+1}|}{|T_j|} \leq \frac{\|\beta_{T_j}\|_1^2}{|T_j|} = \frac{\|\beta_{T_j}\|_1^2}{k'}.$$

Then,

$$\sum_{j=1}^{t}\|\beta_{T_j}\|_2 \leq \|\beta\|_2 + \sum_{j=2}^{t}\|\beta_{T_j}\|_2 \leq \|\beta\|_2 + \frac{1}{\sqrt{k'}}\sum_{j=1}^{t-1}\|\beta_{T_j}\|_1 \leq \|\beta\|_2 + \frac{1}{\sqrt{k'}}\|\beta\|_1,$$

and so

$$\sum_{j=1}^{t}\|\beta_{T_j}\|_2 \leq \|\beta\|_2 + \frac{1}{\sqrt{k'}}\sum_{j=1}^{p}\frac{\tilde{\lambda}_j}{\tilde{\lambda}_p}|\beta|_{(j)}. \tag{II.33}$$

In the same way we obtain:

$$\sum_{i=1}^{q}\|\mu_{S_i}\|_2 \leq \|\mu\|_2 + \frac{1}{\sqrt{s'}}\sum_{j=1}^{n}\frac{\lambda_j}{\lambda_n}|\mu|_{(j)}. \tag{II.34}$$

Now if $v = [\beta^\top, \mu^\top]^\top \in \mathbb{R}^{p+n} \in \mathscr{C}(k, s, c_0)$,

$$\begin{aligned}\sum_{j=1}^{t} \|\beta_{T_j}\|_2 \sum_{i=1}^{q} \|\mu_{S_i}\|_2 &\leq \left(\|\beta\|_2 + \frac{1}{\sqrt{k'}}(1+c_0)(\sqrt{k}\|\beta\|_2 + \sqrt{s}\|\mu\|_2)\right) \\ &\quad \times \left(\|\mu\|_2 + \frac{1}{\sqrt{s'}}(1+c_0)(\sqrt{k}\|\beta\|_2 + \sqrt{s}\|\mu\|_2)\right) \\ &\leq \left(2\|\beta\|_2 + \|\mu\|_2\right)\left(2\|\mu\|_2 + \|\beta\|_2\right) \\ &\leq 2\|v\|_2^2 + 5\|\mu\|_2\|\beta\|_2 \\ &\leq 5\|v\|_2^2,\end{aligned}$$

where we chose $k' = s' = (1+c_0)^2(k \vee s)$. Combining this last inequality with Equations (II.31) and (II.32), and using again that $m \geq \min_j \Sigma_{jj}\sqrt{n/2}$ with probability greater than $1 - ne^{-n/16}$, lead to

$$\left|\langle X\beta, \mu\rangle\right| \leq \frac{\sqrt{\lambda_{max}(\Sigma)}}{\min_j \Sigma_{jj}}(2+\tau)\sqrt{\frac{2s'}{n}}5\|v\|_2^2. \tag{II.35}$$

Note that with this choice of s' and k', the assumptions on n and the constant C' defined in the theorem lead to $\binom{p}{k'} \leq (ep/k')^{k'} \leq \exp(n/C')$, and $\binom{n}{s'} \leq (en/s')^{s'} \leq \exp(n/C')$, so we have Equation (II.32) with probability greater than $1-\exp\left(-n\left(\tau^2/2 - 2C'^{-1}\right)\right)$. With the specific assumption on n in the statement of the theorem, the term in the right part of Equation (II.35) is small enough to obtain:

$$2\left|\langle Xb, u\rangle\right| \leq \min\left\{\frac{\lambda_{min}(\Sigma)}{256 \times 5\max_j \Sigma_{jj}}, \frac{1}{16}\right\}\|v\|_2^2 \tag{II.36}$$

Eventually, if $v = [\beta^\top, \mu^\top]^\top \in \mathbb{R}^{p+n} \in \mathscr{C}^p(s, c_0)$, Equation (II.34) still holds, and combining it with Equation (II.32) and Equation (II.31) with $t = 1$ leads to:

$$\begin{aligned}\left|\langle X\beta, \mu\rangle\right| \leq \frac{\sqrt{\lambda_{max}(\Sigma)}}{\min_j \Sigma_{jj}}(\sqrt{p} + (1+\tau)\sqrt{s'})\sqrt{\frac{2}{n}}\|\beta\|_2 \\ \times \left(\|\mu\|_2 + \frac{1}{\sqrt{s'}}(1+c_0)\left(\sqrt{p}\|\beta\|_2 + \sqrt{s}\|\mu\|_2\right)\right).\end{aligned} \tag{II.37}$$

Choosing $s' = (1+c_0)^2(p \vee s)$,

$$\begin{aligned}\left|\langle X\beta, \mu\rangle\right| &\leq \frac{\sqrt{\lambda_{max}(\Sigma)}}{\min_j \Sigma_{jj}}(2+\tau)\sqrt{\frac{2s'}{n}}\|\beta\|_2\left(2\|\mu\|_2 + \|\beta\|_2\right) \\ &\leq \frac{\sqrt{\lambda_{max}(\Sigma)}}{\min_j \Sigma_{jj}}(2+\tau)\sqrt{\frac{2s'}{n}}2\|v\|_2^2.\end{aligned}$$

We conclude as above, thus leading to the second part of the theorem.

9.2 Proof of Theorem II.2

We will actually show a slightly more general result. Let R be any subset of cardinality r containing the support of the true parameter $\mu^\star$ and I_R be the matrix obtained by extracting columns with indices in R from the identity matrix. We consider the following minimization:

$$\hat{\beta}, \hat{\mu} = \underset{\beta,\mu}{\operatorname{argmin}} \|y - X\beta - I_R\mu\|_2^2 + 2\rho J_{\lambda^{[r]}}(\mu),$$

where $\lambda^{[r]}$ contains the first r terms of the sequence of weights defined in Section 3. Obviously, the theorem will result from the case $R = \{1, \dots, n\}$. Note that $\hat{\mu}$ belongs to $\mathbb{R}^r$.

Defining $b = \hat{\beta} - \beta^\star$ and $u = I_R(\hat{\mu} - \mu_R^\star)$ where $\mu_R^\star$ denotes the vector extracted from $\mu^\star$ by selecting coordinates corresponding to indices in R (note that the eliminated coordinates are zeros), we can apply Lemma II.2 to obtain:

$$\begin{aligned}\|Xb + u\|_2^2 &\leq \varepsilon^\top(Xb+u) + \rho J_{\lambda^{[r]}}(\mu^\star) - \rho J_{\lambda^{[r]}}(\hat{\mu}) \\ &= \varepsilon^\top(Xb+u) + \rho J_\lambda(I_R\mu_R^\star) - \rho J_\lambda(I_R\hat{\mu}).\end{aligned}$$

Note that it is crucial to have $\operatorname{supp}(\mu^\star) \subset R$ in order to write $\mu^\star = I_R\mu_R^\star$. Applying now Lemma II.1 we obtain:

$$\|Xb + u\|_n^2 \leq \varepsilon^\top(Xb+u) + \rho\Big(\Lambda(s)\|u\|_2 - \sum_{j=s+1}^{n} \lambda_j |u|_{(j)}\Big), \tag{II.38}$$

where $\Lambda(s)$ is defined as $\sqrt{\sum_{j=1}^{s}\lambda_j^2}$. Hence, using Cauchy-Schwarz inequality we get:

$$\|Xb+u\|_2^2 \leq \|X^\top\varepsilon\|_2\|b\|_2 + \varepsilon^\top u + \rho\big(\Lambda(s)\|u\|_2 - \sum_{j=s+1}^{n}\lambda_j|u|_{(j)}\big).$$

Then, by Lemma II.3, with probability greater than $1-\delta_0/2$ we have (the last inequality is used for the sake of simplicity):

$$\varepsilon^\top u \leq \max(H(u),G(u)) \leq H(u)+G(u),$$

with $H(u)$ and $G(u)$ defined in Lemma II.3. Additionnally, $\frac{1}{\sigma^2}\|X^\top\varepsilon\|_2^2$ follows a χ^2 law with p degrees of freedom, so by the third point in Lemma II.5 with $x = Lp$ this provides, chosing $\delta_0 = (s/2n)^s$, that with probability greater than $1-\frac{1}{2}(s/2n)^s - \exp(-Lp)$:

$$\begin{aligned}
\|Xb+u\|_2^2 &\leq c_L\sigma\sqrt{p}\|b\|_2 + H(u) + G(u) + \rho\big(\Lambda(s)\|u\|_2 - \sum_{j=s+1}^{n}\lambda_j|u|_{(j)}\big)\\
&\leq c_L\sigma\sqrt{p}\|b\|_2 + \frac{\rho}{2}\sum_{j=1}^{n}\lambda_j|u|_{(j)}\\
&\quad + \frac{\rho}{2}\sqrt{s\log(2n/s)}\|u\|_2 + \rho\big(\Lambda(s)\|u\|_2 - \sum_{j=s+1}^{n}\lambda_j|u|_{(j)}\big)\\
&\leq c_L\sigma\sqrt{p}\|b\|_2 + \big(2\rho\Lambda(s)\|u\|_2 - \frac{\rho}{2}\sum_{j=s+1}^{n}\lambda_j|u|_{(j)}\big),
\end{aligned}$$

where $c_L = \sqrt{1+2L+2\sqrt{L}}$ and where we used Equation (II.40) to obtain the last inequality. The fact that the left part of the last inequality is positive gives:

$$\sum_{j=1}^{n}\lambda_j|u|_{(j)} \leq \sum_{j=s+1}^{n}\lambda_j|u|_{(j)} + \Lambda(s)\|u\|_2 \leq \frac{2}{\rho}c_L\sigma\sqrt{p}\|b\|_2 + 5\Lambda(s)\|u\|_2,$$

where the left part of the inequality is obtained using Cauchy-Schwarz inequality. Hence,

$$\sum_{j=1}^{n}\frac{\lambda_j}{\lambda_n}|u|_{(j)} \leq \frac{2c_L}{\rho}\sqrt{\frac{p}{\log 2}}\|b\|_2 + 5\sqrt{\frac{s\log(2en/s)}{\log 2}}\|u\|_2, \tag{II.39}$$

where we used the right part of the following inequality [14]

$$s\log\Big(\frac{2n}{s}\Big) \le \sum_{j=1}^{s} \log\Big(\frac{2n}{j}\Big) = s\log(2n) - \log(s!) \le s\log\Big(\frac{2en}{s}\Big). \tag{II.40}$$

Choosing $L = 1$ lead to $c_L = \sqrt{5}$, and reminding that $\rho \ge 2(4+\sqrt{2})$ we conclude that $[b^\top, u^\top]^\top \in \mathscr{C}^p(s_1, 4)$ (see Definition II.6) with $s_1 = \frac{s\log(2en/s)}{\log 2}$. Therefore, by Condition II.1 and the definition of κ therein :

$$\begin{aligned}
2\|Xb+u\|_2^2 &\le 2\sqrt{5}\sigma\sqrt{p}\|b\|_2 + 4\rho\Lambda(s)\|u\|_2 \\
&\le \frac{5\sigma^2}{\kappa^2}p + \kappa^2\|b\|_2^2 + \frac{4\rho^2\Lambda(s)^2}{\kappa^2} + \kappa^2\|u\|_2^2 \\
&\le \frac{4\rho^2}{\kappa^2}\Lambda(s)^2 + \frac{5\sigma^2}{\kappa^2}p + \kappa^2\|v\|_2^2 \\
&\le \frac{4\rho^2}{\kappa^2}\Lambda(s)^2 + \frac{5\sigma^2}{\kappa^2}p + \|Xb+u\|_2^2.
\end{aligned}$$

Thus,

$$\|Xb+u\|_2^2 \le \frac{4\rho^2}{\kappa^2}\Lambda(s)^2 + \frac{5\sigma^2}{\kappa^2}p,$$

and

$$\|b\|_2^2 + \|u\|_2^2 \le \frac{4\rho^2}{\kappa^4}\Lambda(s)^2 + \frac{5\sigma^2}{\kappa^4}p. \tag{II.41}$$

The proof of Theorem II.2 concludes by the inequality of Equation II.40.

