On the probabilistic properties of the solutions of some high-dimensional optimization problems arising in Statistics

Noureddine El Karoui
Department of Statistics, UC Berkeley*

Abstract

We study the probabilistic properties of the solutions of certain high-dimensional optimization problems arising in statistics. More specifically, if for $1 \leq i \leq n$, $X_i \in \mathbb{R}^p$ and $\epsilon_i \in \mathbb{R}$, we study the properties of

$$\widehat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho(\epsilon_i - X_i'\beta) + \frac{\tau}{2} \|\beta\|^2 \ ,$$

in the high-dimensional setting where p/n tends to a finite non-zero limit.

While most the work is done for $\tau > 0$, we show that under some extra assumptions on ρ, it is possible to recover the case $\tau = 0$ as a limiting case when $p/n < 1$. This implies that we can derive results for the unpenalized (i.e $\tau = 0$) standard "regression M-estimate" problem where ϵ_i is replaced by $Y_i = X_i'\beta_0 + \epsilon_i$, with β_0 an arbitrary deterministic vector in $\mathbb{R}^p$, characterizing in this case the behavior of $\widehat{\beta} - \beta_0$.

Our assumptions on X_i's are very general and cover for instance cases where X_i's are i.i.d with independent entries. Importantly, our proof handles the case where these entries are not Gaussian.

While our main focus is on the case of i.i.d ϵ_i's, our proof technique can also handle the case of ϵ_i's with different distributions and we give some details on this problem at the end of the paper.

1 Introduction

The last 15-20 years have seen renewed interest in statistics and machine learning for the use of convex methods in data analysis. Hence, in many applied situations, practitioners now often solve a non-trivial optimization problem to estimate or approximate a parameter or quantity of interest. A natural question is therefore to understand the probabilistic properties of the solutions of these optimization problems, if we assume for instance that the data is generated by an underlying probabilistic mechanism. We study this issue in this paper for a certain class of optimization problems which are "natural" in high-dimensional statistics and demonstrate that the tools and intuition developed in random matrix theory can be brought to bear on some of these problems. Our focus in this paper is probabilistic, but we now give some statistical background for the problems we consider.

We will focus on the basic statistical problem of regression M-estimates in high-dimension. Regression M-estimates have been of interest in statistics for at least five decades (Anscombe (1967); Relles (1968); Huber (1973)). They are natural extensions of the least-squares problem: namely one estimates a regression vector by solving the optimization problem

$$\widehat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho(Y_i - X_i'\beta) \ , \tag{1}$$

*I would like to thank Peter Bickel, Bin Yu and Derek Bean for many interesting discussions on related problems. I thank anonymous referees for comments that led to improvement of the presentation of the paper. Support from NSF grant DMS-0847647 (CAREER) is gratefully acknowledged. **AMS 2010 MSC:** Primary: 60F99; Secondary: 62E20.**Key words and Phrases :** high-dimensional optimization, random matrix theory, concentration of measure, proximal mapping, regression M-estimates. **Contact :** nkaroui@berkeley.edu

for ρ a function chosen by the user. Here, $X_i \in \mathbb{R}^p$ is a vector of (observed) predictors and $Y_i \in \mathbb{R}$ is an observed scalar response. In this paper, ρ is a convex function from $\mathbb{R}$ to $\mathbb{R}_+$. Typically once assumes that there is a linear relationship between X_i and Y_i, i.e

$$Y_i = X_i'\beta_0 + \epsilon_i \ ,$$

where ϵ_i are unobserved random errors, and β_0 is an unknown fixed vector one wishes to estimate. The $n \times p$ matrix X whose i-th row is X_i' is called the design matrix. A very natural question is to understand how close $\widehat{\beta}$ is to β_0, since $\widehat{\beta}$ is an estimate of β_0.

Huber's papers from the 1970's (Huber (1972), Huber (1973)) contain a number of very interesting results, including limiting behavior for $\widehat{\beta}$ as $n \to \infty$ when p is held fixed. Huber also raised the question of understanding the behavior of the estimators when p is large and obtained partial results in the least-squares case. Further interesting contributions happened in the mid to late 80's with work of Portnoy (Portnoy (1984), Portnoy (1985), Portnoy (1987)) and Mammen (Mammen (1989)). In these studies, the authors studied the behavior of regression M-estimates when p and n are both large, but $p/n \to 0$ at various rates. Some of the papers refer to fixed design (i.e X_i's are non-random and the only source of randomness in the problem are ϵ_i's), others treat the random design case (i.e both X_i's and ϵ_i's are random).

A central result of Huber (see e.g Huber and Ronchetti (2009)) is that when p is held fixed, and ϵ_i's are i.i.d, the optimal ρ one can use is $-\log f_\epsilon$, where f_ϵ is the density of the errors ϵ_i's - at least when we measure quality of the estimator by the size of cov $\left(\widehat{\beta}\right)$. This essentially means that maximum likelihood methods are asymptotically optimal for these problems in low-dimension. In El Karoui et al. (2011), El Karoui et al. (2013) a group of us looked at corresponding questions in the high-dimensional setting where p/n is not small and found the situation to be very different. Indeed, it was clear that one could do significantly better than using $-\log f_\epsilon$, in many metrics, including ones that would typically favor maximum likelihood methods in low dimension. More specifically, we found, in the situation where f_ϵ is a log-concave density that, at least for certain design matrices X, instead of considering $\rho = -\log(f_\epsilon)$, it was more natural to consider

$$\rho_\# = (p_2 + r_\#^2 \log \phi_{r_\#} \star f_\epsilon)^* - p_2 \ .$$

where $r_\# = \min\{r : r^2 I_\epsilon(r) = p/n\}$. Here p_2 is the function such that $p_2(x) = x^2/2$, ϕ_r is the Gaussian density with variance r^2, $f \star g$ represents the convolution of f and g, $I_\epsilon(r)$ is the Fisher information (Lehmann and Casella (1998), p. 115) of $\phi_r \star f_\epsilon$ and $g^*(x) = \sup_y(xy - g(y))$, is the Fenchel-Legendre dual of g. This convinced us that, on top of their great theoretical interest, these high-dimensional investigations could uncover unexpected insights that could be put to use, in perhaps slightly different form, in statistical practice. We also note that this result is a key motivation for the current paper. In particular, to develop a probabilistic theory that is relevant to the result just discussed, we need to be able to work with errors ϵ_i's that are log-concave and functions ρ that grow at infinity at least polynomially fast - the latter creating a number of mathematical problems. This is what we do in this paper. (A more statistical discussion explaining the setup of our work is in Appendix D-2.)

In El Karoui et al. (2011), we proposed a probabilistic heuristic - i.e a heuristic based on sound mathematical arguments - to understand the behavior of $\widehat{\beta}$ and verified the high quality of its predictions in simulations and computations. As an aside, we note that this heuristic approach was also the route taken by Huber (Huber (1973)) in his seminal work. Our heuristic was also developed because certain physics-based heuristic produced wrong predictions and we thought it would be helpful to develop other methods to guide practitioners' intuition. We believe that our methods are more reliable in the hands of mathematically-minded researchers. Our heuristic led to the formulation of a natural variational problem to optimize $\|\widehat{\beta} - \beta_0\|$, which we rigorously solved in Bean et al. (2013). Interestingly, the solution of the variational problem - i.e $\rho_\#$ mentioned above - depends in general on p/n, i.e the dimensionality of the problem and is convex. In other words, in general, "good" ρ's depend in high-dimension of the dimensionality of the problem. (El Karoui et al. (2011) is the long form of the paper El Karoui et al. (2013), which is very short due to page-limit requirements in the journal where it appeared.)

In El Karoui et al. (2011) pp. 4-5, we showed that when X_i's are i.i.d $\mathcal{N}(0,\Sigma)$ with Σ invertible, when $p < n$ and $Y_i = X_i'\beta_0 + \epsilon_i$, with $\{\epsilon_i\}_{i=1}^n$ independent of $X = \{X_i\}_{i=1}^n$, the solution of Equation (1) has a simple stochastic representation, namely

$$\widehat{\beta}-\beta_0 \stackrel{\mathcal{L}}{=} \Sigma^{-1/2}\|\widehat{\beta}(0;\mathrm{Id}_p)\|u ,$$

where u is uniform on the unit sphere in $\mathbb{R}^p$ and independent of $\|\widehat{\beta}(0;\mathrm{Id}_p)\|$, which is the norm of the solution of Equation (1) when $\beta_0 = 0$ and $\Sigma = \mathrm{Id}_p$. As we explained in that paper and detailed in Bean et al. (2013), this stochastic representation can be used to create confidence intervals for $v'\beta_0$, where v is a fixed deterministic vector of Euclidean norm 1, based on $v'\widehat{\beta}$ - i.e attach "error-bars" to this quantity. The width of the interval, and hence the accuracy of $v'\widehat{\beta}$ as an approximation of $v'\beta_0$, depends on $\|\widehat{\beta}(0;\mathrm{Id}_p)\|$ and so understanding this quantity - which is one of the foci of the current paper - is interesting and useful (the width of these confidence intervals naturally measures the quality of the corresponding statistical statements). So, even though $\|\widehat{\beta}-\beta_0\|$ does not go to zero asymptotically in the problems we consider, $\widehat{\beta}$ can be used to build confidence intervals of width of order $n^{-1/2}$ for the linear contrasts $v'\beta_0$ we just mentioned. This is, remarkably, similar to the situation in low-dimension, i.e p fixed and $n \to \infty$. We refer the reader to the supplementary material of Bean et al. (2013) for precise details. The characterization of $\|\widehat{\beta}(0;\mathrm{Id}_p)\|$ allowed us to optimize the width of these intervals over ρ in Bean et al. (2013), illustrating the need to understand well $\|\widehat{\beta}(0;\mathrm{Id}_p)\|$. Finally, $\|\widehat{\beta}(0;\mathrm{Id}_p)\|$ evidently plays a key role in measuring prediction error, another reason to study and understand it - see Appendix D-2.3 for more details. The fact that $\|\widehat{\beta}-\beta_0\|$ does not tend to 0 asymptotically is what renders the problem interesting probabilistically but has been a source of misunderstanding among some statisticians. We address these misunderstandings in Appendix D-2, since statistical issues are not the main concern of this paper.

Beside these performance issues, our investigation also pointed to an interesting behavior for the residuals, i.e $R_i = Y_i - X_i'\widehat{\beta}$, suggesting that the natural idea of "bootstrapping from the residuals" is problematic in the situation we investigate. (See also Theorem 3.1 of the current paper.) We refer to El Karoui and Purdom (2015) for more details and solutions to some of those statistical problems.

The idea of looking at asymptotics for p and n large is motivated by their probabilistic interest, naturalness, and the fact that these asymptotic results might yield better approximation in finite samples than their traditional "small p, large n" counterparts - see e.g Johnstone (2001). At a high-level, this is explained by the fact in this type of work, we need to keep track of "higher-order" quantities that are typically neglected in classical asymptotics. As we just discussed, this different perspective also sheds new light on quantities of statistical interest that up to now were thought to be well understood. We note that Huber (1973) already raised the "large p, large n" question.

Our heuristic in El Karoui et al. (2011) was based on random matrix theory, convex analysis and concentration of measure ideas. We prove in this paper that these tools can be used to obtain a rigorous understanding of various aspects of the problems of interest in great generality. The proof presented here does not simply follow from the heuristic - i.e we are not filling some "minor technical details". Rather, the heuristic provided some insights which helped us design the proof presented in the current article. Also, the problem on which we focus most of our attention here - see Equation (2) - is more general than the one studied in El Karoui et al. (2013) - the latter having no penalization. The introduction of penalization created conceptual challenges which are tackled here, while also simplifying some technical issues, resulting in quite general results.

The assumptions under which we operate for the design matrix reflect the central role played by the concentration of measure phenomenon (Ledoux (2001)) in this problem - see also our discussion on page 11 and the proof. Concentration of certain quadratic forms in X_i's is especially important here. We thought it important to do the proof at this level of generality to show the scope (or lack thereof) of potential "universality" results. This is also the level of generality that is now standard in random matrix theory. For an example where our concentration assumptions on X_i's are not satisfied and the corresponding results are completely different, see El Karoui et al. (2011) and our discussion in Subsection 6.2.