9.3 Proof of Theorem II.3

As in the previous proof, the more general version still holds and in the same way we obtained (II.38), with the same definition of b and u, we now have:

$$\|Xb+u\|_2^2 \le \varepsilon^\top(Xb+u) + \nu(\|\beta^\star\|_1 - \|\hat{\beta}\|_1) + \rho(\Lambda(s)\|u\|_2 - \sum_{j=s+1}^{n} \lambda_j |u|_{(j)}).$$

With T being the support of the true regression vector $\beta^\star$ we have, using the triangle inequality:

$$\|\beta^\star\|_1 - \|\hat{\beta}\|_1 = \|\beta^\star_T\|_1 - \|b+\beta^\star\|_1 = \|\beta^\star_T\|_1 - \|b_T + \beta^\star_T\|_1 - \|b_{T^c}\|_1 \le \|b_T\|_1 - \|b_{T^c}\|_1.$$

Hence we can write:

$$\begin{aligned}\|Xb+u\|_2^2 &\leq \|X^\top\varepsilon\|_\infty\|b\|_1 + \nu\left(\|b_T\|_1 - \|b_{T^c}\|_1\right) + \varepsilon^\top u \\ &\quad + \rho\Lambda(s)\|u\|_2 - \rho\sum_{j=s+1}^{n}\lambda_j|u|_{(j)} \\ &\leq \|b_T\|_1(\nu + \|X^\top\varepsilon\|_\infty) - \|b_{T^c}\|_1(\nu - \|X^\top\varepsilon\|_\infty) + \varepsilon^\top u \\ &\quad + \rho\Lambda(s)\|u\|_2 - \rho\sum_{j=s+1}^{n}\lambda_j|u|_{(j)}.\end{aligned}$$

With the choice $\nu = 4\sigma\sqrt{\log p}$ we have $\|X^\top\varepsilon\|_\infty \leq \nu/2$ according to Lemma II.5, with probability greater than $1-\frac{1}{p}$. Using again Lemma II.3 to bound $\varepsilon^\top u$, we obtain that with probability greater than $1-\frac{1}{2}\left(\frac{s}{2n}\right)^s - \frac{1}{p}$:

$$\|Xb+u\|_2^2 \leq \|b_T\|_1(6\sigma\sqrt{\log p}) - \|b_{T^c}\|_1(2\sigma\sqrt{\log p}) + 2\rho\Lambda(s)\|u\|_2 - \frac{\rho}{2}\sum_{j=s+1}^{n}\lambda_j|u|_{(j)}. \tag{II.42}$$

The fact that the left part of the inequality is positive gives:

$$\frac{4}{\rho}\sigma\sqrt{\log p}\|b_{T^c}\|_1 + \sum_{j=s+1}^{n}\lambda_j|u|_{(j)} \leq \frac{12}{\rho}\sigma\sqrt{\log p}\|b_T\|_1 + 4\Lambda(s)\|u\|_2,$$

and using Cauchy-Schwarz inequality, this leads to:

$$\frac{4}{\rho}\sigma\sqrt{\log p}\|b\|_1 + \sum_{j=1}^{n}\lambda_j|u|_{(j)} \leq \frac{16}{\rho}\sigma\sqrt{k\log p}\|b\|_2 + 5\Lambda(s)\|u\|_2$$

Eventually we obtain, because $\lambda_n = \sigma\sqrt{\log 2}$ and $\sqrt{\log p} \geq \frac{\rho\log 2}{4}$:

$$\|b\|_1 + \sum_{j=1}^{n}\frac{\lambda_j}{\lambda_n}|u|_{(j)} \leq \frac{4\sigma\sqrt{\log p}}{\rho\lambda_n}\|b\|_1 + \sum_{j=1}^{n}\frac{\lambda_j}{\lambda_n}|u|_{(j)} \leq \frac{16\sigma\sqrt{k\log p}}{\rho\lambda_n}\|b\|_2 + \frac{5\Lambda(s)}{\lambda_n}\|u\|_2 \tag{II.43}$$

and the concatenated vector of b and u is therefore in the cone $\mathscr{C}(k_1, s_1, 4)$ with $k_1 = 16k\log p/\log 2$ and $s_1 = s\log(2en/s)/\log 2$. Starting from (II.42), we obtain,

using again κ as the capacity constant in Condition II.1:

$$\begin{aligned}
2\|Xb+u\|_2^2 &\le \|b_T\|_1 12\sigma\sqrt{\log p} + 4\rho\Lambda(s)\|u\|_2 \\
&\le 12\sigma\sqrt{k\log p}\|b\|_2 + 4\rho\Lambda(s)\|u\|_2 \\
&\le \frac{36}{\kappa^2}\sigma^2 k\log p + \kappa^2\|b\|_2^2 + \frac{4\rho^2}{\kappa^2}\Lambda(s)^2 + \kappa^2\|u\|_2^2 \\
&\le \frac{36}{\kappa^2}\sigma^2 k\log p + \frac{4\rho^2}{\kappa^2}\Lambda(s)^2 + \kappa^2\|v\|_2^2 \\
&\le \frac{36}{\kappa^2}\sigma^2 k\log p + \frac{4\rho^2}{\kappa^2}\Lambda(s)^2 + \|Xb+u\|_2^2.
\end{aligned}$$

Thus,

$$\|Xb+u\|_2^2 \le \frac{36}{\kappa^2}\sigma^2 k\log p + \frac{4\rho^2}{\kappa^2}\Lambda(s)^2$$

and using again Condition II.1 and the remark after:

$$\|b\|_2^2 + \|u\|_2^2 \le \frac{36}{\kappa^4}\sigma^2 k\log p + \frac{4}{\kappa^4}\Lambda(s)^2. \tag{II.44}$$

9.4 Proof of Theorem II.4

In the same way we obtained (II.38), we now have:

$$\|Xb+u\|_2^2 \le \varepsilon^\top(Xb+u) + \rho\big(\tilde{\Lambda}(k)\|b\|_2 - \sum_{j=k+1}^{p}\tilde{\lambda}_j|b|_{(j)}\big) + \rho\big(\Lambda(s)\|u\|_2 - \sum_{j=s+1}^{n}\lambda_j|u|_{(j)}\big)$$

We use twice Lemma II.3 to bound $\varepsilon^\top Xb$ and $\varepsilon^\top u$ with $(k/2p)^k$ and $(s/2n)^s$ as respective choices of δ_0, so that with probability $1-\frac{1}{2}\left(\frac{s}{2n}\right)^s-\frac{1}{2}\left(\frac{k}{2p}\right)^k$:

$$\begin{aligned}
\|Xb+u\|_2^2 &\le H(b)+G(b)+H(u)+G(u)\\
&\quad+\rho\big(\tilde\Lambda(k)\|b\|_2-\sum_{j=k+1}^{p}\tilde\lambda_j|b|_{(j)}\big)+\rho\big(\Lambda(s)\|u\|_2-\sum_{j=s+1}^{n}\lambda_j|u|_{(j)}\big)\\
&\le\frac{\rho}{2}\sum_{j=1}^{p}\tilde\lambda_j|b|_{(j)}+\frac{\rho}{2}\sqrt{k\log(2p/k)}\|b\|_2+\rho\big(\tilde\Lambda(k)\|b\|_2-\sum_{j=k+1}^{p}\tilde\lambda_j|b|_{(j)}\big)\\
&\quad+2\rho\Lambda(s)\|u\|_2-\frac{\rho}{2}\sum_{j=s+1}^{n}\lambda_j|u|_{(j)}\\
&\le\frac{\rho}{2}4\tilde\Lambda(k)\|b\|_2-\frac{\rho}{2}\sum_{j=k+1}^{p}\tilde\lambda_j|b|_{(j)}+2\rho\Lambda(s)\|u\|_2-\frac{\rho}{2}\sum_{j=s+1}^{n}\lambda_j|u|_{(j)},
\end{aligned}$$

where we use Equation (II.40) to obtain the last inequality. The left part of the inequality is positive so

$$\sum_{j=k+1}^{p}\tilde\lambda_j|b|_{(j)}+\sum_{j=s+1}^{n}\lambda_j|u|_{(j)}\le 4\tilde\Lambda(k)\|b\|_2+4\Lambda(s)\|u\|_2, \tag{II.45}$$

and

$$2\|Xb+u\|_2^2\le 4\rho\tilde\Lambda(k)\|b\|_2+4\rho\Lambda(s)\|u\|_2. \tag{II.46}$$

Equation (II.45) together with the Cauchy-Schwarz inequality leads to

$$\sum_{j=1}^{p}\tilde\lambda_j|b|_{(j)}+\sum_{j=1}^{n}\lambda_j|u|_{(j)}\le 5\tilde\Lambda(k)\|b\|_2+5\Lambda(s)\|u\|_2. \tag{II.47}$$

Combining the equation above with Equation (II.40) shows that the concatenated estimator is in $\mathscr{C}(k_1,s_1,4)$ with s_1 and k_1 as in the statement of the theorem (note that $\tilde\lambda_n=\ \lambda_n=\ \sigma\sqrt{\log 2}$) and so, noting κ the capacity constant of Condition II.1, Equation (II.46) leads to:

$$2\|Xb+u\|_2^2 \le (3+C)^2\frac{\tilde{\Lambda}(k)^2}{2\kappa^2}+\kappa^2\|b\|_2^2+4\rho^2\frac{\Lambda(s)^2}{\kappa^2}+\kappa^2\|u\|_2^2$$
$$\le \frac{C'}{\kappa^2}\left(\tilde{\Lambda}(k)^2+\Lambda(s)^2\right)+\|Xb+u\|_2^2,$$

where $C' = 4\rho^2 \vee (3+C)^2/2$. Finally:

$$\|Xb+u\|_2^2 \le \frac{C'}{\kappa^2}\left(\tilde{\Lambda}(k)^2+\Lambda(s)^2\right),$$

and

$$\|b\|_2^2+\|u\|_2^2 \le \frac{C'}{\kappa^4}\left(\tilde{\Lambda}(k)^2+\Lambda(s)^2\right). \tag{II.48}$$

10 Proof of Theorem II.5

In this section, we give the proof of the asymptotic FDR control presented in Theorem II.5. In the following, for a given matrix A and a given subset T, A_T denotes the extracted matrix formed by the columns of A with indices in T, whereas $A_{T,\cdot}$ denotes the extracted matrix formed by the rows of A with indices in T. For vectors, there is no ambiguity. Moreover, S (of cardinal s) denotes the support of the true parameter $\mu^\star$.

We first recall some properties on the dual of the sorted ℓ_1 norm, and also a lemma taken from [76] and stated here without proof:

Definition II.1 ([76]). *A vector $a \in \mathbb{R}^n$ is said to majorize $b \in \mathbb{R}^n$ (denoted $b \preccurlyeq a$) if they satisfy for all $i \in \{1,\dots,n\}$:*

$$|a|_{(1)}+\cdots+|a|_{(i)} \ge |b|_{(1)}+\cdots+|b|_{(i)}.$$

Proposition II.1 ([11]). *Let J_λ be the sorted ℓ_1 norm for a certain non-increasing sequence λ of length n. The unit ball of the dual norm is:*

$$\mathscr{C}_\lambda = \{v \in \mathbb{R}^n : v \preccurlyeq \lambda\}.$$

Lemma II.7 ([76], Lemma A.9). *Given any constant $\alpha > 1/(1-q)$, suppose* $\max\{\alpha s, s+d\} \le s^\star < n$ *for any (deterministic) sequence d that diverges to ∞. Let $\zeta_1, \dots, \zeta_{n-s}$ be i.i.d $\mathcal{N}(0,1)$. Then*

$$(|\zeta|_{(s^\star - s+1)}, |\zeta|_{(s^\star - s+2)}, \dots, |\zeta|_{(n-s)}) \preccurlyeq (\lambda^{\mathrm{BH}}_{s^\star+1}, \lambda^{\mathrm{BH}}_{s^\star+2}, \dots, \lambda^{\mathrm{BH}}_{n})$$

with probability approaching one.

We adapt from [76] the definition of a resolvent set below, useful to determine the true support of the mean-shift parameters.

Definition II.2. *Let $s^\star$ be an integer obeying $s < s^\star < n$. The set $S^\star(S, s^\star)$ is said to be a resolvent set if it is the union of S and of the $s^\star - s$ indices corresponding to the largest entries of the error term ε restricted to $\bar{S}$.*

Let c be any positive constant and fix $s^\star \ge s(1+c)/(1-q)$ (q being the target FDR level), so that assumptions of Lemma II.7 are satisfied. For clarity, we denote $S^\star = S^\star(S, s^\star)$. For a resolvent set $S^\star$ of cardinality $s^\star$, define the reduced minimization as:

$$\beta^{S^\star}, \mu^{S^\star} = \underset{\beta \in \mathbb{R}^p, \mu \in \mathbb{R}^{s^\star}}{\operatorname{argmin}} \left\{ \|y - X\beta - I_{S^\star}\mu\|_2^2 + 2\rho J_{\tilde{\lambda}}(\beta) + 2\rho J_{\lambda^{[s^\star]}}(\mu) \right\}, \tag{II.49}$$

where $\lambda^{[s^\star]}$ is the beginning (the first $s^\star$ terms) of the sequence of weights in the global problem. Note that a resolvent set contains the support of the true parameter $\mu^\star$, so the generalized versions of the main results in Section 3, considered in the proof in Section 9, hold.

We want to show that the estimator of the unreduced problem $\hat{\mu}$ has null values for coordinates which indices are not in $S^\star$. Precisely, we will show that $\hat{\mu} = I_{S^\star}\mu^{S^\star}$. The first order conditions for global and reduced minimisation problem above are respectively:

$$\begin{cases} X^\top(y - X\hat{\beta} - \hat{\mu}) \in \rho \partial J_{\tilde{\lambda}}(\hat{\beta}) & \text{(II.50)} \\ y - X\hat{\beta} - \hat{\mu} \in \rho \partial J_{\lambda}(\hat{\mu}) & \text{(II.51)} \end{cases}$$

and

$$\begin{cases} X^\top(y - X\beta^{S^\star} - I_{S^\star}\mu^{S^\star}) \in \rho \partial J_{\tilde{\lambda}}(\beta^{S^\star}) & \text{(II.52)} \\ I_{S^\star}^\top(y - X\beta^{S^\star} - I_{S^\star}\mu^{S^\star}) \in \rho \partial J_{\lambda^{[s^\star]}}(\mu^{S^\star}) & \text{(II.53)} \end{cases}$$

Clearly, Equation (II.52) leads to Equation (II.50) taking $\hat{\beta} = \beta^{S^\star}$ and $\hat{\mu} = I_{S^\star}\mu^{S^\star}$. We must now show that this choice of $\hat{\beta}$ and $\hat{\mu}$ satisfies Equation (II.51).

First, $y - X\hat{\beta} - \hat{\mu}$ must be in the unit ball of the dual norm, that is $y - X\hat{\beta} - \hat{\mu} \preccurlyeq \rho\lambda$. Because $y - X\hat{\beta} - \hat{\mu} \in \mathbb{R}^n$ is the concatenation of $I_{S^\star}^\top(y - X\hat{\beta} - \hat{\mu})$ and $I_{\overline{S^\star}}^\top(y - X\hat{\beta} - \hat{\mu})$, we must check that $S^\star$ satisfies:

$$I_{\overline{S^\star}}^\top(y - X\hat{\beta}^{S^\star} - I_{S^\star}\hat{\mu}^{S^\star}) \preccurlyeq \rho\lambda^{-[s^\star]},$$

where $\lambda^{-[s^\star]}$ is the end of the sequence in the global problem (omitting the first $s^\star$ terms). If so, noting that if $a_1 \preccurlyeq b_1$ and $a_2 \preccurlyeq b_2$ then $a \preccurlyeq b$ (with a and b being the respective concatenation of a_1, a_2 and b_1, b_2) and combining it with Equation (II.53) will lead to the belonging at the unit ball of the dual norm.

Equivalently, we must check that

$$y_{\overline{S^\star}} - X_{\overline{S^\star},\cdot}\beta^{S^\star} \preccurlyeq \rho\lambda^{-[s^\star]},$$

or also

$$X_{\overline{S^\star},\cdot}(\beta^\star - \beta^{S^\star}) + I_{\overline{S^\star},\cdot}\varepsilon \preccurlyeq \rho\lambda^{-[s^\star]} \tag{II.54}$$

Lemma II.7, together with the definition of the resolvent set $S^\star$ given in Definition II.2, allows us to handle the second term to obtain, with probability tending to one:

$$I_{\overline{S^\star},\cdot}\varepsilon \preccurlyeq (\lambda^{\mathrm{BH}})^{-[s^\star]} \preccurlyeq \rho(\lambda^{\mathrm{BH}})^{-[s^\star]}.$$

It remains to control the term $X_{\overline{S^\star},\cdot}(\beta^\star - \beta^{S^\star})$. For our purpose, it is sufficient to show that $\|X_{\overline{S^\star},\cdot}(\beta^\star - \beta^{S^\star})\|_\infty$ tends to zero when n goes to infinity, because in this case we would have $X_{\overline{S^\star},\cdot}(\beta^\star - \beta^{S^\star}) \preccurlyeq \rho\epsilon(\lambda^{\mathrm{BH}})^{-[s^\star]}$ if n is large enough. Thus, let $i \in \{1, \dots, n\}$ and x_i the i^{th} row of X, then we have:

$$|\langle x_i, \beta^\star - \beta^{S^\star}\rangle| \leq \sum_{j=1}^{p} |x_{i,j}||\beta^\star - \beta^{S^\star}|_j \leq \frac{M}{\sqrt{n}}\|\beta^\star - \beta^{S^\star}\|_1. \tag{II.55}$$

Now we distinguish the three cases. For Equation (II.10), we do not assume sparsity on β so we rely on the Cauchy-Schwarz inequality to obtain

$$|\langle x_i, \beta^\star - \beta^{S^\star}\rangle| \leq \frac{M}{\sqrt{n}}\sqrt{p}\|\beta^\star - \beta^{S^\star}\|_2.$$

For procedure of Equation (II.11), Equation (II.43) (with $R = S^\star$, $b = \beta^\star - \beta^{S^\star}$ and $u = \mu^\star - \mu^{S^\star}$) allows to upper-bound $\|\beta^\star - \beta^{S^\star}\|_1$ as follows:

$$\|\beta^\star - \beta^{S^\star}\|_1 \leq \frac{16\sigma\sqrt{k\log p}}{\rho\lambda_n}\|\beta^\star - \beta^{S^\star}\|_2 + \frac{5\Lambda(s)}{\lambda_n}\|\mu^\star - \mu^{S^\star}\|_2.$$

Then, using this bound in Equation (II.55) leads to the bound

$$|\langle x_i, \beta^\star - \beta^{S^\star}\rangle| \leq \frac{M}{\sqrt{n}}C\Big(\sqrt{k\log p} \vee \sqrt{s\log\Big(\frac{2en}{s}\Big)}\Big)\big(\|\beta^\star - \beta^{S^\star}\|_2 \vee \|\mu^\star - \mu^{S^\star}\|_2\big),$$

with C being some positive constant and where we recall that $\Lambda(s) \leq \sigma\sqrt{s\log\left(\frac{2en}{s}\right)}$ according to Equation (II.40).

For procedure of Equation (II.12), since

$$\|\beta^\star - \beta^{S^\star}\|_1 \leq \sum_{j=1}^{p}\frac{\tilde{\lambda}_j}{\tilde{\lambda}_n}\left|\beta^\star - \beta^{S^\star}\right|_{(j)},$$

the same arguments show that Equation (II.47) leads to the bound

$$|\langle x_i, \beta^\star - \beta^{S^\star}\rangle| \leq \frac{M}{\sqrt{n}}C'\big(\sqrt{k\log(2ep/k)} \vee \sqrt{s\log(2en/s)}\big)\big(\|\beta^\star - \beta^{S^\star}\|_2 \vee \|\mu^\star - \mu^{S^\star}\|_2\big),$$

with C' being some positive constant.

Therefore the coordinates are uniformly bounded by a quantity tending to zero in each of the three cases of the theorem, thanks to the upper bounds obtained in the proofs of Section 9, in Equations (II.41), (II.44), and (II.48). Now it is sufficient to choose n such that $\left|\langle x_i, \beta^\star - \beta^{S^\star}\rangle\right| \leq \rho\epsilon\lambda_n^{\mathrm{BH}}$ (it is important to notice that the right term does not depend on n and equals to $\rho\epsilon\Phi^{-1}(1-q/2)$) to finally obtain Equation (II.54). Note that Equation (II.54) is the necessary condition for $y - X\beta^{S^\star} - I_{S^\star}\mu^{S^\star}$ to be feasible (meaning in the unit ball $\mathscr{C}_\lambda$ of the dual norm of J_λ) but this is also sufficient for being in the subdifferential because

$$\partial J_\lambda(x) = \{\omega \in \mathscr{C}_\lambda \,:\, \langle\omega, x\rangle = J_\lambda(x)\},$$

and as we have, due to Equation (II.53):

$$\langle I_{S^\star}^\top (y - X\beta^{S^\star} - I_{S^\star}\mu^{S^\star}), \mu^{S^\star} \rangle = J_{\lambda^{[s^\star]}}(\mu^{S^\star}),$$

then:

$$\langle y - X\beta^{S^\star} - I_{S^\star}\mu^{S^\star}, I_{S^\star}\mu^{S^\star} \rangle = J_\lambda(I_{S^\star}\mu^{S^\star}).$$

Therefore, with probability tending to one, $\hat{\mu} = I_{S^*}\mu^{S^\star}$ and in particular

$$\mathrm{supp}(\hat{\mu}) \subset S^\star. \tag{II.56}$$

We now show that the support of $\hat{\mu}$ contains the support of $\mu^\star$. Considering Equation (II.53) we have in particular the belonging to the unit ball of the dual norm, that is to say:

$$I_{S^\star}^\top (y - X\beta^{S^\star} - I_{S^\star}\mu^{S^\star}) \preccurlyeq \rho\lambda^{[s^\star]}.$$

In particular we have

$$\| I_{S^\star}^\top (y - X\beta^{S^\star} - I_{S^\star}\mu^{S^\star}) \|_\infty \leq \rho\lambda_1.$$

Having $y = X\beta^\star + \mu^\star + \varepsilon = X\beta^\star + I_{S^\star}(\mu^\star)_{S^\star} + \varepsilon$, the inequality above re-writes as:

$$\| X_{S^\star,\cdot}(\beta^\star - \beta^{S^\star}) + \mu^\star_{S^\star} - \mu^{S^\star} + I_{S^\star}^\top \varepsilon \|_\infty \leq \rho\lambda_1.$$

By the triangle inequality, we obtain:

$$\| \mu^\star_{S^\star} - \mu^{S^\star} \|_\infty \leq \rho\lambda_1 + \| X_{S^\star,\cdot}(\beta^\star - \beta^{S^\star}) + I_{S^\star}^\top \varepsilon \|_\infty \leq \rho\lambda_1 + \| X_{S^\star,\cdot}(\beta^\star - \beta^{S^\star}) \|_\infty + \| I_{S^\star}^\top \varepsilon \|_\infty$$

Now, we already said that we have $\| X_{S^\star,\cdot}(\beta^\star - \beta^{S^\star}) \|_\infty \leq \rho\epsilon\lambda_n^{\mathrm{BH}} \leq \rho\epsilon\lambda_1^{\mathrm{BH}}$, and using the standard bound on the norm of a Gaussian noise (see Lemma II.5), we also have, with probability tending to one (precisely with probability $1 - 1/n$):

$$\| I_{S^\star}^\top \varepsilon \|_\infty \leq \| \varepsilon \|_\infty \leq 2\sigma\sqrt{\log n}.$$

Combining the previous inequalities leads to:

$$\| \mu^\star_{S^\star} - \mu^{S^\star} \|_\infty \leq \rho\lambda_1 + \rho\epsilon\lambda_1^{\mathrm{BH}} + 2\sigma\sqrt{\log n} = \rho(1 + 2\epsilon)\lambda_1^{\mathrm{BH}} + 2\sigma\sqrt{\log n}.$$

A standard bound for the Gaussian quantile function gives $\lambda_1^{\mathrm{BH}} \leq \sigma\sqrt{2\log(2n/q)}$, so with $q \geq 2/n$ (this is quite artificial, q is generally more than 0.01) we obtain:

$$\|(\mu^\star)_{S^\star} - \mu^{S^\star}\|_\infty \leq (1+\rho(1+2\epsilon))2\sigma\sqrt{\log n}.$$

Therefore, because the entries of $\mu^\star$ are of absolute values greater than the right bound of the above inequality we obtain:

$$S \subset \mathrm{supp}((\mu^\star)_{S^\star}) \subset \mathrm{supp}(\mu^{S^\star}) \subset \mathrm{supp}(\hat{\mu}),$$

and so the Power tends to one.

It remains to prove the FDR control, using Equation (II.56). Define the False Discovery Proportion (FDP) as $V/(R \vee 1)$, where R and V are defined in Equation (II.14). Because of the inclusion $S \subset \mathrm{supp}(\hat{\mu})$, the FDP is $(R-s)/R = 1 - s/R$ with probability tending to one. According to Equation (II.56) and the assumption on $s^\star$,

$$\mathrm{FDP} = 1 - \frac{s}{R} \leq 1 - \frac{s}{s^\star} \leq 1 - \frac{1-q}{1+c} = \frac{q+c}{1+c} \leq q + c,$$

with probability tending to one. In expectation, and with n tending to infinity, we obtain:

$$\limsup_{n\to+\infty} \mathrm{FDR}(\hat{\mu}) \leq q + c,$$

and because c is arbitrarily close to zero, it leads to the conclusion.

11 Supplementary simulations

We gather here some extra-simulations in low dimension to complete the ones from Section 5.1 with a higher FDR level or/and a higher correlation level for the design matrix. As it is the most challenging case, we focus on experiments with outliers of weak magnitudes.

Influence of the correlation in X Figure II.9 and Figure II.10 below are the same as in Section 5.1 for setting 1, excepted that the correlation in the design matrix is now lower $\rho = 0$ (resp. higher $\rho = 0.8$). Results are similar to those obtained in Section 5.1.

Note that the same conclusion arises in setting 2 (high-dimensional), which is not displayed here.

Influence of the target FDR level Figure II.11 and Figure II.12 below gather the results of simulations in setting 1 and setting 2 of Section 5.1 with target FDR being 10%, with moderate correlation ($\rho = 0.4$) in X. E-LASSO and IPOD do not depend on the target FDR but they are plotted again for the sake of comparison. The results confirm the fact that E-SLOPE provides a high TPR together with a FDR control for various target FDR level.

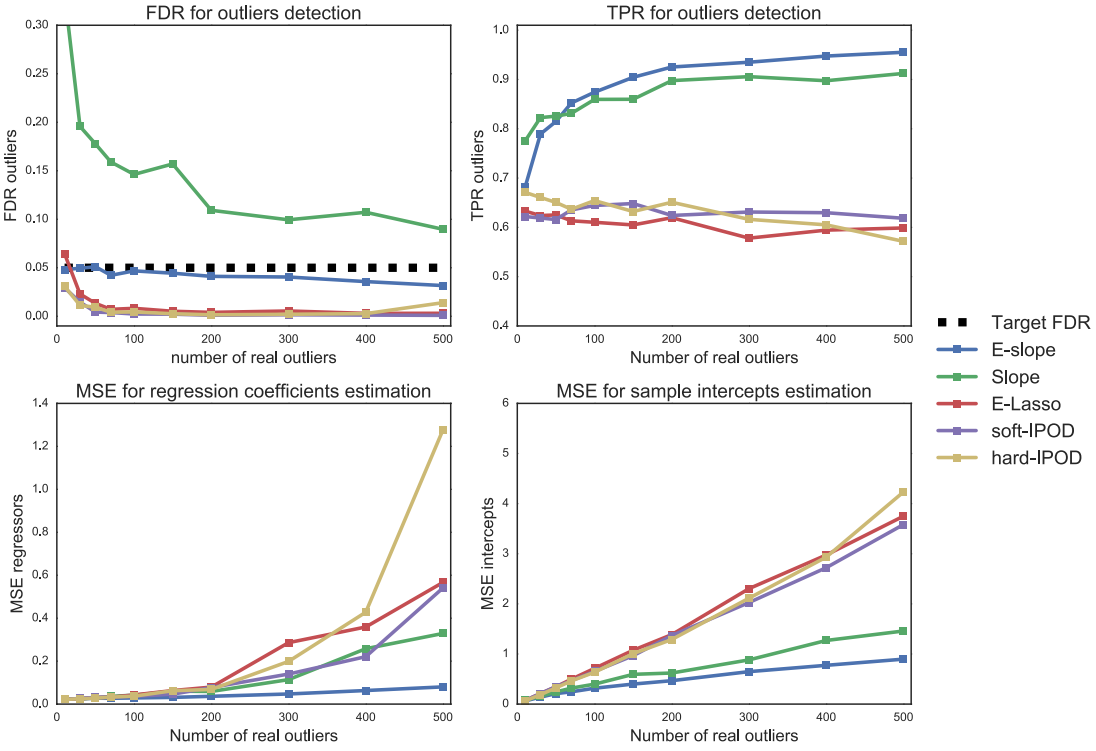

Figure II.9: Results for simulation Setting 1 with low-magnitude outliers and no correlation. The first row gives the FDR (left) and power (right) of each considered procedure for outliers discoveries. The second row gives the MSE for regressors (left) and intercepts (right). E-SLOPE provides high TPR while keeping FDR below the target level, and provides the best MSEs.