Two weeks before this paper was posted on arXiv (with a slightly different presentation), Donoho and Montanari (Donoho and Montanari (2013)), motivated by El Karoui et al. (2011), posted on arXiv a proof of some of the results explained in El Karoui et al. (2013) under the assumption that the design matrix is full of i.i.d Gaussian random variables (i.e X_i's are independent with i.i.d Gaussian entries). Their proof uses different ideas than ours - it is based on the technology of rigorous analysis of approximate message passing algorithms (see Donoho et al. (2009) and Bayati and Montanari (2012)).

By working under concentration assumptions, we are able to show many results without requiring i.i.d-ness of the entries of the vectors X_i's (see Section 3 and Assumption **O4** below) - in fact the entries could be quite dependent. However, to prove all the results of the current paper, we still need the X_i's to have i.i.d entries, but *they do not need to be Gaussian.*

Donoho and Montanari also make interesting connections with rigorous work in statistical physics, namely to the so-called Shcherbina-Tirozzi model (Shcherbina and Tirozzi (2003) and Talagrand (2003)) and other physics-based heuristic approaches based on approximate message passing (Rangan (2011)). Our approach can also be used to have a different point of view on these statistical physics models, in the zero temperature setting.

Our point of view is that the properties of $\widehat{\beta}$ defined in Equation (1) or Equation (2) below can be understood via connections to random matrix theory - which are not obvious a priori. As such, our proof relies on "leave-one-out", martingale and concentration of measure ideas, as some of our previous random matrix theoretic work (see e.g El Karoui (2009)) did. We also use quite a few tools from convex analysis, especially Moreau's proximal mapping (introduced in Moreau (1965)). Leave-one-out ideas have been prevalent in both theoretical and applied Statistics for many decades - though the double leave-one-out idea we used in El Karoui et al. (2011) was a new take on it. Leave-one-out ideas seem to be known in Physics under the name "cavity method", so our general approach falls broadly in that category. A number of the tools we use are also used in the spectral analysis of large random matrices via the Stieltjes transform method (see Marčenko and Pastur (1967), Wachter (1978), Silverstein (1995)). However, the random matrices that appear in the current paper are non-standard from the point of view of random matrix theory: they are weighted covariance matrices with weights depending from the design matrix X in a non-trivial way; the distribution of the weights is also itself a major challenge to understand. By contrast, when similar issues arise in random matrix theory, the weight distribution is typically assumed to be known and independent of the design matrix.

In the notation of the abstract and Equation (2) below, this paper gives a very detailed understanding of the properties of $\widehat{\beta}$, the residuals $R_i = \epsilon_i - X_i'\widehat{\beta}$, and several other interesting quantities.

Beside Theorems 1.1 (p.5) and 6.1 (p.42), our main results are Theorem 3.1 (p.21) and Theorem 4.1 (p. 33). Theorem 3.1 explains how to approximate $\widehat{\beta}$ to high-accuracy using a non-linear function of X_i and explains the behavior of the residuals in the high-dimensional setting (where p and n are both large and p/n is not small) under consideration. The results in Subsection 6.1 explain how to go from results concerning the "ℓ_2-penalized" problem which are at the center of this paper, i.e $\tau > 0$ to the unpenalized problem, i.e $\tau = 0$. This in turn allows us to study the non-null case, i.e $\beta_0 \neq 0$, in the unpenalized case ($\tau = 0$). In other words, we obtain some of the properties of the solution of Equation (1) by understanding the solution of Equation (2). We also explain in Section 6 that our techniques can be adapted to the case where ϵ_i's have different distributions (the heteroskedastic case) and mention briefly potential extensions to the elliptical setting, as well as to weighted robust regression for instance. For the convenience of the reader, we summarize the key results of our analysis in the next few subsections. Sections 2, 3, 4 and 5 are devoted to proofs, the main contribution of the current paper. Section 6 describes results in the unpenalized case and extensions. The Appendix contains mathematical and statistical background. Our notations are standard but we redefine most of them on p. 12.

Probabilistically, this paper is part of an effort to understand the probabilistic properties of solutions of high-dimensional optimization problems depending on random observations, which we believe is an interesting endeavor from both a purely probabilistic and a more applied point of view. This paper shows that it is possible to do so using a random matrix point of view, which we believe also opens a new set of interesting questions for random matrix theorists.

Regression M-estimates are quite widely used, despite the fact that in some random design cases, they are known to have undesirable inadmissibility properties (Stein (1960), Baranchik (1973)) even in simple (Gaussian) situations. We do not dwell more on these otherwise interesting issues, since they are tangential to the main aim of this particular paper, which is to obtain a very detailed probabilistic understanding of $\widehat{\beta}$ and other quantities of interest.

1.1 Main focus of the paper and results

The focus of the paper is the problem of understanding the probabilistic properties of

$$\widehat{\beta} = \text{argmin}_{\beta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \rho(\epsilon_i - X_i'\beta) + \frac{\tau}{2}\|\beta\|^2 \tag{2}$$

where $\tau > 0$. For all $1 \leq i \leq n$, we have $\epsilon_i \in \mathbb{R}$ and $X_i \in \mathbb{R}^p$. We will see later (see Subsubsection 1.1.1 and Section 6.1) that under certain conditions on ρ, the understanding of $\widehat{\beta}$ for various $\tau > 0$ will lead us to rigorous understanding of $\widehat{\beta}$ when $\tau = 0$. This will then allow us to study the solution of Equation (1) in the standard linear regression model, i.e $Y_i = X_i'\beta_0 + \epsilon_i$.

The aim of the paper is to prove the following theorem:

Theorem 1.1. *Consider $\widehat{\beta}$ defined in Equation* (2) *and assume that $\tau > 0$ is given, i.e does not change with p and n. Under Assumptions **O1-O7**, **P1** and **F1-F2**, detailed below, we have: as p, n tend to infinity while $p/n \to \kappa \in (0,\infty)$, $\text{var}\left(\|\widehat{\beta}\|^2\right) \to 0$. Furthermore, for $r_\rho(\kappa)$ a deterministic scalar, if $\widehat{z}_\epsilon = \epsilon + r_\rho(\kappa) Z$, where ϵ is a random variable with the same distribution as ϵ_i's and Z is a $\mathcal{N}(0,1)$ random variable independent of ϵ, we have: $\|\widehat{\beta}\| \to r_\rho(\kappa)$ in probability and there exists a constant $c_\rho(\kappa) \geq 0$ such that*

$$\begin{cases} \mathbf{E}\left([prox_{c_\rho(\kappa)}(\rho)]'(\widehat{z}_\epsilon)\right) & = 1 - \kappa + \tau c_\rho(\kappa) \\ \kappa r_\rho^2(\kappa) & = \mathbf{E}\left((\widehat{z}_\epsilon - prox_{c_\rho(\kappa)}(\rho)(\widehat{z}_\epsilon))^2\right) \end{cases} . \tag{3}$$

We use the notation $\text{prox}_c(\rho)$ to denote the proximal mapping of the function $c\rho$, where $c \geq 0$. ρ is assumed to be convex. This notion was introduced in Moreau (1965). We recall that

$$\text{prox}_c(\rho)(x) = \text{argmin}_{y\in\mathbb{R}}(c\rho(y) + \frac{1}{2}(x-y)^2) \text{ , or equivalently,}$$
$$\text{prox}_c(\rho)(x) = (\text{Id} + c\psi)^{-1}(x) \text{ , where } \psi = \partial\rho$$

is the sub-differential of ρ (see Schirotzek (2007), p.59). The proximal mapping is an important notion in convex analysis and convex optimization (beside the very thorough and nice Moreau (1965), see for instance Beck and Teboulle (2010) or Ruszczyński (2006), Section 7.3). We note that even when ρ is not differentiable, $\text{prox}_c(\rho)(x)$ is a well-defined function.

As explained in Bean et al. (2013), the previous system can be reformulated in terms of $\text{prox}_1((c_\rho(\kappa)\rho)^*)$, where f^* represents the Fenchel-Legendre dual of f. Indeed, Moreau's prox identity (Moreau (1965)) gives

$$\text{prox}_1((c\rho)^*)(x) = x - \text{prox}_1(c\rho)(x) \ .$$

Example : our assumptions are for instance satisfied when

- X_i's are i.i.d with i.i.d entries. Those entries have mean 0 and variance 1 and have for instance a "strongly log-concave density" (i.e a density f_ϵ of the form $f_\epsilon = \exp(-g_\epsilon)$, where g_ϵ is convex with $g_\epsilon'' \geq C$ for some $C > 0$; an example is the Gaussian distribution) or are bounded. (See p. 11 or Appendix D-4 for justification.)
- ϵ_i's are i.i.d, independent of X_i's, and have a log-concave distribution that is symmetric around 0.
- the function ρ is twice-differentiable, convex and grows at most polynomially at ∞. Furthermore, its unique minimizer is at 0 where $\rho(0) = 0$.

As explained below, our assumptions are in fact much less restrictive than what is stated in the example just given (chosen mostly because it is simple to state). In case the reader is unfamiliar with Moreau's proximal mapping, we give a few examples in Appendix A.

1.1.1 From $\tau \neq 0$ and $\beta_0 = 0$ to $\tau = 0$ and $\beta_0 \neq 0$

The standard linear model in statistics assumes that the statistician observes $Y_i = X_i'\beta_0 + \epsilon_i$. Then it is of interest to understand the properties of $\widehat{\beta}(\{Y_i, X_i\}_{i=1}^n) - \beta_0$, where $\widehat{\beta}(\{Y_i, X_i\}_{i=1}^n)$ is defined as

$$\widehat{\beta}(\{Y_i, X_i\}_{i=1}^n) = \text{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \rho(Y_i - X_i'\beta) \ .$$

We have the following simple lemma.

Lemma (A)**.** *Suppose that span*$\{X_i\}_{i=1}^n = \mathbb{R}^p$ *and* $Y_i = X_i'\beta_0 + \epsilon_i$. *Suppose that* ρ *is strongly convex and call*

$$\widehat{\beta}(\{Y_i, X_i\}_{i=1}^n) = \text{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \rho(Y_i - X_i'\beta) \ ,$$
$$\widehat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \rho(\epsilon_i - X_i'\beta) \ .$$

Then

$$\widehat{\beta}(\{Y_i, X_i\}_{i=1}^n) - \beta_0 = \widehat{\beta} \ .$$

In particular, using Theorem 6.1, we see that when Assumptions ***O1-O7****,* ***P1*** *and* ***F1-F2*** *are satisfied,* $\lim p/n = \kappa < 1$*, and* ρ *is strongly convex, we have,*

$$\lim_{n,p\to\infty} \left| \|\widehat{\beta}\| - r_\rho(\kappa; 0) \right| \to 0 \textit{ in probability} \ ,$$

where $r_\rho(\kappa; 0) = \lim_{\tau \to 0} r_\rho(\kappa; \tau)$ *and* $r_\rho(\kappa; \tau)$ *is the quantity denoted by* $r_\rho(\kappa)$ *in Theorem 1.1.*

More information, details and justifications are provided in Subsection 6.1 and Theorem 6.1 on p. 42. (We refer the reader to Hiriart-Urruty and Lemaréchal (2001) p. 73 for a definition of strong convexity.)

The previous lemma simply states that to understand the properties of $\widehat{\beta}(\{Y_i, X_i\}_{i=1}^n) - \beta_0$, it is enough to understand those of $\widehat{\beta}_\tau$ as $\tau \to 0$, where $\widehat{\beta}_\tau$ is defined as

$$\widehat{\beta}_\tau = \text{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \rho(\epsilon_i - X_i'\beta) + \frac{\tau}{2} \|\beta\|^2 \ .$$

1.2 Key intermediate results

We present here some key intermediate results - of probabilistic, analytic and statistical interest - that appear in our proof for the convenience of the reader. Different parts of the proof require different assumptions, so we label the assumptions accordingly. The assumptions are progressively more restrictive. We decided to state them separately to show what aspects of the proof held under the less restrictive assumptions, something that would have been lost if we had just stated the most restrictive assumptions at once.

These results apply to $\widehat{\beta}$ as defined in Equation (2). τ is held fixed in our asymptotics and we choose to not index $\widehat{\beta}$ by τ to avoid cumbersome notations.

We believe our notations are standard, but definitions, if needed, can be found on p. 12.