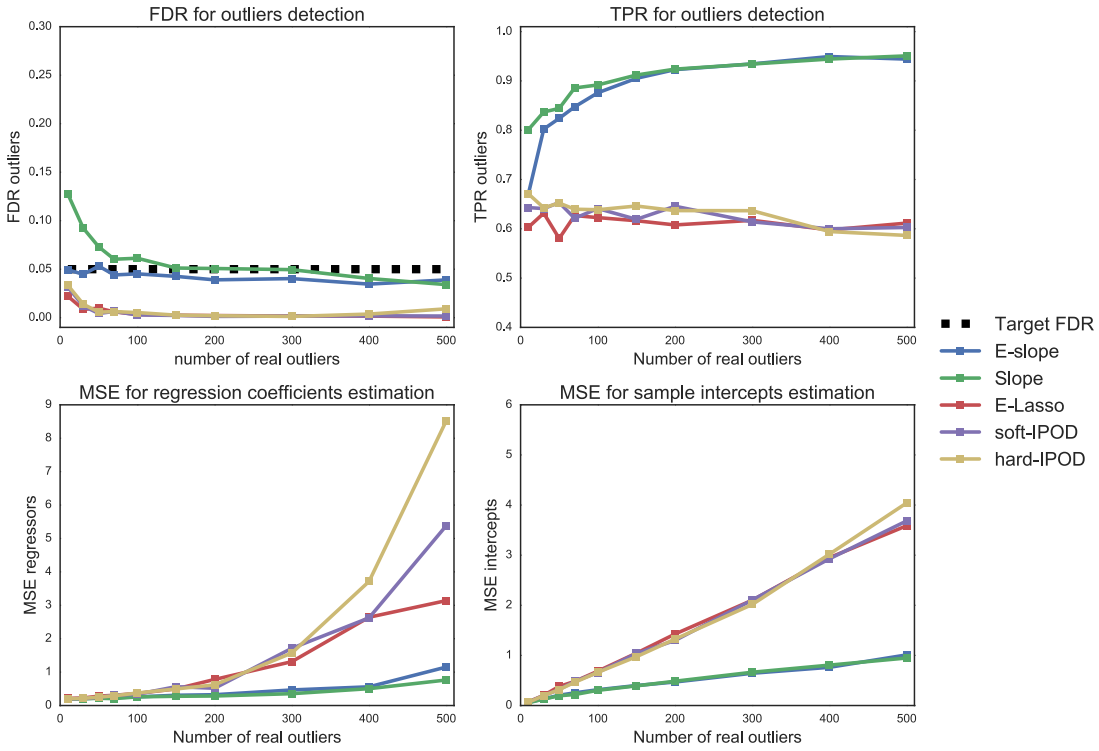

Figure II.10: Results for simulation Setting 1 with low-magnitude outliers and high correlation ($\rho = 0.8$). The first row gives the FDR (left) and power (right) of each considered procedure for outliers discoveries. The second row gives the MSE for regressors (left) and intercepts (right). E-SLOPE provides high TPR while keeping FDR below the target level, and provides the best MSEs.

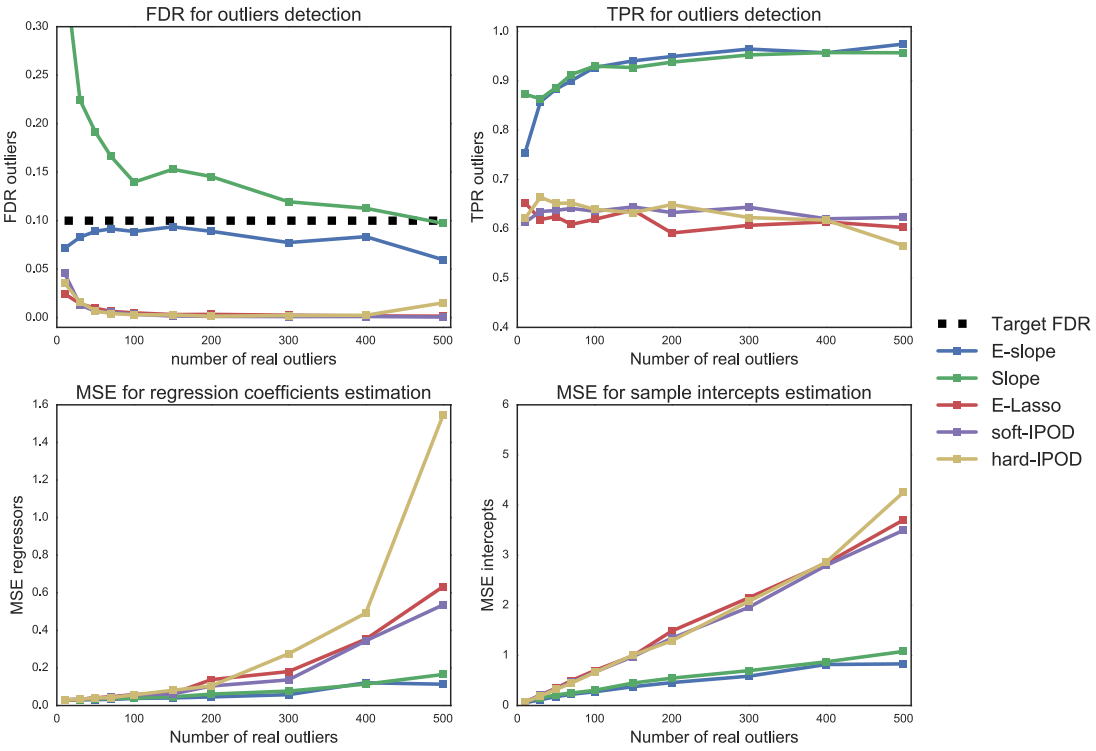

Figure II.11: Results for simulation Setting 1 with low-magnitude outliers, correlation $\rho = 0.4$. The first row gives the FDR (left) and power (right) of each considered procedure for outliers discoveries. The second row gives the MSE for regressors (left) and intercepts (right).

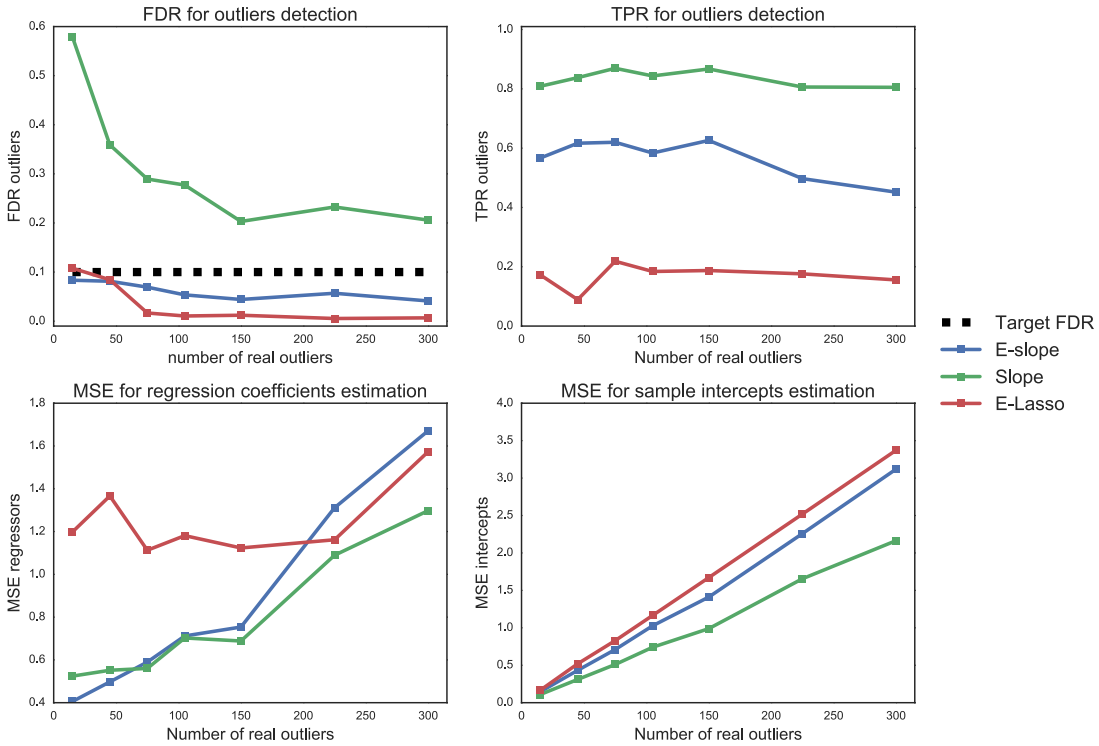

Figure II.12: Results for simulation Setting 2 with low-magnitude outliers, correlation $\rho = 0.4$. The first row gives the FDR (left) and power (right) of each considered procedure for outliers discoveries. The second row gives the MSE for regressors (left) and intercepts (right).

CHAPTER III

Extension to Generalized Linear Models

Abstract

Generalized Linear Models [61] (GLMs) are routinely used in supervised learning. The classical procedures for estimation are based on Maximum Likelihood and it is well known that the presence of outliers can have a large impact on this estimator. Robust procedures are presented in the literature on GLMs but they need a robust initial estimate in order to be computed [17, 83]. Here we study robust estimation in GLMs, when a small number k of the n observations are arbitrarily corrupted. There has been some recent work connecting robustness and sparsity in the context of linear regression with corrupted observations, by using the mean-shift outlier model (see Chapter II and [73, 32]) which explicitly models the outliers. Following these papers, we propose a procedure based on an explicit outlier response modeling in the GLM settings. We establish an asymptotic control on the False Discovery Rate (FDR) and statistical power for support selection of the individual intercepts. As a consequence, our procedure is the first proposition with guaranteed FDR and statistical power control for outliers detection under the mean-shift model in Generalized Linear Model. Numerical illustrations are provided on both simulated and real-world datasets. Experiments are conducted using an open-source software written in Python and C++.

1 Introduction

We consider Generalized Linear Models (GLMs) and we study a robust method for estimating its parameters. Robust estimators for GLMs have been intensively studied in the past few years [83, 17, 2]. However, these proposals either lack robustness or require a robust initial estimator. GLMs are a very general class of models for predicting a response given a covariate vector, and include many classical conditional distributions such as Gaussian, logistic, etc. In such models, the data points are typically low dimensional and are all assumed drawn from this model. In our setting, some observations are outliers, and could have arbitrary values with no quantitative relationship to the assumed generalized linear model.

The past few years have actually led to an understanding that outlier robust estimation is intimately connected to sparse signal recovery [47, 73]. The main insight here is that if the number of outliers is small, it could be cast as a sparse error vector that is added to the standard noise. For the task of high dimensional robust linear regression, there has been some interesting recent works that have provided bounds on the performance of the convex regularization based estimators for general high-dimensional robust estimation [32, 84] together with measures on the performance of these estimators regarding outliers detection [84].

In this paper, we provide a part of such an analysis for GLMs beyond the standard Gaussian linear model.

2 Contribution of the paper

We consider GLMs with canonical link functions and no dispersion parameter. Namely we assume that we have a response vector $y = (y_1, ..., y_n)^\top$ whose elements are observations of independent random variables $Y_1, \ldots, Y_n$ from a distribution with conditional log-density given for any $i = 1, \ldots, n$ by

$$\log f(y_i; x_i, \beta^\star, \mu_i^\star) = y_i(x_i\beta^\star + \mu_i^\star) - b(x_i\beta^\star + \mu_i^\star) + c(y_i), \tag{III.1}$$

where n is the sample size, b is a twice continuously differentiable function with derivative b' being a one-to-one function, and $\beta^\star \in \mathbb{R}^p$, $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$ respectively stand for the regression coefficients, vector of covariates, label of sample i. A non-

zero $\mu_i^\star$ means that observation i is an outlier and we assume that $\mu^\star \in \mathbb{R}^n$ is sparse with support S with $|S| = s < n$.

According to GLM theory [61] we recall that:

- $\mathbb{E}[Y_i] = b'(x_i\beta^\star + \mu_i^\star)$
- $\mathrm{Var}(Y_i) = b''(x_i\beta^\star + \mu_i^\star)$
- The link function is given by the inverse of b'.

2.1 Related works

As explained in Section 1, robust estimators for GLMs have been intensively studied in the past few years. The Cook statistic [22], used to measure the influence of an observation in linear models, can be extended to GLMs (see [61], Chapter 12). This statistic is a measure of the distance between the maximum likelihood estimator $\hat{\beta}$ and the maximum likelihood estimator computed without observation i, $\hat{\beta}_{(i)}$. However, this measure is non-robust and therefore, when there are several ouliers, it may suffer from the same masking effect [34] as in linear regression.

To provide a higher robustness, recent approaches [83] rely on a type of M-estimator, that may lead to a non-convex procedure and thus multiple solutions. To overcome this issue, one must initialize the algorithm at an initial estimator which is a very good approximation of the true parameter $\beta^\star$ [82]. However, this approach does not focus on the outliers detection but only on robust regression.

The model (III.1) above have been studied recently in [88] and allows to study both robust regression and outliers detection. They propose a convex optimization problem based on the minimization of a penalized negative log-likelihood, using Lasso as the penalization. Error bounds are obtained in [88] for both the regression and intercept coefficients, but this does not answer to the outliers detection problem we focus on in this Chapter.

2.2 Main contribution

In the same spirit as in Chapter II, the goal is to define a convex minimization problem that leads to the identification of outliers, which again are defined by the non-zero coordinates of $\mu^\star$.

We focus on low-dimensional setting (small p) and perform penalized negative log-likelihood minimization similarly to [88], using SLOPE penalization in the following optimization problem:

$$\min_{\beta\in\mathbb{R}^p,\mu\in\mathbb{R}^n} -\frac{1}{n}\sum_{i=1}^{n}\big(y_i(x_i\beta+\mu)-b(x_i\beta+\mu)\big)+J_\lambda(\mu), \tag{III.2}$$

with, for any $x\in\mathbb{R}^n$:

$$J_\lambda(x)=\sum_{j=1}^{n}\lambda_j|x|_{(j)}, \tag{III.3}$$

where $|x|_{(1)}\geq|x|_{(2)}\geq\cdots\geq|x|_{(n)}$ and $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_n>0$.

In Section 3 we provide a sequence λ which allows to obtain, under some asymptotic regime and a control of the error bound, a control of the FDR for the support selection of $\mu^\star$, and such that the power of the procedure (III.2) converges to one. In Section 4 we study the particular case of Binomial model, for which Assumption 1 and Assumption 2 given below are met. We then perform numerical experiments in Section 5 to illustrate the theoretical findings of Section 3 and Section 4, and we apply our procedure to whole exome sequencing data for colorectal cancer tumors.

2.3 Assumptions

In Section 3 we establish theoretical results under the following assumptions:

Assumption III.1. *There exists $M>0$ such that for each sample i and each covariates j $|x_{i,j}|\leq\frac{M}{\sqrt{n}}$, where $x_{i,j}$ is the value of the j-th covariates of sample i.*

The boundedness assumption on the entries of X are typically satisfied with a large probability when X has random uniform (and uniformly bounded) entries, with columns normalized to 1. Note that we could allow Gaussian distribution for the rows of X by assuming $|x_{i,j}|\leq\frac{M\log(p\vee n)}{\sqrt{n}}$. Then, we should simply add the logarithmic factor $\log(p\vee n)$ in the asymptotic assumption of Equation (III.6) for our results to remain valid. The second assumption is on the conditional distribution of labels:

Assumption III.2. *We assume that:*

- *The inverse of the link function, that is b', is L-lipschitz.*

- *For all $i \in \{1,\dots,n\}$, $y_i - \mathbb{E}[y_i]$ is c_i sub-Gaussian and the c_i's are uniformly bounded from above by a constant c, namely:*

$$\forall i \in \{1,\dots,n\}, \forall t > 0, \mathbb{P}(|y_i - \mathbb{E}[y_i]| \leq t) \geq 1 - 2\exp^{-c_i t^2} \geq 1 - 2\exp^{-ct^2}. \quad \text{(III.4)}$$

We then define for any $t > 0$, $F(t) = 1 - 2\exp^{-ct^2}$.

The assumption above are satisfied for large class of GLMs, including Linear Model and models with logit as link function (Bernoulli, Binomial, Categorical). Note in particular that bounded variables satisfy the second assumption in Assumption III.2 according to Hoeffding's inequality. Note also that some GLMs, such as Poisson regression, do not meet the assumptions above. In Section 4 we study specifically the Binomial model, which is well-suited for the real-world dataset we study in Section 5.

The assumption below is a result to be established but that we will assume true in the next sections. This result is similar to the result we proved in Theorem II.2 of Section 3, Chapter II, in the linear regression context. It is also closed to a result used in [88] in a context of GLM.

Assumption III.3. *Let R be any subset containing the true support S of $\mu^\star$ and $\lambda^{[|R|]}$ be the beginning (the first $|R|$ terms) of the sequence of weights in the global problem* (III.2).

With β^R, μ^R being solution of the following reduced minimization problem:

$$\min_{\beta \in \mathbb{R}^p, \mu \in \mathbb{R}^{|R|}} -\frac{1}{n}\sum_{i=1}^{n}\left(y_i(x_i\beta + I_R\mu) - b(x_i\beta + I_R\mu)\right) + J_{\lambda^{[|R|]}}(\mu), \quad \text{(III.5)}$$

we assume that

$$\|\beta^\star - \beta^R\|_2^2 + \|\mu^\star - \mu^R\|_2^2 = O(p\log p \vee s\log n).$$

Finally, we work under the following asymptotic setting, which is similar to the setting in which we obtained the results of Chapter II:

$$(p^2\log p)/n, (s\log n)/n \xrightarrow[n\to+\infty]{} 0. \quad \text{(III.6)}$$

This is typically satisfied when p is fixed and the number of outliers increases slower than linearly as a function of the number of observations. This asymptotic setting alows us to bound any term of the form $\langle x_i, \beta^\star - \beta^R\rangle$ by a quantity that vanishes.

Under the three assumptions above we establish in the next section the analoguous of the FDR control result obtained in the linear regression context in Theorem II.5 of Section 4, Chapter II.

3 Theoretical results

We first establish a preliminary result based on a lemma in [76] that we recall below.

Lemma III.1 ([76], Lemma A.9). *Given any constant $q > 1/(1-\alpha)$, suppose that there exists sequences $s, s^\star$ and a (deterministic) sequence d that diverges to ∞ such that*

$$\max\{qs, s+d\} \leq s^\star < n.$$

Let $U_1, U_2, \ldots, U_n$ i.i.d. uniform random variables on $[0,1]$. Note $U_{[1]} \leq U_{[2]} \leq \cdots \leq U_{[n]}$ the corresponding ordered statistics. Then:

$$\mathbb{P}\big(\forall\, j \in \{1, ..., n - s^\star\},\, U_{[s^\star - s + j]} \geq \alpha(s^\star + j)/n\big) \xrightarrow[n\to+\infty]{} 1. \tag{III.7}$$

Corollary III.1 below is a generalization of a result for Gaussian random variables in [76], Lemma A.9, to sub-Gaussian random variables. The proof of Corollary III.1 is given in Section 7.

Corollary III.1. *Assume assumptions of Lemma III.1 are met. Let $\zeta_1, \ldots, \zeta_n$ be independent sub-Gaussian random variables with the same constant c, namely the cumulative distribution functions $F_{|\zeta_i|}$ of the $|\zeta_i|$'s verify:*

$$\forall\, i \in \{1, \ldots, n\}, \forall\, t > 0, F_{|\zeta_i|}(t) \geq 1 - 2\exp^{-ct^2} = F(t). \tag{III.8}$$

We then have:

$$\mathbb{P}\Big(\forall\, j \in \{1, ..., n - s^\star\},\, |\zeta|_{(s^\star - s + j)} \leq \sqrt{\frac{1}{c}\log\frac{2n}{\alpha(s^\star + j)}}\Big) \xrightarrow[n\to+\infty]{} 1. \tag{III.9}$$

We now consider the multi-test problem with null-hypotheses

$$H_i : \mu_i^* = 0$$

for $i = 1, \dots, n$, and we consider the multi-test that rejects H_i whenever $\hat{\mu}_i \neq 0$, where $\hat{\mu}$ (and $\hat{\beta}$) are given by (III.2). When H_i is rejected, or "discovered", we consider that sample i is an outlier. Note however that in this case, the value of $\hat{\mu}_i$ gives extra information on how much sample i is outlying.

We use the FDR as a standard Type I error for this multi-test problem [8]. The FDR is the expectation of the proportion of falses discoveries among all discoveries. Letting V (resp. R) be the number of false rejections (resp. the number of rejections), the FDR is defined as

$$\mathrm{FDR}(\hat{\mu}) = \mathbb{E}\left[\frac{V}{R \vee 1}\right] = \mathbb{E}\left[\frac{\#\{i : \mu_i^\star = 0, \hat{\mu}_i \neq 0\}}{\#\{i : \hat{\mu}_i \neq 0\}}\right]. \tag{III.10}$$

We use the Power (or True Positive Rate, TPR) to measure the Type II error for this multi-test problem. The TPR is the expectation of the proportion of true discoveries. It is defined as

$$\mathrm{TPR}(\hat{\mu}) = \mathbb{E}\left[\frac{\#\{i : \mu_i^\star \neq 0, \hat{\mu}_i \neq 0\}}{\#\{i : \mu_i^\star \neq 0\}}\right]. \tag{III.11}$$

The Type II error is then given by $1 - \mathrm{TPR}(\hat{\mu})$.

Theorem III.1. *Suppose that Assumptions III.1, III.2, III.3 of Section 2.3 are met. Suppose also that non-zero coordinates of $\mu^\star$ are greater than $C\sqrt{p \log p \vee s \log n}$ and that $s \to \infty$. Consider then $\hat{\beta}, \hat{\mu}$ given by procedure* (III.2) *with $\lambda = (1+\epsilon)\lambda^0(\alpha)$ where $\epsilon > 0$ and*

$$\lambda_i^0(\alpha) = \frac{1}{n}\sqrt{\frac{1}{c}\log\frac{2n}{i\alpha}},$$

with c given by Assumption III.2, any $\alpha \in]0,1[$ and any $i \in \{1, \dots, n\}$. Then, the following properties hold:

$$\mathrm{TPR}(\hat{\mu}) \to 1, \qquad \limsup \mathrm{FDR}(\hat{\mu}) \leq \alpha. \tag{III.12}$$

Note that the magnitude of the outliers is required to grow as $\sqrt{s \log n}$ in Theorem III.1 to perform outlier detection. The assumption seems necessary in linear regression setting to distinguish outlier from pure random noise. However, there is no reason why this assumption would be necessary to control the FDR. We illustrate in numerical simulation of Section 5 that the FDR is indeed below a given level even for lower magnitudes.

The proof of Theorem III.1 is given in Section 7. Note that when $n \to +\infty$, it is also natural to assume that $s \to +\infty$ (let us recall that s stands for the sparsity of the sample outliers $\mu \in \mathbb{R}^n$). We emphasize that good numerical performances are actually obtained with lower magnitudes, as illustrated in Section 5.

4 The Binomial model

A particular model we will focus on is the binomial model, where the elements y_i of the response vector $y = (y_1, ..., y_n)^\top$ are observations of independent random variables from a Binomial distribution $\mathscr{B}(n_s, \sigma(x_i\beta^\star + \mu_i^\star))$, where n_s is a fixed integer and σ is the sigmoid function, defined for all $x \in \mathbb{R}$ as $\sigma(x) = \frac{1}{1+e^{-x}}$. These distributions have the form of Equation (III.1) with $b = -n_s \log(1-\sigma)$. Note, in addition, that $b' = n_s\sigma$ is Lipschitz.

We emphasize that Assumption III.2 is satisfied in this case with $c = \frac{2}{n_s}$ since for all $t > 0$:

$$\mathbb{P}(|y_i - E[y_i]| > t) \leq 2e^{-\frac{2t^2}{n_s}} \tag{III.13}$$

by Hoëffding deviation inequality of a sum of independent random variables [39]. In this particular case, the minimization problem rewrites as:

$$\min_{\beta\in\mathbb{R}^p, \mu\in\mathbb{R}^n} \frac{1}{n}\sum_{i=1}^{n} \big(- y_i(x_i\beta + \mu_i) - n_s \log(1 - \sigma(x_i\beta + \mu_i))\big) + J_\lambda(\mu), \tag{III.14}$$

with $\lambda_i = (1+\epsilon)\frac{1}{n}\sqrt{\frac{n_s}{2}\log\frac{2n}{i\alpha}}$.

Note that we could choose to take the proportions as observations, instead of the counts. In this situation, observations are sub-Gaussian with constant $c = 2n_s$ and that would lead to the following equivalent minimization problem:

$$\min_{\beta\in\mathbb{R}^p, \mu\in\mathbb{R}^n} \frac{1}{n}\sum_{i=1}^{n} \big(- \frac{y_i}{n_s}(x_i\beta + \mu_i) - \log(1 - \sigma(x_i\beta + \mu_i))\big) + J_\lambda(\mu), \tag{III.15}$$

with $\lambda_i = (1+\epsilon)\frac{1}{n}\sqrt{\frac{1}{2n_s}\log\frac{2n}{i\alpha}}$.

Finally, one could be more familiar with $\sqrt{n}$ normalization of the feature matrix. In this situation, again a equivalent problem would be obtained:

$$\min_{\beta\in\mathbb{R}^p,\mu\in\mathbb{R}^n} \frac{1}{n}\sum_{i=1}^{n}\Big(-\frac{y_i}{n_s}(x_i\beta+\sqrt{n}\mu_i)-\log(1-\sigma(x_i\beta+\sqrt{n}\mu_i))\Big)+J_\lambda(\mu), \tag{III.16}$$

with $\lambda_i = (1+\epsilon)\sqrt{\frac{1}{2nn_s}\log\frac{2n}{i\alpha}}$

5 Numerical experiments

In this section, we consider the binomial model given in the previous section and illustrate the performance of procedure (III.14) both on simulated and real-world datasets. Experiments are done using the open-source `tick` library [6], available at https://x-datainitiative.github.io/tick/.