1.2.1 Impact of leaving one observation out

We first consider the situation where we leave the i-th observation, (X_i, ϵ_i), out. We call, with standard notation,

$$\widehat{\beta}_{(i)} = \text{argmin}_{\beta \in \mathbb{R}^p} F_i(\beta) \ , \text{ where } F_i(\beta) = \frac{1}{n} \sum_{j \neq i} \rho(\epsilon_j - X_j'\beta) + \frac{\tau}{2} \|\beta\|^2 \ .$$

Definition. We call, assuming that ρ is twice differentiable and $\psi = \rho'$,

$$R_i = \epsilon_i - X_i'\widehat{\beta}\ , \text{ and } S = \frac{1}{n}\sum_{i=1}^{n} \psi'(R_i) X_i X_i'\ ,$$
$$\tilde{r}_{j,(i)} = \epsilon_j - X_j'\widehat{\beta}_{(i)} \text{ and } S_i = \frac{1}{n}\sum_{j\neq i} \psi'(\tilde{r}_{j,(i)}) X_j X_j'\ ,$$
$$f_i(\beta) = -\frac{1}{n}\sum_{j\neq i} X_j \psi(\epsilon_j - X_j'\beta) + \tau\beta \triangleq f(\beta) + \frac{1}{n} X_i \psi(\epsilon_i - X_i'\beta)\ .$$

Using the above definitions, let us now consider

$$\widetilde{\beta}_i = \widehat{\beta}_{(i)} + \frac{1}{n}(S_i + \tau \mathrm{Id})^{-1} X_i \psi(\mathrm{prox}_{c_i}(\rho)(\tilde{r}_{i,(i)})) \triangleq \widehat{\beta}_{(i)} + \eta_i\ , \tag{4}$$

where

$$c_i = \frac{1}{n} X_i'(S_i + \tau \mathrm{Id})^{-1} X_i\ , \text{ and} \tag{5}$$
$$\eta_i = \frac{1}{n}(S_i + \tau \mathrm{Id})^{-1} X_i \psi(\mathrm{prox}_{c_i}(\rho)(\tilde{r}_{i,(i)}))\ . \tag{6}$$

Our main results from Section 3, Theorem 3.1 and Proposition 3.4 state the following:

Theorem (B)**.** *Under Assumptions **O1-O7** stated below, we have, for any fixed k, when τ is held fixed,*

$$\sup_{1\leq i\leq n} \|\widehat{\beta} - \widetilde{\beta}_i\| = \mathrm{O}_{L_k}(\frac{\mathit{polyLog}(n)}{n})\ .$$

In particular, we have

$$\forall 1 \leq i \leq n\ , \mathbf{E}\left(\|\widehat{\beta} - \widetilde{\beta}_i\|^2\right) = \mathrm{O}(\mathit{polyLog}(n)/n^2)\ .$$

Also,

$$\sup_{1\leq i\leq n}\ \sup_{j\neq i} |\tilde{r}_{j,(i)} - R_j| = \mathrm{O}_{L_k}(\frac{\mathit{polyLog}(n)}{n^{1/2}})\ ,$$
$$\sup_i |R_i - \mathit{prox}_{c_i}(\rho)(\tilde{r}_{i,(i)})| = \mathrm{O}_{L_k}(\frac{\mathit{polyLog}(n)}{n^{1/2}})\ ,$$
$$\mathrm{var}\left(\|\widehat{\beta}\|^2\right) = \mathrm{O}(\frac{\mathit{polyLog}(n)}{n})\ .$$

Beside their probabilistic interest, these results give us fine insights about how to perform "online updates" for $\widehat{\beta}$ (i.e how to change $\widehat{\beta}$ if a new observation becomes available without solving the optimization problem anew, something that is very useful computationally) and the error made in doing so. They also show that the behavior of the residuals R_i is very different from the classical setting $\lim p/n \to 0$ (in that setting, we basically have $R_i \simeq \epsilon_i$), and that Moreau's proximal mapping is a key ingredient for understanding the marginal behavior of the residuals. Under the assumptions detailed below, the probabilistic behavior of $\tilde{r}_{i,(i)}$ is much simpler to understand than that of R_i - that is one of the motivations for relating these two quantities.

For this part of the proof (i.e "leave-one-**O**bservation-out"), we work under the following assumptions:

- **O1**: p/n has a finite non-zero limit.
- **O2**: ρ is twice differentiable, convex and non-linear. $\psi = \rho'$. Note that $\psi' \geq 0$ since ρ is convex. We assume that $\rho \geq 0$, $\rho(0) = 0$ and 0 is the unique minimizer of ρ. Note that this implies that $\mathrm{sign}(\psi(x)) = \mathrm{sign}(x)$.
- **O3**: ψ' is $L(u)$-Lipschitz on $(-|u|, |u|)$, where $L(|u|) \leq K|u|^{m_1}$ as $|u| \to \infty$. Hence, $\psi(|x|) = \mathrm{O}(|x|^m)$ at infinity for some m and ρ grows at most polynomially at ∞.

- **O4**: X_i's are independent and identically distributed. $X_i \in \mathbb{R}^p$. Their distribution is allowed to change with p and n. Furthermore, for any 1-Lipschitz (with respect to Euclidean norm) convex function G, if $m_{G(X_i)}$ is a median of $G(X_i)$, for any $t > 0$, $P(|G(X_i) - m_{G(X_i)}| > t) \leq C_n \exp(-\mathsf{c}_n t^2)$, C_n and c_n can vary with n. For simplicity, we assume that, $1/\mathsf{c}_n = \mathrm{O}(\log(n)^\alpha)$ for some $\alpha \geq 0$ and C_n is bounded in n. X_i's have mean 0 and $\mathrm{cov}\,(X_i) = \mathrm{Id}_p$. Furthermore, for any given k, we assume that the k-th moment of the entries of X_i is bounded independently of n and p.
- **O5**: $\{X_i\}_{i=1}^n$ are independent of $\{\epsilon_i\}_{i=1}^n$
- **O6**: for any fixed k , $\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left(\psi^{2k}(\epsilon_i)\right)$ remains uniformly bounded in p and n, as both grow to infinity.
- **O7**: $\sup_{1\leq i\leq n} |\epsilon_i| \triangleq \mathcal{E}_n = \mathrm{O}_{L_k}((\log n)^\beta)$ for some $\beta > 0$ and ϵ_i's are independent. In other words, we assume that for any given $k \geq 1$, $\mathbf{E}\left(|\mathcal{E}_n|^k\right) = \mathrm{O}((\log n)^{\beta k})$.

We do not assume at this point that ϵ_i's have identical distribution. (Note that if ϵ_i's are log-concave or bounded, **O6-O7** are satisfied.)

1.2.2 Impact of leaving one predictor out

For the second part of the proof (i.e "leave-one-**P**redictor-out"), we need all the previous assumptions and

- **P1**: X_i's have i.i.d entries. We call $X_i(k)$ the k-th coordinate of X_i. Furthermore, the vectors $\Theta_k = (X_1(k), \ldots, X_n(k))$ in $\mathbb{R}^n$ satisfy: for any 1-Lipschitz (with respect to Euclidean norm) convex function G, if $m_{G(\Theta_k)}$ is a median of $G(\Theta_k)$, for any $t > 0$, $P(|G(\Theta_k) - m_{G(\Theta_k)}| > t) \leq C_n \exp(-\mathsf{c}_n t^2)$, C_n and c_n can vary with n. For simplicity, we assume that $1/\mathsf{c}_n = \mathrm{O}(\log(n)^\alpha)$ for some $\alpha \geq 0$ and C_n is bounded in n.

We note that according to Corollary 4.10 and the discussion that follows in Ledoux (2001), Assumptions **O4** and **P1** are compatible. **O4** and **P1** are for instance satisfied if the entries of X_i's are independent and bounded by $\mathrm{O}((\log n)^{\alpha/2})$. Another example is the case of $X_i \sim \mathcal{N}(0, \mathrm{Id}_p)$. We note that as the proof will make clear, the assumption that X_i's have the same distribution at given n and p could be relaxed. By contrast, we use strongly the assumption of independence.

Let us now state the main approximation results we get by leaving one predictor out.

We call V the $n \times (p-1)$ matrix corresponding to the first $(p-1)$ columns of the design matrix X. We call V_i in $\mathbb{R}^{p-1}$ the vector corresponding to the first $p-1$ entries of X_i, i.e $V_i' = (X_i(1), \ldots, X_i(p-1))$. We call $X(p)$ the vector in $\mathbb{R}^n$ with j-th entry $X_j(p)$.

Let us call $\widehat{\gamma}$ the solution of our optimization problem when we use the design matrix V instead of X. In other words,

$$\widehat{\gamma} = \mathrm{argmin}_{\gamma\in\mathbb{R}^{p-1}} \frac{1}{n}\sum_{i=1}^n \rho(\epsilon_i - V_i'\gamma) + \frac{\tau}{2}\|\gamma\|^2 \, . \tag{7}$$

(It is easy to see that $\begin{pmatrix}\widehat{\gamma}\\ 0\end{pmatrix}$ is the solution of the original optimization problem (2) when $X_i(p)$ is replaced by 0.)

The corresponding residuals are $\{r_{i,[p]}\}_{i=1}^n$, i.e

$$r_{i,[p]} = \epsilon_i - V_i'\widehat{\gamma} \, .$$

We call

$$u_p = \frac{1}{n}\sum_{i=1}^n \psi'(r_{i,[p]}) V_i X_i(p) \, , \text{ and } \mathfrak{S}_p = \frac{1}{n}\sum_{i=1}^n \psi'(r_{i,[p]}) V_i V_i' \, .$$

Note that $u_p \in \mathbb{R}^{p-1}$ and $\mathfrak{S}_p$ is $(p-1)\times(p-1)$. We call

$$\xi_n \triangleq \frac{1}{n}\sum_{i=1}^n X_i^2(p)\psi'(r_{i,[p]}) - u_p'(\mathfrak{S}_p + \tau\mathrm{Id})^{-1}u_p \, , \tag{8}$$

and

$$N_p \triangleq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i(p) \psi(r_{i,[p]}) \ . \tag{9}$$

It is shown in Subsubsection 4.2.2 that $\xi_n \geq 0$. We consider

$$\mathfrak{b}_p \triangleq \frac{1}{\sqrt{n}} \frac{N_p}{\tau + \xi_n} \ , \tag{10}$$

and call

$$\widetilde{b} = \begin{bmatrix} \widehat{\gamma} \\ 0 \end{bmatrix} + \mathfrak{b}_p \begin{bmatrix} -(\mathfrak{S}_p + \tau \mathrm{Id})^{-1} u_p \\ 1 \end{bmatrix} \ . \tag{11}$$

We have, if $\widehat{\beta}_p$ denotes the last coordinate of $\widehat{\beta}$:

Theorem (C)**.** *Under Assumptions* ***O1-O7*** *and* ***P1****, we have, for any fixed* $\tau > 0$*,*

$$\|\widehat{\beta} - \widetilde{b}\| \leq \mathrm{O}_{L_k}\left(\frac{polyLog(n)}{n}\right)$$

Furthermore,

$$\sqrt{n}(\widehat{\beta}_p - \mathfrak{b}_p) = \mathrm{O}_{L_k}(polyLog(n)/\sqrt{n}) \ ,$$

$$\sup_i |X_i'(\widehat{\beta} - \widetilde{b})| = \mathrm{O}_{L_k}\left(\frac{polyLog(n)}{\sqrt{n}}\right) \ ,$$

$$\sup_i |R_i - r_{i,[p]}| = \mathrm{O}_{L_k}\left(\frac{polyLog(n)}{\sqrt{n}}\right) \ .$$

This is the statement of Theorem 4.1 on p.33.

The last coordinate of $\widetilde{b}$, $\mathfrak{b}_p$, has a much simpler probabilistic structure under Assumption **P1** than $\widehat{\beta}_p$, the last coordinate of $\widehat{\beta}$. Because our approximations in the previous theorem are sufficiently good, we will be able to transfer our insights about $\mathfrak{b}_p$ to $\widehat{\beta}_p$.