5.1 Simulation settings

The matrix X is simulated as a matrix with i.i.d row distributed as $\mathcal{N}(0,\Sigma)$, with Toeplitz covariance $\Sigma_{i,j} = \rho^{|i-j|}$ for $1 \le i,j \le p$, with moderate correlation $\rho = 0.4$. The columns of X are normalized to 1. We set non-zero elements of $\mu^\star$ to $\mu_i^\star = \sqrt{2\log n}$. In all reported results based on simulated datasets, the sparsity of $\mu^\star$ varies between 1% to 20%, and we display the averages of FDR, MSE and power over 100 replications.

Setting 1 (small n_s) This is the setting described above with $n = 500$, $n_s = 30$ and $p = 10$.

Setting 2 (large n_s) This is the setting described above with $n = 500$, $n_s = 100$ and $p = 10$.

5.2 Metrics

In our experiments, we report the "MSE coefficients", namely $\|\hat\beta-\beta^\star\|_2^2$ and the "MSE intercepts", namely $\|\hat\mu-\mu^\star\|_2^2$. We report also the FDR (III.10) and the TPR (III.11) to assess the procedures for the problem of outliers detection, where the expectations are approximated by averages over 100 simulations.

5.3 Results and conclusions on simulated datasets

We comment the displays provided in Figures III.1 and III.2 below.

- In each setting, Slope provides FDR control.
- The FDR seems too low, meaning that power could be increased with a more careful tuning of the weights.
- Simulations in both settings behave similarly except in term of Power, which is better in Setting 2. This is expected since observations are then the mean of much more realizations of Bernoulli random variables.

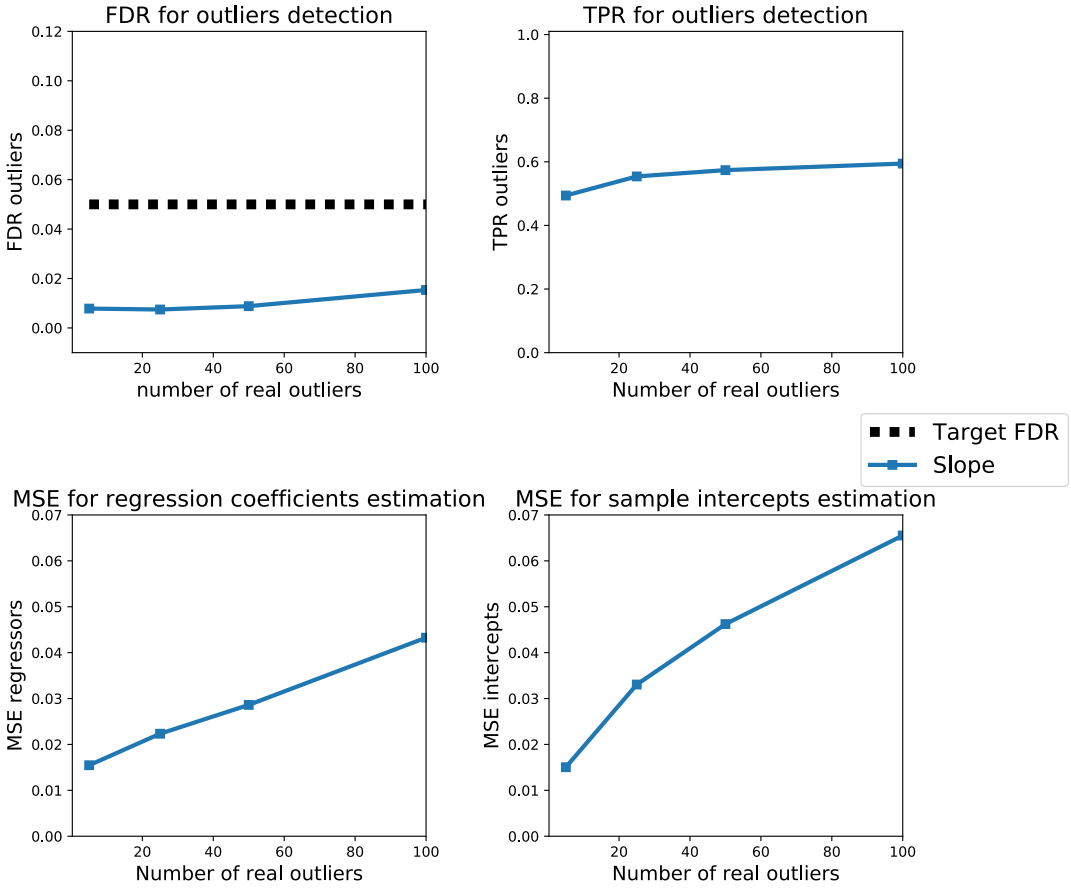

Figure III.1: Results for Simulation Setting 1. First row gives the FDR (left) and TPR (right) of our procedure for outliers discoveries. Second row gives the MSE for regressors (left) and intercepts (right).

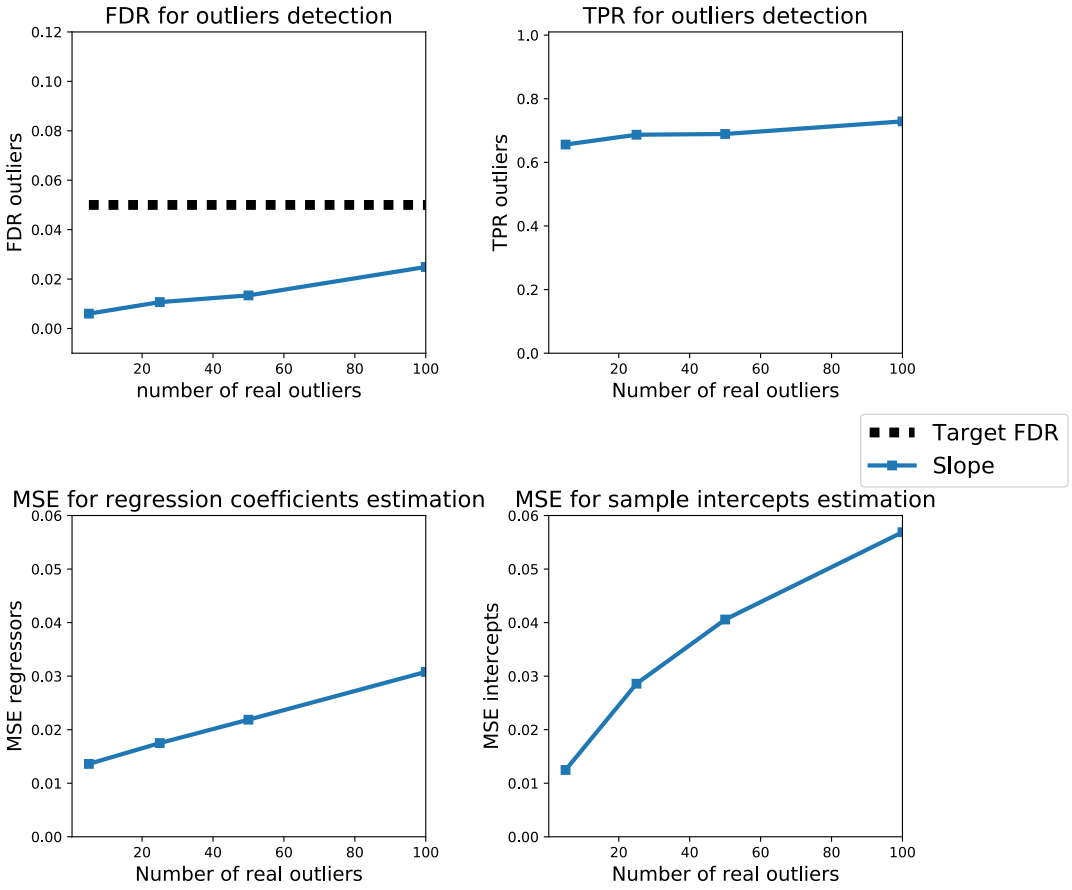

Figure III.2: Results for Simulation Setting 2. First row gives the FDR (left) and TPR (right) of our procedure for outliers discoveries. Second row gives the MSE for regressors (left) and intercepts (right).

5.4 Application to colorectal cancer tumors

5.4.1 Biological context

We consider whole exome sequencing data for 47 primary colorectal cancer tumors, characterized by a global genomic instability affecting repetitive DNA sequences (also known as microsatellite unstable tumors, see [24]).

In details, micro-satellites are portions of DNA sequence that are composed of a base motif (one or several nucleotides) repeated several times (generally 5 to 50). For example, $AAAAA$ is a micro-satellite with the base motif A (Adenine) repeated five times. Such portions of DNA have higher mutation rate than other DNA sequences,

leading to genetic diversity (instability). The colorectal cancer affects genes called *mismatch repair* (MMR) genes, responsible for the correction of transcription errors, therefore the analysis of micro-satellites is particularly relevant as their instability is directly connected to the behaviour of the MMR genes. The aim of the analysis is to find two categories of sequences: survivors (multi-satellites that mutated less than expected) and transformators (multi-satellites that mutated more than expected), with the idea that those sequences may play a key role (in good or bad) in the cancer development.

In what follows, we restrict ourselves to repetitive sequences whose base motif is the single nucleotide A, and which are in regulatory regions (following the coding regions) that influence gene expression (3' UTR). The same analysis could have been run with different base motifs and different regions (exonic, intronic). It has been shown in recent publications (see [79]), that the probability of mutation of a sequence is dependent of the length of the repeat. The figure below show the mutation rate as a function of the number of repeats of the base motif.

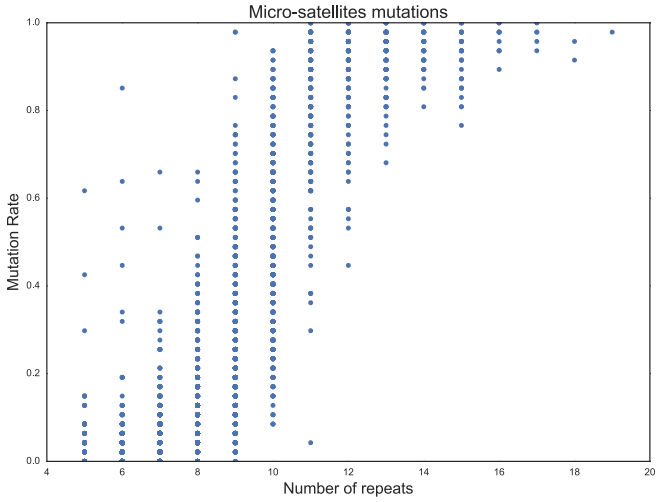

Figure III.3: Multi-satellites with base motif A: mutation rate over 47 cancer tumors plotted versus the number of repeats of the base motif.

Based on Figure III.3 it clearly appears that the linear regression is not well suited for this type of data, particularly because the output of interest is a mutation rate, which is a measure in $[0,1]$. Here the task is close to classification, where one forces

the values of the outcome variable to be bound between 0 and 1, applying a sigmoid (or logistic) function and where the bounded values are then interpreted as the probability of belonging to one of the categories in which one wants to classify data. Here we do not apply classification but a natural extension of the linear regression is to consider Generalized Linear Model [61]. As a first approach, our cancer data can be viewed as means of Bernoulli observations over 47 samples, with mutation rate (that is expectation of Bernoulli) depending on covariates including the number of base motif repeats. This is the binomial model we focused on in the previous section.

We perform procedure (III.14) on this dataset, the only feature being the number of repeats of the base motif, apart from the global intercept. Results are shown in Figure III.4. We found 152 "transformators" outliers (portions of DNA sequence that mutated more than expected) and 39 "survivors" outliers (portions of DNA sequence that mutated less than expected), which must be subjected to further biological analysis.

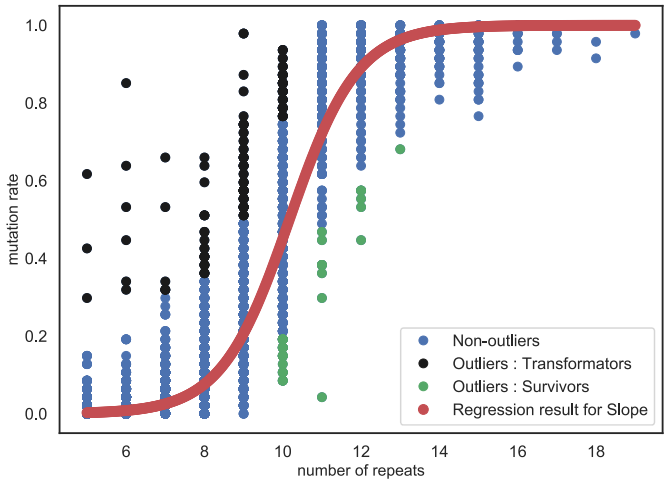

Figure III.4: Multi-satellites with base motif A: identification of 191 outliers for a target FDR $q = 0.05$.

6 Conclusion and prospects

In regression context, dealing with outliers is essential to obtain a good estimation of the regression parameters. As demonstrated in this work, the identification of outliers can also be of even greater importance in some context, because outliers can be the data of interest. In this work, we developed a new procedure to simultaneously estimate the regression coefficients and identify the outliers in particular cases of generalized linear model. The main result of this chapter is the asymptotic FDR control for the outlier detection problem. To the best of our knowledge, this is the first result involving FDR control in this context.

Our theoretical findings are confirmed on intensive experiments both on real and synthetic datasets, with an application to a crucial healthcare problem.

Finally, this work extends the understanding of the deep connection between the SLOPE penalization and FDR control, previously studied in linear regression with orthogonal [11] or i.i.d Gaussian [76] features, which distinguishes SLOPE from other popular convex penalization methods.

We conclude by noting that the theory in this chapter is still incomplete and for example does not include all kinds of GLMs. This will be the objective of future work. Moreover, the colorectal cancer dataset seems to suffer from an overdispersion, that could have influenced the results of our analysis. A multi-layers model that take into account this overdispersion (for example a *hierarchical* GLM [59]) could be of great interest. This will also be the objective of future work.

7 Proof of Section 3

7.1 Proof of Corollary III.1

In this proof, given any random variables $Z_1, \dots, Z_n$, we note the corresponding increasing (resp. decreasing) ordered statistics as $Z_{[1]} \leq Z_{[2]} \leq \cdots \leq Z_{[n]}$ (resp. $Z_{(1)} \geq Z_{(2)} \geq \cdots \geq Z_{(n)}$.

Let $F^{-1}_{|\zeta_i|}$ denote the generalized inverse distribution function of the $|\zeta_i|$'s, and $U_1, U_2, \dots, U_n$ i.i.d. uniform random variables on $[0,1]$, with the following equalities in distribution:

$$F^{-1}_{|\zeta_i|}(1 - U_i) = |\zeta_i|. \tag{III.17}$$

$$\tilde{\zeta}_i = F^{-1}(1 - U_{[i]}), \tag{III.18}$$

thus satisfying $\tilde{\zeta}_1 \geq \tilde{\zeta}_2 \geq \cdots \geq \tilde{\zeta}_n$. Note that in the following we apply Lemma 1, which requires sorted random variables. Since the $|\zeta_i|$'s are not identically distributed, their generalized inverse distribution functions are not the same and thus there is no theoretical guarantee that a permutation sorting the U_i's also sorts the $|\zeta_i|$'s. We construct the $\tilde{\zeta}_i$'s to overcome this issue.

According to Lemma III.1, we obtain, since F^{-1} is non-decreasing:

$$\mathbb{P}\big(\forall j \in \{1, ..., n - s^\star\},\, \tilde{\zeta}_{s^\star - s + j} \leq F^{-1}(1 - \alpha(s^\star + j)/n)\big) \longrightarrow 1. \tag{III.19}$$

To conclude it is now sufficient to prove that for all $i \in \{1, \ldots, n\}$, $|\zeta|_{(i)} \leq \tilde{\zeta}_i$. Let σ be a permutation such that for all $i \in \{1, \ldots, n\}$:

$$|\zeta_{\sigma(i)}| = F^{-1}_{|\zeta_{\sigma(i)}|}(1 - U_{[i]}).$$

Then Equation (III.8) guarantees that for all $i \in \{1, \ldots, n\}$,

$$|\zeta_{\sigma(i)}| = F^{-1}_{|\zeta_{\sigma(i)}|}(1 - U_{[i]}) \leq F^{-1}(1 - U_{[i]}) = \tilde{\zeta}_i.$$

Therefore by induction (and because the $\tilde{\zeta}_i$'s are sorted), for all i, $\tilde{\zeta}_i$ is greater than or equal to at least $n - i + 1$ $|\zeta_j|$'s, in particular greater than or equal to the $n - i + 1$ smallest $|\zeta_j|$'s, including $|\zeta|_{(i)}$ which is exactly the $n - i + 1$ smallest. This concludes the proof.

7.2 Proof of Theorem III.1

In the following, for a given matrix A and a given subset T, A_T denotes the extracted matrix formed by the columns of A with indices in T, whereas $A_{T,\cdot}$ denotes the extracted matrix formed by the rows of A with indices in T. For vectors, there is no ambiguity. Moreover, S (of cardinal s) denotes the support of the true parameter $\mu^\star$.

We adapt from [76] the definition of a resolvent set below, useful to determine the true support of the mean-shift parameters.

Definition III.1. *Let $s^\star$ be an integer obeying $s < s^\star < n$. The set $S^\star(S, s^\star)$ is said to be a resolvent set if it is the union of S and the $s^\star - s$ indices corresponding to the largest*

absolute values entries of the residual vector $y - \mathbb{E}[y]$ restricted to $\bar{S}$.

Let ϵ be any positive constant and fix $s^\star \geq s(1+\epsilon)/(1-q)$ (q being the target FDR level), so that assumptions of Lemma III.1 are satisfied. For clarity, we denote $S^\star = S^\star(S, s^\star)$. For a resolvent set $S^\star$ of cardinality $s^\star$, define $\beta^{S^\star}, \mu^{S^\star}$ the solution of the reduced minimization:

$$\min_{\beta \in \mathbb{R}^p, \mu \in \mathbb{R}^{s^\star}} -\frac{1}{n} \sum_{i=1}^{n} \left(y_i (x_i \beta + I_{s^\star} \mu) - b(x_i \beta + I_{s^\star} \mu) \right) + J_{\lambda^{[|s^\star|]}}(\mu), \tag{III.20}$$

where $\lambda^{[s^\star]}$ is the beginning (the first $s^\star$ terms) of the sequence of weights in the global problem (III.2).

We want to show that the estimator of the unreduced problem $\hat{\mu}$ has null values for coordinates which indices are not in $S^\star$. Precisely, we will show that $\hat{\mu} = I_{S^\star} \mu^{S^\star}$. The first order conditions for global and reduced minimisation problem above are respectively:

$$\begin{cases} X^\top (y - b'(X\hat{\beta} + \hat{\mu})) = 0 & \text{(III.21)} \\ \quad y - b'(X\hat{\beta} + \hat{\mu}) \in n \partial J_\lambda(\hat{\mu}) & \text{(III.22)} \end{cases}$$

and

$$\begin{cases} X^\top (y - b'(X\beta^{S^\star} + I_{S^\star} \mu^{S^\star})) = 0 & \text{(III.23)} \\ I_{S^\star}^\top (y - b'(X\beta^{S^\star} - I_{S^\star} \mu^{S^\star})) \in n \partial J_{\lambda^{[s^\star]}}(\mu^{S^\star}), & \text{(III.24)} \end{cases}$$

where b' is applied coordinate-wise. Clearly, Equation (III.23) leads to Equation (III.21) taking $\hat{\beta} = \beta^{S^\star}$ and $\hat{\mu} = I_{S^\star} \mu^{S^\star}$. We must now show that this choice of $\hat{\beta}$ and $\hat{\mu}$ satisfies Equation (III.22).

First, $y - b'(X\hat{\beta} + \hat{\mu})$ must be in the unit ball of the dual norm of the J_λ norm, that is $y - b'(X\hat{\beta} + \hat{\mu}) \preccurlyeq n\lambda$. Because $y - b'(X\hat{\beta} + \hat{\mu}) \in \mathbb{R}^n$ is the concatenation of $I_{S^\star}^\top (y - b'(X\hat{\beta} + \hat{\mu}))$ and $I_{\bar{S^\star}}^\top (y - b'(X\hat{\beta} + \hat{\mu}))$, we must check that $S^\star$ satisfy:

$$I_{\bar{S^\star}}^\top (y - b'(X\beta^{S^\star} + I_{S^\star} \mu^{S^\star})) \preccurlyeq n\lambda^{-[s^\star]},$$

where $\lambda^{-[s^\star]}$ is the end of the sequence in the global problem (omitting the first $s^\star$ terms). If so, noting that if $a_1 \preccurlyeq b_1$ and $a_2 \preccurlyeq b_2$ then $a \preccurlyeq b$ (with a and b being the respective concatenation of a_1, a_2 and b_1, b_2) and combining it with Equation (III.24)

will lead to the belonging to the unit ball of the dual norm.

Equivalently, we must check that

$$y_{\overline{S^\star}} - b'(X_{\overline{S^\star},\cdot}\beta^{S^\star}) \preccurlyeq n\lambda^{-[s^\star]},$$

or also

$$b'(X_{\overline{S^\star},\cdot}\beta^\star) - b'(X_{\overline{S^\star},\cdot}\beta^{S^\star}) + I_{\overline{S^\star},\cdot}z \preccurlyeq n\lambda^{-[s^\star]}, \tag{III.25}$$

where $z = y - \mathbb{E}[y] = y - b'(X\beta^\star + \mu^\star)$ is the vector of residuals. Corollary III.1, together with the definition of the resolvent set $S^\star$ given in Definition III.1, allows us to handle the second term to obtain, with probability tending to one:

$$I_{\overline{S^\star},\cdot}z \preccurlyeq n(\lambda^0)^{-[s^\star]}$$

It remains to control the term $b'(X_{\overline{S^\star},\cdot}\beta^\star) - b'(X_{\overline{S^\star},\cdot}\beta^{S^\star})$. For our purpose, it is sufficient to show that it tends to zero in $\|\|_\infty$ when n goes to infinity, because in this case we would have $b'(X_{\overline{S^\star},\cdot}\beta^\star) - b'(X_{\overline{S^\star},\cdot}\beta^{S^\star}) \preccurlyeq \epsilon(\lambda^0)^{-[s^\star]}$ if n is large enough. Thus, let $i \in \{1,\dots,n\}$ and x_i the i^{th} row of X, then we have:

$$|b'(\langle x_i, \beta^\star\rangle) - b'(\langle x_i, \beta^{S^\star}\rangle)| \leq L|\langle x_i, \beta^\star - \beta^{S^\star}\rangle| \leq \frac{LM\sqrt{p}}{\sqrt{n}}\|\beta^\star - \beta^{S^\star}\|_2. \tag{III.26}$$

When $p/n = o(1)$, the uniform bound above tends to zero under the assumptions made, that concludes the proof.

Note that Equation (III.25) is the necessary condition for $y - b'(X\beta^{S^\star} + I_{S^\star}\mu^{S^\star})$ to be feasible (meaning in the unit ball $\mathscr{C}_\lambda$ of the dual norm of J_λ) but this is also sufficient for being in the subdifferential because

$$\partial J_\lambda(x) = \{\omega \in \mathscr{C}_\lambda \,:\, \langle\omega, x\rangle = J_\lambda(x)\},$$

and as we have, due to Equation (III.24):

$$\langle I_{S^\star}^\top(y - b'(X\beta^{S^\star} + I_{S^\star}\mu^{S^\star})), \mu^{S^\star}\rangle = nJ_{\lambda^{[s^\star]}}(\mu^{S^\star}),$$

then:

$$\langle y - b'(X\beta^{S^\star} + I_{S^\star}\mu^{S^\star}), I_{S^\star}\mu^{S^\star}\rangle = J_\lambda(I_{S^\star}\mu^{S^\star}).$$

Therefore, with probability tending to one, $\hat{\mu} = I_{S^\star}\mu^{S^\star}$ and in particular

$$\mathrm{supp}(\hat{\mu}) \subset S^\star. \tag{III.27}$$

Furthermore, assuming that the non-zero coordinates of $\mu^\star$ have absolute value greater than $C(p\log p \vee s\log n)$, then the assumed estimation rate in Assumption III.3 leads to the fact that $S \subset \mathrm{supp}(\hat{\mu})$ and therefore the TPR is one.

It remains to show the FDR control. Because of the inclusion $S \subset \mathrm{supp}(\hat{\mu})$, the FDP is $(R-s)/R = 1 - s/R$ with probability tending to one. According to Equation (III.27) and the assumption on $s^\star$,

$$\mathrm{FDP} = 1 - \frac{s}{R} \leq 1 - \frac{s}{s^\star} \leq 1 - \frac{1-\alpha}{1+\epsilon} = \frac{\alpha+\epsilon}{1+\epsilon} \leq \alpha + \epsilon,$$

with probability tending to one. In expectation, and with n tending to infinity, we obtain:

$$\limsup_{n\to+\infty} \mathrm{FDR}(\hat{\mu}) \leq \alpha + \epsilon,$$

and ϵ being arbitrarily close to zero leads to the conclusion. ■

Bibliography

[1] C. C. Aggarwal and P. S. Yu. Outlier detection for high dimensional data. In *ACM Sigmod Record*, volume 30, pages 37–46. ACM, 2001.

[2] F. Alqallaf and C. Agostinelli. Robust inference in generalized linear models. *Communications in Statistics - Simulation and Computation*, 45(9):3053–3073, 2016.

[3] R. Andersen. *Modern Methods for Robust Regression.* Quantitative Applications in the Social Sciences. SAGE Publications, 2007.

[4] F. Bach, R. Jenatton, J. Mairal, G. Obozinski, et al. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012.

[5] F. Bach, R. Jenatton, J. Mairal, G. Obozinski, et al. Structured sparsity through convex optimization. *Statistical Science*, 27(4):450–468, 2012.

[6] E. Bacry, M. Bompaire, P. Deegan, S. Gaïffas, and S. V Poulsen. Tick: a python library for statistical learning, with an emphasis on hawkes processes and time-dependent models. *The Journal of Machine Learning Research*, 18(1):7937–7941, 2017.

[7] B. Bah and J. Tanner. Improved bounds on restricted isometry constants for gaussian matrices. *SIAM J. Matrix Analysis Applications*, 31(5):2882–2898, 2010.

[8] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[9] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[10] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

[11] M. Bogdan, E. van den Berg, C. Sabatti, W. Su, and E. J. Candès. Slope - adaptive variable selection via convex optimisation. *Annals of Applied Statistics*, 9(3):1103–1140, 2015.

[12] M. Bogdan, E. van den Berg, W. Su, and E. J. Candès. Statistical estimation and testing via the ordered ℓ_1 norm. *arXiv:1310.1969*, 2013.

[13] P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[14] P. C. Bellec, G. Lecué, and A. B. Tsybakov. Slope meets lasso : improved oracle bounds and optimality. *preprint ArXiv*, 2016.

[15] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

[16] E. J. Candès, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006.