We also have the following results concerning c_i's and $\|\widehat{\beta}\|$. We call, using the definition of S on p.7,

$$\mathsf{c}_{\tau,p} = \frac{1}{n} \mathrm{trace}\left((\mathfrak{S}_p + \tau \mathrm{Id}_{p-1})^{-1}\right) \ ,$$

$$c_\tau = \frac{1}{n} \mathrm{trace}\left([S + \tau \mathrm{Id}_p]^{-1}\right) \ .$$

Then we have, under Assumptions **O1-O7** and **P1**, with c_i defined in (5) on p. 7, the following results.

Proposition (D)**.**

$$\sup_i |c_i - c_\tau| = \mathrm{O}_{L_k}(n^{-1/2} polyLog(n)) \ \textit{and} \ |c_\tau - \mathsf{c}_{\tau,p}| = \mathrm{O}_{L_k}(n^{-1/2} polyLog(n)) \ .$$

Furthermore,

$$\left| \mathsf{c}_{\tau,p}(\xi_n + \tau) - \frac{p-1}{n} \right| = \mathrm{O}_{L_k}\left(\frac{polyLog(n)}{\sqrt{n}}\right) \ .$$

Also,

$$\left(\frac{p}{n}\right)^2 n \mathbf{E}\left(\mathfrak{b}_p^2\right) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{E}\left((\mathsf{c}_{\tau,p} \psi(r_{i,[p]})^2\right) + \mathrm{o}(1) \ .$$

And finally,

$$\frac{p}{n} \mathbf{E}\left(\|\widehat{\beta}\|^2\right) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{E}\left((c_\tau \psi(prox_{c_\tau}(\rho)(\tilde{r}_{i,(i)})))^2\right) + \mathrm{o}(1) \ .$$

This last equation is at a high-level what gives us the second equation in System (3). The second one, following "Furthermore,", is closely related to the first equation in System (3), as can be understood from reading Section 4.

1.2.3 Final steps

The last steps of the proof in Section 5 are divided into two: first we show that $\widehat{\beta}'_{(i)}X_i$ is asymptotically normal, with obvious consequences for $\tilde{r}_{i,(i)} = \epsilon_i - \widehat{\beta}'_{(i)}X_i$. This is done in Lemma 5.1. Then some work is needed (under assumptions **F1-F2** below) to show that our system has a unique solution and that c_τ (and hence c_i's) is asymptotically deterministic. This is done in Lemma 5.4.

The assumptions we just mentioned are:

- **F1**: the ϵ_i's have identical distribution and for any $r > 0$, if $Z \sim \mathcal{N}(0,1)$, independent of ϵ_i, $\epsilon_i + rZ$ has a density f which is increasing on $(-\infty, 0)$ and decreasing on $(0, \infty)$. Furthermore, $\lim_{|t|\to\infty} tf(t) = 0$.

- **F2**: For any fixed k, $\mathbf{E}\left(|\epsilon_i|^k\right) < \infty$.

We refer the reader to Lemma C-1 and the discussion immediately following it for examples of densities fulfilling the assumptions made in **F1**. We note that symmetric (around 0) log-concave densities will for instance satisfy all the assumptions we made about the ϵ_i's. See Karlin (1968) and Ibragimov (1956) for instance.

We could relax assumption **F1** to ϵ_i's having identical distribution and many of our arguments would go through, except for the fact that $c_\rho(\kappa)$ in our system (3) would only be shown to be a random variable, with possibly non-zero variance. The proof of Lemma 5.4 explains this in much more details. Equation (39) is especially important to understand c_τ, which is very closely related to $c_\rho(\kappa)$.

Remark : We note that if one is interested in understanding the fluctuation behavior of $\widehat{\beta}_p$, the approximations above (in particular in Theorem (C)) and the definition of $\mathfrak{b}_p$ in Equation (10) lead to fairly easy central limit theorems for $\widehat{\beta}_p$ and, by symmetry, the other coordinates of $\widehat{\beta}$. For space reasons, we leave the minor technical details that need to be filled in to the interested reader. (Of course in the case of i.i.d Gaussian predictors with identity covariance, the rotational invariance arguments given in El Karoui et al. (2011) apply when $p/n < 1$ and allow us to characterize the fluctuation behavior of $\sqrt{n}v'\widehat{\beta}$ for any fixed vector v with $\|v\| = 1$ from simply understanding $\|\widehat{\beta}\|$)

Remarks on the assumptions

- **Assumptions concerning ρ:** in this paper, we wanted to allow ρ to grow reasonably fast at infinity.

One of the motivations for this was to be able to handle a broad class of situations where ϵ_i's are i.i.d with a log-concave density f_ϵ and for our results to hold when $\rho = -\log(f_\epsilon)$. This is a natural choice from the point of view of maximum-likelihood estimation and the paper Bean et al. (2013) which highlighted, in the case of log-concave errors, the importance of the functions $\rho_\#$ described in the introduction. We note that the new "canonical" loss functions $\rho_\#$ tend to be fairly smooth (see Bean et al. (2013)) and hence the fact that the current paper requires ρ to be twice differentiable is not a source of major concerns to us. The classic reference papers Huber (1973); Portnoy (1984, 1985); Mammen (1989) also all require smooth ρ's. We also note that if a function $\psi = \rho'$ of interest is for instance not differentiable at one or two (or a few points), our main result, Theorem 1.1, will apply to a slightly smoothed version of ψ and hence an approximation of ρ. For the purpose of the current paper - where the effort is really probabilistic, trying to handle fairly general X_i's - working with smooth ρ is therefore enough. See nonetheless Appendix D-3. Finally, in the context of Lemma (A) our results are stated for strongly convex ρ. This is not very natural for some (but not all) questions in "robust statistics" - but it is not a problem for the kind of optimality questions that one could tackle using Lemma (A) or Theorem 1.1 - we discuss these issues a bit more in Appendix D-2. However, it is well-known that if ρ is convex, $\rho_\eta = \rho + \eta p_2$, where $p_2(x) = x^2/2$ and $\eta > 0$ is strongly convex, almost by definition (see Hiriart-Urruty and Lemaréchal (2001), p. 73). Once again, it seems that a bit more work of an approximation theoretic nature should allow us to extend the current results of Lemma (A), which apply to ρ_η for any $\eta > 0$ to a ρ that is not strongly convex. Of course, the fact that we can handle ρ's that grow at infinity quite fast is important to allow this kind of approximation arguments. (We also recall that the ϵ_i's we are concerned with in the current paper are not allowed to have very heavy-tails because we allow ρ to grow "fast" at infinity.)

A detailed look at the proof reveals that if we had more restrictive growth conditions on ρ at infinity than the ones we impose in the paper, we could tolerate ϵ_i's with fewer moments and heavier tails. Understanding how heavy the tails of ϵ_i can be and while our system (3) remains valid is interesting statistically, but we leave these considerations for future work since our focus in the current paper is primarily probabilistic and is on the development of mathematical tools and strategy for rigorously tackling this class of problems. We refer the interested reader to Appendix D-2 for a longer discussion of these and related issues explaining why we chose the setup we consider in the current paper.

We finally would like to clarify a little bit a semantic point: the optimization problems we consider in this paper are associated by many researchers in statistics with "robust statistics", which generally deals with ϵ_i's having heavy tails (and hence the functions ρ's that are considered in that field are quite restrictive, from a mathematical point of view). The fact that we consider ϵ_i's having relatively light tails and ρ's that can grow at infinity "fast" is motivated by two factors: one is purely mathematical, since the growth conditions on ρ at infinity create a number of challenges; the other one is that we are not concerned here with the impact of having a few ϵ_i's having a heavy-tailed distribution on the probabilistic properties of $\widehat{\beta}$ (something the current techniques nonetheless seem able to handle when ρ's are not allowed to be as general as the ones considered here). Rather, one motivation for our setup is to show that even in a "simple" context, where ϵ_i's do not have heavy-tails, standard methods of statistics do not perform in high-dimension as low-dimensional (i.e p fixed, $n \to \infty$) intuition would suggest. This is the content of Lemma (A) and the results of the paper Bean et al. (2013), when for instance, ϵ_i's are i.i.d with density $f_\epsilon = \exp(-g)$ and g is strongly convex, symmetric around 0 and has its unique minimum at 0: even in this simple setting, one can improve upon maximum-likelihood techniques, which are basically optimal in low-dimension.

- **Assumptions concerning X_i's** Assumption **O4** is a bit stronger than we will need. For instance, Sections 2 and 3 do not actually require the X_i's to have identical distributions. The functions G we will be dealing with will either be linear or square-root of quadratic forms, so we could limit our assumptions to those functions. However, as documented in Ledoux (2001), a large number of natural or "reasonable" distributions satisfy the **O4** assumptions - see also Appendix D-4. Our choice of having a potentially varying c_n is motivated by the idea that we could, for instance, relax an assumption of boundedness of the entries of X_i's - that guarantees that **O4** and **P1** are satisfied when X_i has independent entries, see Appendix D-4 - and replace it by an assumption concerning the moments of the entries of X_i's: this is what we did for instance in El Karoui (2009) through a truncation of triangular arrays argument (see also Yin et al. (1988)). We also refer the interested reader to El Karoui (2009) for a short list of distributions satisfying **O4**, compiled from various parts of Ledoux (2001). Finally, we could replace the $\exp(-\mathsf{c}_n t^2)$ upper bound in e.g **O4** by $\exp(-\mathsf{c}_n t^\alpha)$ for some fixed $\alpha > 0$ and it seems that all our arguments would go through. We chose not to work under these more general assumptions because it would involve extra book-keeping and does not enlarge the set of distributions we can consider enough to justify this extra technical cost. From a more applied point of view, Assumption **O4** imposes certain restrictions on the Euclidean geometry of the "point cloud" generated by the X_i's - see e.g El Karoui (2009) for more details on this or Appendix D-4. Working at the level of generality of Assumption **O4** allows us to show that this geometry plays a key role in our main results.

Our assumption that $1/\mathsf{c}_n$ increases like a power of $\log(n)$ at most is quite restrictive when it comes to bounded random variables (or truncating random variables) - but is of course satisfied by e.g Gaussian random variables where c_n is a constant independent on n - and motivated by simplifying the book-keeping needed in our proof. The result also applies to $X_{i,j}$'s that are i.i.d and bounded (uniformly in n). Having $1/\mathsf{c}_n$ grow like n^γ for a small γ should be feasible - with γ depending on m and m_1 (see **O3**). In the first part of the proof we keep track of the impact of c_n to illustrate this aspect of the problem.

We would also like to address the question of whether working under **O4** and **P1** is "artificially general". While it is true that we could assume i.i.d-ness of $X_{i,j}$ (and some conditions on their distribution), we would then constantly be using in the proofs the fact that the functionals of $X_{i,j}$ we need in this paper satisfy certain concentration properties. We feel that working under these simpler-to-state assumptions (like i.i.d-ness) would potentially obscure some of the geometric properties of the vectors X_i's - concerning e.g $\|X_i\|/\sqrt{p}$ or $X_i'X_j/p$ - that appear to be fundamental in establishing Theorem 1.1. Working under the assumptions we state makes rather clear (at least in light of previous papers and the short discussion in Appendix D-4) the role of these geometric properties.

Finally, we think the questions raised here are interesting from both a probabilistic and statistical point of views. Hence, we worked under broad assumptions for the sake of mathematical and probabilistic interest, even though for certain statistical tasks, less general assumptions (concerning e.g ψ) are arguably more natural. However, these assumptions (such as having ψ bounded) seem to render the analysis simpler. Hence since our aim was mostly probabilistic in this paper, we chose to work under more general assumptions that enlarge the domain of validity of the results to a large class of interesting situations and also forced us to deal with numerous extra technical and conceptual difficulties - the proof makes this clear.