[17] E. Cantoni and E. Ronchetti. Robust inference for generalized linear models. *Journal of the American Statistical Association*, 96:1022–1030, 02 2001.

[18] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational mathematics*, 12(6):805–849, 2012.

[19] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

[20] J.-T. Chiang. The masking and swamping effects using the planted mean-shift outliers models. *Int. J. Contemp. Math. Sciences*, 2(7):297–307, 2007.

[21] S. Chrétien and S. Darses. Sparse recovery with unknown variance: A lasso-type approach. *IEEE Transactions on Information Theory*, 60(7):3970–3988, 2014.

[22] R. Cook and S. Weisberg. *Residuals and influence in regression.* New York:Chapman and Hall, 1982.

[23] W. J. Dixon. Analysis of extreme values. *The Annals of Mathematical Statistics*, 21(4):488–506, 1950.

[24] A. Duval, S. Rolland, A. Compoint, E. Tubacher, B. Iacopetta, G. Thomas, and R. Hamelin. Evolution of instability at coding and non-coding repeat sequences in human msi-h colorectal cancers. *Hum Mol Genet*, 10(5):513–518, 2001.

[25] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[26] X. Gao and Y. Fang. Penalized weighted least squares for outlier detection and robust regression. ArXiv 2016.

[27] S. A. V. D. Geer and P. Bühlmann. On the conditions used to prove oracle results for the lasso. *Electron. J. Stat*, 2009.

[28] D. Gervini and V. J. Yohai. A class of robust and fully efficient regression estimators. *Annals of Statistics*, pages 583–616, 2002.

[29] C. Giraud. *Introduction to High-Dimensional Statistics.* Chapman and Hall/CRC, 2014.

[30] F. E. Grubbs. Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21, 1969.

[31] A. Gupta and S. Kohli. An mcdm approach towards handling outliers in web data: a case study using owa operators. *Artificial Intelligence Review*, 46(1):59–82, Jun 2016.

[32] N. H. Nguyen and T. D. Tran. Robust lasso with missing and grossly corrupted observations. *IEEE transactions on information theory*, 59(4):2036–2058, April 2013.

[33] A. S. Hadi. A new measure of overall potential influence in linear regression. *Computational Statistics and Data Analysis*, 14:1–27, 1992.

[34] A. S. Hadi and J. S. Simonoff. Procedures for the identification of multiple outlier in linear models. *Journal of the American Statistical Association*, 88(424):1264–1272, December 1993.

[35] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust statistics: the approach based on influence functions*, volume 114. John Wiley & Sons, 2011.

[36] D. M. Hawkins. *Identification of outliers*, volume 11. Springer, 1980.

[37] D. M. Hawkins. *Identification of outliers*. Springer, 1980.

[38] Y. Hochberg and A. C. Tamhane. *Multiple Comparison Procedures*. John Wiley & Sons, Inc., 1987.

[39] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[40] P. J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

[41] P. J. Huber. The 1972 wald lecture robust statistics: A review. *The Annals of Mathematical Statistics*, pages 1041–1067, 1972.

[42] P. J. Huber. Robust statistics. *Wiley series in probability and mathematics statistics*, pages 309–312, 1981.

[43] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, August 2009.

[44] E. J. Candès and Y. Plan. Near-ideal model selection by ℓ_1-minimization. *The Annals of Statistics*, 37(5A):2145–2177, 2009.

[45] E. J. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes rendus mathematique*, 346(9-10):589–592, 2008.

[46] E. J. Candès and P. A. Randall. Highly robust error correction by convex programming. *IEEE Transactions on Information Theory*, 54(7):2829–2840, 2008.

[47] E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[48] P. J. Huber. *Robust Statistical Procedures: Second Edition.* SIAM, 1996.

[49] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, May 2009.

[50] O. Jean Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961.

[51] T. H. Jerome Friedman and R. Tibshirani. *The elements of statistical learning.* Springer series in statistics, 2001.

[52] V. Jonchere et al. Identification of positively and negatively selected driver gene mutations associated with colorectal cancer with microsatellite instability. *Cellular and Molecular Gastroenterology and Hepatology*, 6(3):277–300, 2018.

[53] V. Koltchinskii. Saint flour lectures oracle inequalities in empirical risk minimization and sparse recovery problems. 2009. 2008.

[54] K. L. Ayers and H. J. Cordell. Snp selection in genome-wide and candidate gene studies via penalized logistic regression. *Genetic Epidemiology*, 34(8):879–891, 2010.

[55] D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47(7):2845–2862, 2001.

[56] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28(5):1302–1338, 2000.

[57] G. Lecué and M. Lerasle. Learning from mom's principles : Le cam's approach. *arXiv:1701.01961*, 2017.

[58] G. Lecué and M. Lerasle. Robust machine learning by median-of-means : theory and practice. *arXiv:1711.10306*, 2017.

[59] Y. Lee and J. A. Nelder. Hierarchical generalized linear models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(4):619–656, 1996.

[60] G. Lugosi and S. Mendelson. Risk minimization by median-of-means tournaments. *arXiv:1608.00757*, 2016.

[61] P. McCullagh and J. A. Nelder. *Generalized Linear Models, Second Edition.* CRC Press, 1989.

[62] J. N. Laska, M. A. Davenport, and R. G. Baraniuk. Exact signal recovery from sparsely corrupted measurements through the pursuit of justice. In *43rd Asilomar conference on signals, systems and computers*, pages 1556–1560, Pacific Grove, CA, November 2009. IEEE.

[63] V. N. Vapnik. *The Nature of Statistical Learning Theory.* Springer Science & Business Media, 2013.

[64] N. H. R. D. Cook and S. Weisberg. A note on an alternative outlier model. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(3):370–376, 1982.

[65] G. Raskutti, M. J. Wainwright, and B. Yu. Restricted eigenvalue properties for correlated gaussian design. *Journal of Machine Learning Research*, 11:2241–2259, Aug 2010.

[66] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57:6976–6994, 2011.

[67] K. Ro, C. Zou, Z. Wang, G. Yin, et al. Outlier detection for high-dimensional data. *Biometrika*, 102(3):589–599, 2015.

[68] E. Roquain. Type i error rate control in multiple testing: a survey with proofs. *Journal de la Société Française de Statistique*, 152(2):3–38, 2011.

[69] P. Rousseeuw and V. Yohai. Robust regression by means of s-estimators. In *Robust and nonlinear time series analysis*, pages 256–272. Springer, 1984.

[70] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection.* John Wiley & Sons, 1987.

[71] P. Rousseuww and V. Yohai. *Robust Regression by Means of S-Estimators*, chapter Robust and Nonlinear Time Series Analysis. Lecture notes in Statistics, vol 26. Springer, New York, NY, eds. Franke J., Härdle W., Martin D., 1984.

[72] J. Shao and X. Deng. Estimation in high-dimensional linear models with deterministic design matrices. *The Annals of Statistics*, 40(2):812–831, 2012.

[73] Y. She and A. B. Owen. Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association*, 2010.

[74] A. F. Siegel. Robust regression using repeated medians. *Biometrika*, 69(1):242–244, 1982.

[75] W. Su, M. Bogdan, and E. J. Candès. False discoveries occur early on the lasso path. *Annals of Statistics*, 45(5):2133–2150, 2017.

[76] W. Su and E. J. Candès. Slope is adaptive to unknown sparsity and asymptotically minimax. *Annals of Statistics*, 44(3):1038–1068, 2016.

[77] T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

[78] T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

[79] N. Suraweera, B. Iacopetta, A. Duval, A. Compoint, E. Tubacher, and R. Hamelin. Conservation of mononucleotide repeats within 3' and 5' untranslated regions and their instability in msi-h colorectal cancer. *Oncogene*, 20(51):7472–7477, 2001.

[80] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[81] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1):267–288, 1996.

[82] M. Valdora, C. Agostinelli, and V. J. Yohai. Robust estimation in high dimensional generalized linear models, 2017.

[83] M. Valdora and V. J. Yohai. Robust estimators for generalized linear models. *Journal of Statistical Planning and Inference*, 146:31 – 48, 2014.

[84] A. Virouleau, A. Guilloux, S. Gaïffas, and M. Bogdan. High-dimensional robust regression and outliers detection with slope. *arXiv:1712.02640*, 2017.

[85] H. Wang, G. Li, and G. Jiang. Robust regression shrinkage and consistent variable selection through the lad-lasso. *Journal of Business & Economic Statistics*, 25(3):347–355, Jul. 2007.

[86] T. Wang and Z. Li. Outlier detection in high-dimensional regression model. *Communications in Statistics-Theory and Methods*, 2016. Taylor & Francis.

[87] S. Weisberg. *Applied linear regression*, volume 528. John Wiley & Sons, 2005.

[88] E. Yang, A. Tewari, and P. Ravikumar. On robust estimation of high dimensional generalized linear models. In *International Joint Conference on Artificial Intelligence (IJCAI)*, volume 13, 2013.

[89] V. J. Yohai. High breakdown-point and high efficiency robust estimates for regression. *Annals of Statistics*, 15(2):642–656, 1987.

[90] S. N. M. Yoonkyung Lee and Y. Jung. Regularization of case-specific parameters for robustness and efficiency. *Statistical Science*, 27(3):350–372, 2012.

[91] C. Yu and W. Yao. Robust linear regression : a review and comparison. *Communications in Statistics - Simulation and Computation*, 2016.

[92] Y. Zhang, M. J. Wainwright, and M. I. Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*, pages 921–948, 2014.

[93] P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, (7):2541–2563, 2006.

[94] H. Zou. The adaptive lasso and its oracles properties. *Journal of the American Statistical Association*, 101(476), 2006.

INSTITUT POLYTECHNIQUE DE PARIS

École doctorale de mathématiques Hadamard (EDMH)

Titre : Apprentissage statistique pour la détection de données aberrantes et application en santé

Mots clés : tests multiples, données aberrantes, optimisation convexe, instabilité microsatellitaire

Résumé : Le problème de la détection de données aberrantes et celui de régression robuste dans un contexte de grande dimension est fondamental en statistiques et a de nombreuses applications. Dans la lignée de récents travaux proposant de traiter conjointement ces deux problèmes de régression et de détection, nous considérons dans la première partie de ce travail un modèle linéaire gaussien en grande dimension avec ajout d'un paramètre individuel pour chaque observation. Nous proposons une nouvelle procédure pour simultanément estimer les coefficients de la régression linéaire et les paramètres individuels, en utilisant deux pénalités différentes basées toutes les deux sur une pénalisation convexe ℓ_1 ordonnée, nommée SLOPE. Nous faisons l'analyse théorique de ce problème: nous obtenons dans un premier temps une borne supérieure pour l'erreur d'estimation à la fois pour le vecteur des paramètres individuels et pour le vecteur des coefficients de régression. Puis nous obtenons un résultat asymptotique sur le contrôle du taux de fausse découverte et sur la puissance concernant la détection du support du vecteur des paramètres individuels. Nous comparons numériquement notre procédure avec les alternatives les plus récentes, à la fois sur des données simulées et sur des données réelles.
La seconde partie de ce travail est motivée par un problème issu de la génétique. Des séquences particulières d'ADN, appelées *multi-satellites*, sont des indicateurs du développement d'un type de cancer colorectal. Le but est de trouver parmi ces séquences celles qui ont un taux de mutation bien plus élevé (resp. bien moindre) qu'attendu selon les biologistes. Ce problème mène à une modélisation probabiliste non-linéaire et n'entre ainsi pas dans le cadre abordé dans la première partie de cette thèse. Nous traitons ainsi dans cette partie le cas de modèles linéaires généralisés, avec de nouveau des paramètres individuels en plus du prédicteur linéaire, et analysons les propriétés statistiques d'une nouvelle procédure estimant simultanément les coefficients de régression et les paramètres individuels. Nous utilisons de nouveau la pénalisation SLOPE mais nous nous restreignons au cas de la petite dimension. La performance de l'estimateur est mesuré comme dans la première partie en terme d'erreur d'estimation des paramètres et de taux de fausse découverte concernant la recherche du support du vecteur des paramètres individuels.

Title : Machine Learning and Big Data for outliers detection, and applications

Keywords : multiple testing, outliers detection, convex optimisation, microsatellite instability

Abstract : The problems of outliers detection and robust regression in a high-dimensional setting are fundamental in statistics, and have numerous applications. Following a recent set of works providing methods for simultaneous robust regression and outliers detection, we consider in a first part a model of linear regression with individual intercepts, in a high-dimensional setting. We introduce a new procedure for simultaneous estimation of the linear regression coefficients and intercepts, using two dedicated sorted-ℓ_1 convex penalizations, also called SLOPE. We develop a complete theory for this problem: first, we provide sharp upper bounds on the statistical estimation error of both the vector of individual intercepts and regression coefficients. Second, we give an asymptotic control on the False Discovery Rate (FDR) and statistical power for support selection of the individual intercepts. Numerical illustrations, with a comparison to recent alternative approaches, are provided on both simulated and several real-world datasets.
Our second part is motivated by a genetic problem. Among some particular DNA sequences called *multi-satellites*, which are indicators of the development or colorectal cancer tumors, we want to find the sequences that have a much higher (resp. much lower) rate of mutation than expected by biologist experts. This problem leads to a non-linear probabilistic model and thus goes beyond the scope of the first part. In this second part we thus consider some generalized linear models with individual intercepts added to the linear predictor, and explore the statistical properties of a new procedure for simultaneous estimation of the regression coefficients and intercepts, using again the sorted-ℓ_1 penalization. We focus in this part only on the low-dimensional case and are again interested in the performance of our procedure in terms of statistical estimation error and FDR.

Institut Polytechnique de Paris
91120 Palaiseau, France

INSTITUT POLYTECHNIQUE DE PARIS