Notations

We will repeatedly use the following notations: polyLog(n) is used to replace a power of $\log(n)$; $\lambda_{\max}(M)$ denotes the largest eigenvalue of the matrix M; $|||M|||_2$ denotes the largest singular value of M. We call $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i'$ the usual sample covariance matrix of the X_i's when X_i's are known to have mean 0. We say that $X \leq Y$ in L_k if $\mathbf{E}\left(|X|^k\right) \leq \mathbf{E}\left(|Y|^k\right)$. We write $X \overset{\mathcal{L}}{=} Y$ to say that the random variables X and Y are equal in law. We use the notation $u_n \lesssim v_n$ to say that there exists a constant K independent of n such that $u_n \leq K v_n$ for all n. We use the usual notation $\widehat{\beta}_{(i)}$ to denote the regression vector we obtain when we do not use the pair (X_i, Y_i) or (X_i, ϵ_i) in our optimization problem, a.k.a the leave-one-out estimate. We will also use the notation $X_{(i)}$ to denote $\{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\}$. We use the notation (a, b) for either the interval (a, b) or the interval (b, a): in several situations, we will have to localize quantities in intervals using two values a and b but we will not know whether $a < b$ or $b > a$. We denote by X the $n \times p$ design matrix whose i-th row is X_i'. $\|v\|$ denotes the Euclidean norm of the vector v. $\kappa_l(\xi)$ stands for the k-th cumulant of the random variable ξ. We write $a \wedge b$ for $\min(a, b)$ and $a \vee b$ for $\max(a, b)$. If A and B are two symmetric matrices, $A \succeq B$ means that $A - B$ is positive semi-definite, i.e A is greater than B in the positive-definite/Loewner order. The notations o_P, O_P are used with their standard meanings, but see van der Vaart (1998) p.12 for definitions if needed. For the random variable W, we use the definition $\|W\|_{L_k} = \left[\mathbf{E}\left(|W|^k\right)\right]^{1/k}$. For sequences of random variables W_n, Z_n, we use the notation $W_n = \mathrm{O}_{L_k}(Z_n)$ (resp $W_n = \mathrm{o}_{L_k}(Z_n)$) when $\|W_n\|_{L_k} = \mathrm{O}(\|Z_n\|_{L_k})$ (resp $\|W_n\|_{L_k} = \mathrm{o}(\|Z_n\|_{L_k})$).

Remarks

Note that under our assumptions on ρ, $\widehat{\beta}$, the solution of Equation (2), is defined as the solution of

$$f(\widehat{\beta}) = 0 \text{ with} \tag{12}$$

$$f(\beta) = \frac{1}{n}\sum_{i=1}^n -X_i \psi(\epsilon_i - X_i'\beta) + \tau\beta \ . \tag{13}$$

We call

$$F(\beta) = \frac{1}{n}\sum_{i=1}^n \rho(\epsilon_i - X_i'\beta) + \frac{\tau}{2}\|\beta\|^2 \ . \tag{14}$$

Of course, $f = \nabla_\beta F$.

2 Preliminaries

In all the paper, we work under Assumption **O2**, which is purely about the function ρ, i.e there are no probabilistic elements in this assumption. We also assume **O1**, which guarantees that p/n remains bounded.

In case it is helpful to the reader, we give a very high-level overview of our proof strategy in Appendix D-1.

2.1 General remarks

Proposition 2.1. *Let β_1 and β_2 be two vectors in $\mathbb{R}^p$. Then*

$$\boxed{\|\beta_1 - \beta_2\| \leq \frac{1}{\tau}\|f(\beta_1) - f(\beta_2)\| \;.} \tag{15}$$

When ρ is strongly convex with modulus of convexity $C > 0$, we also have

$$\|\beta_1 - \beta_2\| \leq \frac{1}{C\lambda_{\min}(\widehat{\Sigma}) + \tau}\|f(\beta_1) - f(\beta_2)\| \;.$$

For a definition of modulus of convexity we refer to Proposition 1.1.2 on p. 73 in Hiriart-Urruty and Lemaréchal (2001). When ρ is twice differentiable, the modulus of convexity is a lower bound on its second derivative (see Theorem 4.3.1 on p. 115 in Hiriart-Urruty and Lemaréchal (2001)).

Proof. Let β_1 and β_2 be two vectors in $\mathbb{R}^p$. We have by definition

$$f(\beta_1) - f(\beta_2) = \tau(\beta_1 - \beta_2) + \frac{1}{n}\sum_{i=1}^{n} X_i \left[\psi(\epsilon_i - X_i'\beta_2) - \psi(\epsilon_i - X_i'\beta_1)\right] \;.$$

We can use the mean value theorem to write

$$\psi(\epsilon_i - X_i'\beta_2) - \psi(\epsilon_i - X_i'\beta_1) = \psi'(\gamma^*_{\epsilon_i, X_i'\beta_1, X_i'\beta_2})X_i'(\beta_1 - \beta_2) \;,$$

where $\gamma^*_{\epsilon_i, X_i'\beta_1, X_i'\beta_2}$ is in the interval $(\epsilon_i - X_i'\beta_1, \epsilon_i - X_i'\beta_2)$ - recall that we do not care about the order of the endpoints in our notation.

We therefore have

$$f(\beta_1) - f(\beta_2) = \tau(\beta_1 - \beta_2) + \frac{1}{n}\sum_{i=1}^{n} \psi'(\gamma^*_{\epsilon_i, X_i'\beta_1, X_i'\beta_2})X_i X_i'(\beta_1 - \beta_2) \;,$$

which we write

$$f(\beta_1) - f(\beta_2) = (\mathsf{S}_{\beta_1,\beta_2} + \tau \mathrm{Id}_p)(\beta_1 - \beta_2) \;, \tag{16}$$

where

$$\mathsf{S}_{\beta_1,\beta_2} = \frac{1}{n}\sum_{i=1}^{n} \psi'(\gamma^*_{\epsilon_i, X_i'\beta_1, X_i'\beta_2})X_i X_i' \;.$$

We therefore have

$$\beta_1 - \beta_2 = (\mathsf{S}_{\beta_1,\beta_2} + \tau \mathrm{Id}_p)^{-1}\left(f(\beta_1) - f(\beta_2)\right) \;.$$

Since ρ is convex, $\psi' = \rho''$ is non-negative and $\mathsf{S}_{\beta_1,\beta_2}$ is positive semi-definite. In the semi-definite order, we have $\mathsf{S}_{\beta_1,\beta_2} + \tau\mathrm{Id}_p \succeq \tau\mathrm{Id}_p$. When ρ is strongly convex with modulus C, we have $\psi'(x) \geq C$ (see Theorem 4.3.1 p. 115 in Hiriart-Urruty and Lemaréchal (2001)) and therefore, $\mathsf{S}_{\beta_1,\beta_2} + \tau\mathrm{Id}_p \succeq C\widehat{\Sigma} + \tau\mathrm{Id}_p \succeq (C\lambda_{\min}(\widehat{\Sigma}) + \tau)\mathrm{Id}_p$. In particular,

$$\|\beta_1 - \beta_2\| \leq \frac{1}{\tau}\|f(\beta_1) - f(\beta_2)\| \;.$$

In the strongly convex case, we have

$$\|\beta_1 - \beta_2\| \leq \frac{1}{C\lambda_{\min}(\widehat{\Sigma}) + \tau}\|f(\beta_1) - f(\beta_2)\| \;.$$

□

Proposition 2.1 yields the following lemma.

Lemma 2.1. *For any β_1,*

$$\|\widehat{\beta} - \beta_1\| \leq \frac{1}{\tau}\|f(\beta_1)\| \;.$$

The lemma is a simple consequence of Equation (15) since by definition $f(\widehat{\beta}) = 0$.

In the following, we will strive to find approximations of $\widehat{\beta}$. We will therefore use Lemma 2.1 repeatedly.

2.2 Boundedness of $\|\widehat{\beta}\|$

We have the following lemma.

Lemma 2.2. *Let us call $W_n = \frac{1}{n}\sum_{i=1}^n X_i\psi(\epsilon_i)$, $W_n \in \mathbb{R}^p$. We have*

$$\|\widehat{\beta}\| \leq \frac{1}{\tau}\|W_n\| \ .$$

*In particular, under Assumptions **O4** and **O5**,*

$$\mathbf{E}\left(\|\widehat{\beta}\|^2\right) \leq \frac{1}{\tau^2}\frac{p}{n}\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left(\psi^2(\epsilon_i)\right) \ . \tag{17}$$

If $k \geq 1$, a similar result holds in L_{2k} - provided the entries of X_i's have cumulants of order $2k$ bounded in n. In other words,

$$\mathbf{E}\left(\|\widehat{\beta}\|^{2k}\right) = \mathrm{O}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left(\psi^{2k}(\epsilon_i)\right)\right) \ .$$

*These conditions are automatically satisfied under our assumptions **O4** and **O6**.*

This guarantees that $\|\widehat{\beta}\|$ is bounded in L_{2k} provided $\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left(|\psi(\epsilon_i)|^{2k}\right)$ is bounded. If this latter quantity is polyLog(n) so is $\mathbf{E}\left(\|\widehat{\beta}\|^{2k}\right)$.

We also have

$$\|\widehat{\beta}\| \leq \sqrt{\frac{2}{\tau}}\sqrt{\frac{1}{n}\sum_{i=1}^n \rho(\epsilon_i)} \ , \tag{18}$$

and hence

$$\mathbf{E}\left(\|\widehat{\beta}\|^{2k}\right) \leq \frac{2^k}{\tau^k}\mathbf{E}\left(\left[\frac{1}{n}\sum_{i=1}^n \rho(\epsilon_i)\right]^k\right) \ .$$

Though from a probabilistic point of view our various bounds might look interchangeable, it is important to have both from the point of view of potential statistical applications (beyond the scope of this paper). Indeed, in some robust regression problems, where ϵ_i's can have heavy tails, one would typically used bounded ψ functions (for instance the Huber functions or smoothed version of the Huber functions - see Huber and Ronchetti (2009), p. 84, Equation (4.51) for a definition of the exponential of the Huber functions). The bound based on Equation (17) will then be particularly helpful.

Proof. The first inequality follows easily from taking $\beta_1 = 0$ in Lemma 2.1 and realizing that $W_n = -f(0)$. The second inequality follows from the fact that, if $\mathbf{e}$ is an n-dimensional vector with entries all equal to 1, $W_n = X'D_\psi\mathbf{e}/n$, where X is $n \times p$ and D_ψ is a diagonal matrix whose (i,i) entry is $\psi(\epsilon_i)$. Hence,

$$\|W_n\|^2 = \frac{1}{n^2}\mathbf{e}'D_\psi XX'D_\psi\mathbf{e} \ ,$$

and therefore, $\mathbf{E}\left(\|W_n\|^2\right) = \frac{p}{n^2}\sum_{i=1}^n \mathbf{E}\left(\psi^2(\epsilon_i)\right)$, since $\mathbf{E}\left(XX'\right) = p\mathrm{Id}_n$ and $\{\epsilon_i\}_{i=1}^n$ is independent of $\{X_i\}_{i=1}^n$.

For the L_{2k} bound, we can use $\mathbf{E}\left(\|W_n\|^{2k}\right) \leq p^{k-1}\sum_{j=1}^p \mathbf{E}\left(W_n^{2k}(j)\right)$, because for $\alpha_i \geq 0$, $(\sum_{i=1}^p \alpha_i)^k \leq p^{k-1}\sum_{i=1}^p \alpha_i^k$ by convexity.

Let us work temporarily conditional on ϵ_i. We control $\mathbf{E}\left(W_n^{2k}(j)|\{\epsilon_i\}_{i=1}^n\right)$ through the use of cumulants since $W_n(j) = \sum_{i=1}^n X_i(j)\psi(\epsilon_j)/n$, so the $2k$-th cumulant of $W_n(j)$ - conditional on $\{\epsilon_i\}_{i=1}^n$ - is $\sum_{i=1}^n \psi^{2k}(\epsilon_i)/n^{2k}\kappa_{2k}(X_i(j))$. These cumulants are all of order n^{1-2k}, if $\sum \psi^{2k}(\epsilon_i)/n = O(1)$. By the classical connection between moments and cumulants, we see that $\mathbf{E}\left(W_n^{2k}(j)\right) = \mathrm{O}(n^{-k})$ if $\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left(\psi^{2k}(\epsilon_i)\right)$ is uniformly bounded. Hence, $\mathbf{E}\left(\|W_n\|^{2k}\right) = \mathrm{O}(p^{k-1}pn^{-k}) = \mathrm{O}(1)$.

The proof of Equation (18) simply follows from observing that, since $\rho \geq 0$,

$$\frac{\tau}{2}\|\widehat{\beta}\|^2 \leq \frac{1}{n}\sum_{i=1}^{n}\rho(\epsilon_i - X_i'\widehat{\beta}) + \frac{\tau}{2}\|\widehat{\beta}\|^2 = F(\widehat{\beta})$$
$$\leq \frac{1}{n}\sum_{i=1}^{n}\rho(\epsilon_i) = F(0) \;.$$

Indeed, since, according to Equation (14), $\widehat{\beta} = \operatorname{argmin}_{\beta\in\mathbb{R}^p} F(\beta)$, we have $F(\widehat{\beta}) \leq F(0) = \frac{1}{n}\sum_{i=1}^{n}\rho(\epsilon_i)$, and the result follows immediately.

□

3 Approximating $\widehat{\beta}$ by $\widehat{\beta}_{(i)}$: leave-one-observation-out

We consider the situation where we leave the i-th observation, (X_i, ϵ_i), out and refer the reader to Subsection 1.2 for definitions of the quantities that will play a key role in our analysis.

These definitions and the approximations they will imply can be understood in light of the probabilistic heuristics we derived for this problem in El Karoui et al. (2011) and El Karoui et al. (2013) - so we refer the reader to these papers for explanations and intuition about why we choose to introduce these quantities; see also Appendix D-1.

With the definitions introduced in Subsection 1.2 and Subsubsection 1.2.1, the aim of the work in this section to show that $\widehat{\beta}$ can be very well approximated by $\widetilde{\beta}_i$. In Theorem 3.1, we show that the approximation is accurate to order polyLog$(n)/n$ in Euclidean norm, if for instance $1/\mathsf{c}_n = \mathrm{O}(\mathrm{polyLog}(n))$. We refer the reader to Theorem 3.1 for full details.

3.1 Deterministic bounds

We refer the reader to Subsubsection 1.2.1 where the important quantities $\widetilde{\beta}_i$, η_i, $\tilde{r}_{j,(i)}$ and f_i are defined.

Proposition 3.1. *We have, with $\widetilde{\beta}_i$ defined in Equation (4),*

$$\|\widehat{\beta} - \widetilde{\beta}_i\| \leq \frac{1}{\tau}\|\mathcal{R}_i\| \;, \tag{19}$$

where

$$\mathcal{R}_i = \frac{1}{n}\sum_{j\neq i}\left[\psi'(\gamma^*(X_j, \widehat{\beta}_{(i)}, \eta_i)) - \psi'(\tilde{r}_{j,(i)})\right] X_j X_j' \eta_i \;, \tag{20}$$

and $\gamma^*(X_j, \widehat{\beta}_{(i)}, \eta_i)$ is in the ("unordered") interval $(\tilde{r}_{j,(i)}, \tilde{r}_{j,(i)} - X_j'\eta_i)$=$(\epsilon_j - X_j'\widehat{\beta}_{(i)}, \epsilon_j - X_j'\widetilde{\beta}_i)$.

Proof. We have of course, since $f_i(\widehat{\beta}_{(i)}) = 0$, and $\widetilde{\beta}_i = \widehat{\beta}_{(i)} + \eta_i$,

$$f(\widetilde{\beta}_i) = f(\widetilde{\beta}_i) - f_i(\widehat{\beta}_{(i)}) = -\frac{1}{n}X_i\psi(\epsilon_i - X_i'\widetilde{\beta}_i) + \frac{1}{n}\sum_{j\neq i}X_j\left[\psi(\epsilon_j - X_j'\widehat{\beta}_{(i)}) - \psi(\epsilon_j - X_j'(\widehat{\beta}_{(i)} + \eta_i))\right] + \tau\eta_i \;.$$

By the mean-value theorem, we also have

$$\psi(\epsilon_j - X_j'\widehat{\beta}_{(i)}) - \psi(\epsilon_j - X_j'(\widehat{\beta}_{(i)} + \eta_i)) = \psi'(\tilde{r}_{j,(i)})X_j'\eta_i + \left[\psi'(\gamma^*(X_j, \widehat{\beta}_{(i)}, \eta_i)) - \psi'(\tilde{r}_{j,(i)})\right] X_j'\eta_i \;,$$

where $\gamma^*(X_j, \widehat{\beta}_{(i)}, \eta_i)$ is in the ("unordered") interval $(\epsilon_j - X_j'\widehat{\beta}_{(i)}, \epsilon_j - X_j'(\widehat{\beta}_{(i)} + \eta_i))$, i.e $(\tilde{r}_{j,(i)}, \tilde{r}_{j,(i)} - X_j'\eta_i)$.

Hence, if $\mathcal{R}_i$ is the quantity defined in Equation (20),

$$\frac{1}{n}\sum_{j\neq i}X_j\left[\psi(\epsilon_j - X_j'\widehat{\beta}_{(i)}) - \psi(\epsilon_j - X_j'(\widehat{\beta}_{(i)} + \eta_i))\right] = \frac{1}{n}\sum_{j\neq i}\psi'(\tilde{r}_{j,(i)})X_jX_j'\eta_i + \mathcal{R}_i \;,$$
$$= S_i\eta_i + \mathcal{R}_i \;.$$

In light of the previous simplifications, we have, using

$$f(\beta) = f_i(\beta) - \frac{1}{n} X_i \psi(\epsilon_i - X_i'\beta) \text{ and } f_i(\widehat{\beta}_{(i)}) = 0 \;,$$

the equality

$$f(\widetilde{\beta}_i) = -\frac{1}{n} X_i \psi(\epsilon_i - X_i'\widetilde{\beta}_i) + (S_i + \tau \mathrm{Id})\eta_i + \mathcal{R}_i \;.$$

Since by definition, $\eta_i = \frac{1}{n}(S_i + \tau\mathrm{Id})^{-1} X_i \psi(\mathrm{prox}_{c_i}(\rho)(\tilde{r}_{i,(i)}))$,

$$(S_i + \tau\mathrm{Id})\eta_i = \frac{1}{n} X_i \psi(\mathrm{prox}_{c_i}(\rho)(\tilde{r}_{i,(i)})) \;.$$

In other respects,

$$\epsilon_i - X_i'\widetilde{\beta}_i = \tilde{r}_{i,(i)} - c_i \psi(\mathrm{prox}_{c_i}(\rho)(\tilde{r}_{i,(i)})) \;.$$

When ρ is differentiable, $x - c\psi(\mathrm{prox}_c(\rho)(x)) = \mathrm{prox}_c(\rho)(x)$ almost by definition of the proximal mapping (see Lemma A-1 and its proof). Therefore, $\epsilon_i - X_i'\widetilde{\beta}_i = \mathrm{prox}_{c_i}(\rho)(\tilde{r}_{i,(i)})$ and

$$-\frac{1}{n} X_i \psi(\epsilon_i - X_i'\widetilde{\beta}_i) + (S_i + \tau\mathrm{Id})\eta_i = \frac{1}{n} X_i \left[-\psi(\mathrm{prox}_{c_i}(\rho)(\tilde{r}_{i,(i)})) + \psi(\mathrm{prox}_{c_i}(\rho)(\tilde{r}_{i,(i)})) \right] = 0.$$

We conclude that

$$f(\widetilde{\beta}_i) = \mathcal{R}_i \;.$$

Applying Lemma 2.1, we see that

$$\|\widehat{\beta} - \widetilde{\beta}_i\| \leq \frac{1}{\tau} \|\mathcal{R}_i\| \;.$$

□

3.1.1 On $\mathcal{R}_i$

Lemma 3.1. *We have*

$$\|\eta_i\| \leq \frac{1}{\sqrt{n}\tau} \frac{\|X_i\|}{\sqrt{n}} \left[|\psi(\tilde{r}_{i,(i)})| \wedge \frac{|\tilde{r}_{i,(i)}|}{c_i} \right] \;, \tag{21}$$

and

$$\|\mathcal{R}_i\| \leq |||\widehat{\Sigma}|||_2 \sup_{j \neq i} \left| \psi'(\gamma^*(X_j, \widehat{\beta}_{(i)}, \eta_i)) - \psi'(\tilde{r}_{j,(i)}) \right| \frac{1}{\sqrt{n}\tau} \frac{\|X_i\|}{\sqrt{n}} \left[|\psi(\tilde{r}_{i,(i)})| \wedge |\tilde{r}_{i,(i)}|/c_i \right] \;. \tag{22}$$

Proof. We have

$$\mathcal{R}_i = \frac{1}{n} \sum_{j \neq i} \left[\psi'(\gamma^*(X_j, \widehat{\beta}_{(i)}, \eta_i)) - \psi'(\tilde{r}_{j,(i)}) \right] X_j X_j' \eta_i \;.$$

Of course, $\mathcal{S} = \frac{1}{n} \sum_{j \neq i} \left[\psi'(\gamma^*(X_j, \widehat{\beta}_{(i)}, \eta_i)) - \psi'(\tilde{r}_{j,(i)}) \right] X_j X_j'$ can be written $\mathcal{S} = \frac{1}{n} X' \mathsf{D} X$, where D is a diagonal matrix with (j,j) entry $\left[\psi'(\gamma^*(X_j, \widehat{\beta}_{(i)}, \eta_i)) - \psi'(\tilde{r}_{j,(i)}) \right]$ and (i,i) entry 0.

Using the fact that $||| \cdot |||_2$ is a matrix norm, we see that $|||\mathcal{S}|||_2 \leq |||\widehat{\Sigma}|||_2 |||\mathsf{D}|||_2$. This implies that

$$\|\mathcal{R}_i\| \leq |||\widehat{\Sigma}|||_2 \sup_{j \neq i} \left| \psi'(\gamma^*(X_j, \widehat{\beta}_{(i)}, \eta_i)) - \psi'(\tilde{r}_{j,(i)}) \right| \|\eta_i\| \;,$$

where $\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n X_i X_i'$ is the usual sample covariance matrix.

We note that, since $|||(S_i + \tau \mathrm{Id}_p)^{-1}|||_2 \leq \tau^{-1}$,

$$\|\eta_i\| \leq \frac{1}{\sqrt{n}\tau} \frac{\|X_i\|}{\sqrt{n}} |\psi(\mathrm{prox}_{c_i}(\rho)(\tilde{r}_{i,(i)}))| \;.$$

Using Lemma A-1, we see that

$$|\psi(\mathrm{prox}_{c_i}(\rho)(\tilde{r}_{i,(i)}))| \leq |\psi(\tilde{r}_{i,(i)})| \wedge \frac{|\tilde{r}_{i,(i)}|}{c_i} \;.$$

The lemma is shown. □

3.1.2 On $\gamma^*(X_j, \widehat{\beta}_{(i)}, \eta_i)$ and related quantities

We now show how to control $\frac{1}{\sqrt{n}} \sup_{j \neq i} \left|\psi'(\gamma^*(X_j, \widehat{\beta}_{(i)}, \eta_i)) - \psi'(\tilde{r}_{j,(i)})\right|$

Lemma 3.2. *Let us call*

$$\mathcal{B}_n(i) = \sup_{j \neq i} \left[|\epsilon_j - X_j' \widehat{\beta}_{(i)}| + |X_j' \eta_i|\right] \ .$$

*Suppose, as in our assumption **O3**, that ψ' is $L(\mathcal{B}_n(i))$ Lipschitz on $(-\mathcal{B}_n(i), \mathcal{B}_n(i))$. Then,*

$$\sup_{j \neq i} \left|\psi'(\gamma^*(X_j, \widehat{\beta}_{(i)}, \eta_i)) - \psi'(\tilde{r}_{j,(i)})\right| \leq L(\mathcal{B}_n(i)) \sup_{j \neq i} |X_j' \eta_i| \ .$$

It follows that

$$\|\mathcal{R}_i\| \leq \sup_{j \neq i} |X_j' \eta_i| \frac{L(\mathcal{B}_n(i))}{\sqrt{n}\tau} \frac{\|X_i\|}{\sqrt{n}} |||\widehat{\Sigma}|||_2 \left[|\psi(\tilde{r}_{i,(i)})| \wedge |\tilde{r}_{i,(i)}|/c_i\right] \ .$$

We note that we could replace the assumption concerning the Lipschitz property of ψ' on $(-\mathcal{B}_n(i), \mathcal{B}_n(i))$ by saying that ψ' has modulus of continuity ω_n when restricted to this interval and putting growth condition on this modulus. We chose not to do this to simplify the exposition.

Proof. By definition, we have

$$|\gamma^*(X_j, \widehat{\beta}_{(i)}, \eta_i) - \tilde{r}_{j,(i)}| \leq |X_j' \eta_i| \ .$$

Therefore,

$$\sup_{j \neq i} |\gamma^*(X_j, \widehat{\beta}_{(i)}, \eta_i)| \leq \sup_{j \neq i} \left[|\epsilon_j - X_j' \widehat{\beta}_{(i)}| + |X_j' \eta_i|\right]$$

Recall that

$$\mathcal{B}_n(i) = \sup_{j \neq i} \left[|\epsilon_j - X_j' \widehat{\beta}_{(i)}| + |X_j' \eta_i|\right] \ .$$

ψ' is $L(\mathcal{B}_n(i))$-Lipschitz on $(-\mathcal{B}_n(i), \mathcal{B}_n(i))$ by assumption. Therefore,

$$\sup_{j \neq i} \left|\psi'(\gamma^*(X_j, \widehat{\beta}_{(i)}, \eta_i)) - \psi'(\tilde{r}_{j,(i)})\right| \leq L(\mathcal{B}_n(i)) \sup_{j \neq i} |X_j' \eta_i| \ .$$

The bound for $\|\mathcal{R}_i\|$ follows immediately.

□

3.2 Probabilistic aspects

Note that

$$X_j' \eta_i = \psi(\text{prox}_{c_i}(\rho)(\tilde{r}_{i,(i)})) \frac{1}{n} X_j' (S_i + \tau \text{Id}_p)^{-1} X_i \ .$$

We can rewrite the bound on $\|\mathcal{R}_i\|$ as

$$\|\mathcal{R}_i\| \leq \left[\sup_{j \neq i} \frac{|X_j'(S_i + \tau \text{Id}_p)^{-1} X_i|}{n}\right] \frac{L(\mathcal{B}_n(i))}{\sqrt{n}\tau} \frac{\|X_i\|}{\sqrt{n}} |||\widehat{\Sigma}|||_2 \left(\left[|\psi(\tilde{r}_{i,(i)})| \wedge |\tilde{r}_{i,(i)}|/c_i\right] |\psi(\text{prox}_{c_i}(\rho)(\tilde{r}_{i,(i)})|\right) \ .$$

In light of Proposition 3.1, the bound on $\|\mathcal{R}_i\|$ is encouraging since it shows that we can control $\|\widehat{\beta} - \widetilde{\beta}_i\|$ in L_k provided we can control each terms in the above product in L_{5k}: indeed, for a product of m random variables $\{W_j\}_{j=1}^m$, we have $\mathbf{E}\left(|\prod_{j=1}^m W_j|\right) \leq \prod_{j=1}^m \left[\mathbf{E}\left(|W_j|^m\right)\right]^{1/m}$ by Hölder's inequality. In particular, we will later need control of $\mathbf{E}\left(\|\widehat{\beta} - \widetilde{\beta}_i\|^2\right)$ and will therefore require subsequent bounds to in L_{10}.

3.2.1 On $\sup_{j\neq i}|X_j'(S_i+\tau\mathrm{Id})^{-1}X_i/n|$

We will control $X_j'(S_i+\tau\mathrm{Id})^{-1}X_i/n$ by appealing to Lemma B-2, which is designed to handle problems of the kind we are encountering here.

Lemma 3.3. *Suppose X_i are independent and satisfy the concentration assumptions mentioned in Assumption **O4**. Then*

$$\sup_{j\neq i}\frac{\|X_j\|}{\tau\sqrt{n}}=\mathrm{O}_{L_{20}}(1)$$

and

$$\sup_{j\neq i}|X_j'(S_i+\tau\mathrm{Id})^{-1}X_i/n|=\mathrm{O}_{L_{10}}\left(\frac{polyLog(n)}{\mathsf{c}_n^{1/2}}\right).$$

Proof. • **First part** Using the fact that $X_j\to\|X_j\|/\sqrt{n}$ is $n^{-1/2}$-Lipschitz with respect to Euclidean norm we see, using Lemma B-1, that

$$\sup_{j\neq i}|\|X_j\|/\sqrt{n}-m_{\|X_j\|/\sqrt{n}}|\leq\mathrm{polyLog}(n)/(\sqrt{n\mathsf{c}_n})\text{ in }L_{20}.$$

Recall that $\mathrm{cov}\,(X_i)=\mathrm{Id}_p$. So $m_{\|X_j\|/\sqrt{n}}$ is of order 1 in the case we are interested in (using for instance Proposition 1.9 in Ledoux (2001)). Hence,

$$\sup_{j\neq i}|\|X_j\|/\sqrt{n}|=\mathrm{O}_{L_{20}}(1),$$

provided $1/(n\mathsf{c}_n)=\mathrm{O}(\mathrm{polyLog}(n))$. This is clearly the case under our assumptions. This shows the first result of the Lemma.

• **Second part** Let us work conditionally on $X_{(i)}=\{X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n\}$. Call $v_{j,(i)}=(S_i+\tau\mathrm{Id})^{-1}X_j$. The map $F_j(X_i)=X_j'(S_i+\tau\mathrm{Id})^{-1}X_i=X_i'v_{j,(i)}$ is Lipschitz (as a function of X_i) with Lipschitz constant $\sqrt{X_j'(S_i+\tau\mathrm{Id})^{-2}X_j}\leq\|X_j\|/\tau$. Indeed, it is linear in X_i. Call m_{F_j} its mean, conditional on $X_{(i)}$. Since X_i has mean 0, we see that $m_{F_j}=0$.

Therefore, using Lemma B-2, we see that

$$\left[\mathbf{E}\left(\left[\frac{1}{n}\sup_{j\neq i}|X_j'(S_i+\tau\mathrm{Id})^{-1}X_i|\right]^{10}\right)\right]^{1/10}=\mathrm{O}\left(\left[\mathbf{E}\left(\left[\frac{1}{\sqrt{n}}\sup_j\frac{\|X_j\|}{\tau\sqrt{n}}\right]^{20}\right)\right]^{1/20}\sqrt{\mathrm{polyLog}(n)/\mathsf{c}_n}\right),$$

and therefore

$$\frac{1}{n}\sup_{j\neq i}|X_j'(S_i+\tau\mathrm{Id})^{-1}X_i|=\mathrm{O}_{L_{10}}\left(\sqrt{\mathrm{polyLog}(n)/\mathsf{c}_n}\right),$$

since we have established earlier that $\sup_{j\neq i}|\|X_j\|/\sqrt{n}|=\mathrm{O}_{L_{20}}(1)$. □

3.2.2 Control of the residuals R_i and $\tilde{r}_{i,(i)}$

Our aim here is to show that we can control $\sup_i|R_i|$, where $R_i=\epsilon_i-X_i'\widehat{\beta}$ are the residuals from the full robust ridge-regression problem. This will allow us to achieve control of $\mathcal{B}_n(i)$. As $\tilde{r}_{i,(i)}$ is much easier to understand than R_i, our strategy is to relate the two.

Lemma 3.4. *We have the deterministic bound*

$$|R_i|\leq|\tilde{r}_{i,(i)}|+\frac{\|X_i\|^2}{n}\frac{1}{\tau}|\psi(\tilde{r}_{i,(i)})|. \tag{23}$$

Denoting by $\mathcal{E}_n=\sup_{1\leq i\leq n}|\epsilon_i|$, we have under our assumptions on $\{X_i\}_{i=1}^n$,

$$\sup_{1\leq i\leq n}|\tilde{r}_{i,(i)}|\leq\mathcal{E}_n+[\|W_n\|+\frac{1}{n}\sup_{1\leq i\leq n}\|X_i\||\psi(\mathcal{E}_n)\vee\psi(-\mathcal{E}_n)|]polyLog(n)/[\tau\sqrt{\mathsf{c}_n}]\ \text{in } L_k.$$

*Under the assumption (see **O3**) that $|\psi(x)| = \mathrm{O}(|x|^m)$ for some fixed m at infinity, we have, for some constant K,*

$$\sup_i |R_i| \leq K(\sup_i |\tilde{r}_{i,(i)}|)^{m\vee 1} \text{ in } L_k \;,$$

and $\|W_n\| + \frac{1}{n}\sup_{1\leq i\leq n}\|X_i\||\psi(\mathcal{E}_n) \vee \psi(-\mathcal{E}_n)| = \mathrm{O}_{L_k}(\|W_n\| + \textit{polyLog}(n)\mathcal{E}_n^m/\sqrt{n})$.

Proof. Recall the representation

$$\beta_1 - \beta_2 = (\mathsf{S}_{\beta_1,\beta_2} + \tau \mathrm{Id}_p)^{-1}\left(f(\beta_1) - f(\beta_2)\right) \;.$$

Take $\beta_1 = \widehat{\beta}$ and $\beta_2 = \widehat{\beta}_{(i)}$. Note that since $f_i(\widehat{\beta}_{(i)}) = 0$,

$$f(\widehat{\beta}_{(i)}) = -\frac{1}{n}X_i\psi(\tilde{r}_{i,(i)}) - \frac{1}{n}\sum_{j\neq i} X_i\psi(\tilde{r}_{j,(i)}) + \tau\widehat{\beta}_{(i)} = -\frac{1}{n}X_i\psi(\tilde{r}_{i,(i)}) \;.$$

Therefore,

$$\widehat{\beta} - \widehat{\beta}_{(i)} = \frac{1}{n}(\mathsf{S}_{\widehat{\beta},\widehat{\beta}_{(i)}} + \tau\mathrm{Id}_p)^{-1}X_i\psi(\tilde{r}_{i,(i)}) \;.$$

Since $\tilde{r}_{i,(i)} - R_i = X_i'(\widehat{\beta} - \widehat{\beta}_{(i)})$, we also have

$$|\tilde{r}_{i,(i)} - R_i| \leq \frac{\|X_i\|^2}{n}\frac{1}{\tau}|\psi(\tilde{r}_{i,(i)})| \;.$$

We conclude that

$$|R_i| \leq |\tilde{r}_{i,(i)}| + \frac{\|X_i\|^2}{n}\frac{1}{\tau}|\psi(\tilde{r}_{i,(i)})| \;.$$

Now under assumptions, we have $\sup_{1\leq i\leq n}|\|X_i\|^2/n - \frac{p}{n}| = \mathrm{O}_{L_k}(\mathrm{polyLog}(n)/\sqrt{n\mathsf{c}_n})$, according to Lemma B-3. Using the fact that $\|\widehat{\beta}_{(i)}\| \leq \|W_{n,(i)}\|/\tau$ (see Lemma 2.2 with obvious modifications of notations), the independence of X_i and $\widehat{\beta}_{(i)}$, we have, through Lemma B-2,

$$\sup_{1\leq i\leq n}|X_i'\widehat{\beta}_{(i)}| \leq \sup_{1\leq i\leq n}\frac{\|W_{n,(i)}\|}{\tau}\mathrm{polyLog}(n)/\sqrt{\mathsf{c}_n} \;.$$

Since $\|W_{n,(i)}\| \leq \|W_n\| + \|X_i\||\psi(\epsilon_i)|/n$, we have

$$|\tilde{r}_{i,(i)}| \leq |\epsilon_i| + [\|W_n\| + \sup_i\|X_i\||\psi(\epsilon_i)|/n)]\mathrm{polyLog}(n)/[\tau\sqrt{\mathsf{c}_n}] \text{ in } L_k \;.$$

Denoting by $\mathcal{E}_n = \sup_{1\leq i\leq n}|\epsilon_i|$, we have, using the fact that ψ is non-decreasing,

$$\sup_{1\leq i\leq n}|\tilde{r}_{i,(i)}| \leq \mathcal{E}_n + \left[\|W_n\| + \frac{1}{n^{1/2}}\sup_{1\leq i\leq n}\frac{\|X_i\|}{n^{1/2}}|\psi(\mathcal{E}_n)\vee\psi(-\mathcal{E}_n)|\right]\mathrm{polyLog}(n)/[\tau\sqrt{\mathsf{c}_n}] \text{ in } L_k \;,$$

for any given k. We note that if $|\psi(x)| = \mathrm{O}(|x|^m)$ at ∞, since we have shown that

$$|R_i| \leq |\tilde{r}_{i,(i)}| + \frac{\|X_i\|^2}{n}\frac{1}{\tau}|\psi(\tilde{r}_{i,(i)})| \;,$$

we have the bound $\sup_{1\leq i\leq n}|R_i| \lesssim \sup_{1\leq i\leq n}|\tilde{r}_{i,(i)}|^{m\vee 1}$ and therefore,

$$\sup_{1\leq i\leq n}|R_i| \lesssim \left[\mathcal{E}_n + \mathrm{polyLog}(n)/[\tau\sqrt{\mathsf{c}_n}][\|W_n\| + \frac{1}{n}\sup_{1\leq i\leq n}\|X_i\||\psi(\mathcal{E}_n)\vee\psi(-\mathcal{E}_n)|\right]^{m\vee 1} \text{ in } L_k \;,$$

provided the bound on $\sup_{1\leq i\leq n}|\tilde{r}_{i,(i)}|$ holds in L_{mk}. Note that this is guaranteed under our assumptions. Of course, here we are using control of $\sup_i\|X_i\|^2/n$, which we get by controlling $\|X_i\|/\sqrt{n}$ through concentration arguments. The fact that $\sup_i\|X_i\|/\sqrt{n} = \mathrm{O}_{L_k}(1)$ gives us the last statement of the lemma. □

Remark 1: at the gist of the bound on $\tilde{r}_{i,(i)}$ is a uniform bound on $\|\widehat{\beta}_{(i)}\|$ in L_k. We could also have used the bound $\sup_i \|\widehat{\beta}_{(i)}\| \leq \sqrt{2/\tau}\sqrt{1/n\sum_{i=1}^n \rho(\epsilon_i)}$ which is immediate from Lemma 2.2. This would change slightly the appearance of our bounds on $\sup_i |\tilde{r}_{i,(i)}|$.

In the case where ρ grows like $|x|^{1+\epsilon}$ at infinity, it seems preferable to have bounds that depend on $\psi(\epsilon_i)$ and not $\rho(\epsilon_i)$, which is why we demonstrated how to use those ψ-based bounds, instead of using the slightly simpler ones based on ρ. This difference will likely be more important when $\mathcal{E}_n$ is allowed to grow much faster with n than under our assumption **O7** - but we leave this variant of statistical interest to another paper.

Remark 2: We note that a similar result holds of course for $\tilde{r}_{j,(i)}$. More precisely,

$$|\tilde{r}_{j,(i)} - R_j| \leq \left|\frac{1}{n} X_j' (\mathsf{S}_{\widehat{\beta},\widehat{\beta}_{(i)}} + \tau \mathrm{Id}_p)^{-1} X_i\right| \left|\psi(\tilde{r}_{i,(i)})\right| ,$$

and hence,

$$|\tilde{r}_{j,(i)} - R_j| \leq \frac{\|X_j\|\|X_i\|}{n\tau} \left|\psi(\tilde{r}_{i,(i)})\right| .$$

Of course, this bound is very coarse and we will see that we can get a better one later.

However, this finally allows us to have the following proposition

Proposition 3.2. *Under the assumption that $|\psi(x)| = \mathrm{O}(|x|^m)$, as in **O3**, we have the bound*

$$\mathcal{B}_n(i) \leq K \left[\mathcal{E}_n + (\|W_n\| + \frac{\mathcal{E}_n^m}{\sqrt{n}}) \textit{polyLog}(n)/[\tau\sqrt{\mathsf{c}_n}]\right]^{m\vee 1} \text{ in } L_k ,$$

where K is a constant independent of p and n. When $\|W_n\|$ and $\frac{\mathcal{E}_n^m}{\sqrt{n}}$ are bounded in L_k, this bound simply becomes

$$\mathcal{B}_n(i) \leq K \left[\mathcal{E}_n \vee \textit{polyLog}(n)/(\tau\sqrt{\mathsf{c}_n})\right]^{m\vee 1} \text{ in } L_k .$$

The same bound holds for $\sup_i \mathcal{B}_n(i)$ in L_k .

Proof. The result follows easily from the fact that

$$\mathcal{B}_n(i) = \sup_{j\neq i} \left[|\tilde{r}_{j,(i)}| + |X_j'\eta_i|\right] ,$$

the fact that

$$\sup_i \sup_{j\neq i} |\tilde{r}_{j,(i)} - R_j| \leq \sup_i \sup_j \frac{\|X_j\|\|X_i\|}{n\tau} \left|\psi(\tilde{r}_{i,(i)})\right| ,$$

and the bounds on $\sup_i |R_i|$ we have derived earlier. In more details, we have

$$\mathcal{B}_n(i) \leq \sup_i R_i + \sup_i \sup_{j\neq i} |\tilde{r}_{j,(i)} - R_j| + \sup_i \sup_{j\neq i} |X_j'\eta_i| .$$

The same bound is of course true for $\sup_i \mathcal{B}_n(i)$. Now we recall that

$$X_j'\eta_i = \frac{X_j'(S_i + \tau\mathrm{Id})^{-1}X_i}{n} \psi(\mathrm{prox}_{c_i}(\rho)(\tilde{r}_{i,(i)})) .$$

Using our previous investigations concerning $X_j'(S_i + \tau\mathrm{Id}_p)^{-1}X_i$, the part concerning $\sup_i \sup_{j\neq i} |X_j'\eta_i|$ is easily shown to be negligible compared to the bound on

$$\sup_i \sup_{j\neq i} |\tilde{r}_{j,(i)} - R_j| .$$

□

3.2.3 Consequences

We have the following result. Recall that ψ' is assumed to be Lipschitz with Lipschitz constant $L(u)$ on $(-|u|, |u|)$.

Proposition 3.3. *Suppose, as is consistent with **O3**, **O6** and **O7**, that $|\psi(x)| = \mathrm{O}(|x|^m)$, $\|W_n\|$ is bounded in L_k and $\mathcal{E}_n^m = \mathrm{o}(\sqrt{n})$ in L_k. Suppose further that $L(x) \leq K|x|^{m_1}$. Then we have*

$$\|\mathcal{R}_i\| \leq K \frac{polyLog(n)}{n\tau^2 \mathsf{c}_n^{3/2}} \left(\mathcal{E}_n \vee (\tau \mathsf{c}_n)^{-1/2} polyLog(n)\right)^{2m + m_1(m\vee 1)} \text{ in } L_k \ .$$

In particular, if $\mathcal{E}_n = \mathrm{O}_{L_k}(polyLog(n))$ and $1/\mathsf{c}_n = \mathrm{O}(polyLog(n))$, we have, since τ is assumed to be fixed,

$$\|\mathcal{R}_i\| \leq K \frac{polyLog(n)}{n} \text{ in } L_k \ .$$

Furthermore, the same bounds hold for $\sup_i \|\mathcal{R}_i\|$.

Proof. The proof follows by aggregating all the intermediate results we had and noticing that under our assumptions, $|||\widehat{\Sigma}|||_2 = \mathrm{O}_{L_k}(\mathsf{c}_n^{-1})$. This latter result follows easily from a standard ϵ-net and union bound argument for controlling $|||\widehat{\Sigma}|||_2$ - see e.g Talagrand (2003), Appendix A.4. We provide some details on this bound in Lemma B-4.

The statement concerning $\sup_i \|\mathcal{R}_i\|$ follows by the same method. □

We have the following theorem, which is very important for this paper. We recall that $\widetilde{\beta}_i$ is defined in Equation (4) on p.7.

Theorem 3.1. *Under Assumptions **O1-O7**, we have, for any fixed k, when τ is held fixed,*

$$\sup_{1\leq i\leq n} \|\widehat{\beta} - \widetilde{\beta}_i\| = \mathrm{O}_{L_k}(\frac{polyLog(n)}{n}) \ .$$

In particular, we have

$$\forall 1 \leq i \leq n \ , \mathbf{E}\left(\|\widehat{\beta} - \widetilde{\beta}_i\|^2\right) = \mathrm{O}(polyLog(n)/n^2) \ .$$

Also,

$$\sup_{1\leq i\leq n} \sup_{j\neq i} |\tilde{r}_{j,(i)} - R_j| = \mathrm{O}_{L_k}(\frac{polyLog(n)}{n^{1/2}}) \ .$$

Finally,

$$\sup_i |R_i - prox_{c_i}(\rho)(\tilde{r}_{i,(i)})| = \mathrm{O}_{L_k}(\frac{polyLog(n)}{n^{1/2}}) \ .$$

Proof. The only parts that may require a discussion are the ones involving the residuals. However, they follow easily from the very coarse bound

$$\begin{aligned}\sup_{j\neq i} |\tilde{r}_{j,(i)} - R_j| = \sup_{j\neq i} \left|X_j'(\widehat{\beta} - \widehat{\beta}_i)\right| &\leq \sup_{j\neq i} \left|X_j'(\widehat{\beta} - \widetilde{\beta}_i)\right| + \sup_{j\neq i} |X_j'(\widetilde{\beta}_i - \widehat{\beta}_i)| \ , \\ &\leq \left(\sup_{1\leq j\leq n} \frac{\|X_j\|}{\sqrt{n}}\right)\sqrt{n}\|\widehat{\beta} - \widetilde{\beta}_i\| + \sup_{j\neq i} |X_j'\eta_i| \ ,\end{aligned}$$

and the fact that $\left(\sup_{1\leq j\leq n} \frac{\|X_j\|}{\sqrt{n}}\right) = \mathrm{O}_{L_k}(1)$ under our assumptions. Recalling that $\|\widehat{\beta} - \widetilde{\beta}_i\| \leq \|\mathcal{R}_i\|/\tau$ and hence $\sup_i \|\widehat{\beta} - \widetilde{\beta}_i\| \leq \sup_i \|\mathcal{R}_i\|/\tau$ gives control of the first term. Control of the second term follows from Lemma 3.3 and our bounds on $\sup_i |\tilde{r}_{i,(i)}|$ in Lemma 3.4.

Concerning the fine approximation of R_i, recall that

$$R_i = \epsilon_i - X_i'\widehat{\beta} = \epsilon_i - X_i'\widetilde{\beta}_i - X_i'(\widehat{\beta} - \widetilde{\beta}_i) \ .$$

Now, given the definition of $\widetilde{\beta}_i$, we have

$$X_i'\widetilde{\beta}_i = X_i'\widehat{\beta}_{(i)} + c_i\psi[\text{prox}_{c_i}(\rho)(\tilde{r}_{i,(i)})] .$$

Hence,

$$\epsilon_i - X_i'\widetilde{\beta}_i = \tilde{r}_{i,(i)} - c_i\psi[\text{prox}_{c_i}(\rho)(\tilde{r}_{i,(i)})] = \text{prox}_{c_i}(\rho)(\tilde{r}_{i,(i)}) ,$$

where the last equality is a standard property of the proximal mapping (see Lemma A-1 if needed). So we have established that

$$\sup_i \left|R_i - \text{prox}_{c_i}(\rho)(\tilde{r}_{i,(i)})\right| = \sup_i \left|X_i'(\widetilde{\beta}_i - \widehat{\beta})\right|$$

and the result follows from our previous bounds. □

3.3 Asymptotically deterministic character of $\|\widehat{\beta}\|^2$

Proposition 3.4. *Under our assumptions **O1-O7**,*

$$\text{var}\left(\|\widehat{\beta}\|^2\right) \to 0 \textit{ as } n \to \infty .$$

Therefore $\|\widehat{\beta}\|^2$ has a deterministic equivalent in probability and in L_2.

More specifically, when $1/\mathsf{c}_n = \mathrm{O}(polyLog(n))$, we have

$$\text{var}\left(\|\widehat{\beta}\|^2\right) = \mathrm{O}(\frac{polyLog(n)}{n}) .$$

Proof. We will use the Efron-Stein inequality - a martingale inequality - to show that var $\left(\|\widehat{\beta}\|^2\right)$ goes to 0 as $n \to \infty$. In what follows, we rely on our assumptions, which imply that $\psi(\epsilon_i)$ have enough moments for all the expectations of the type $\mathbf{E}\left(\|\widehat{\beta}\|^{2k}\right)$ to be bounded like $1/\tau^{2k}$. Note that this the content of our Lemma 2.2.

Recall that the Efron-Stein inequality (Efron and Stein (1981)) gives, if Y is a function of n independent random variables, and $Y_{(i)}$ is any function of all those random variables except the i-th,

$$\text{var}\left(Y\right) \leq \sum_{i=1}^n \text{var}\left(Y - Y_{(i)}\right) \leq \sum_{i=1}^n \mathbf{E}\left((Y - Y_{(i)})^2\right) .$$

We first observe that

$$\mathbf{E}\left(|\|\widehat{\beta}\|^2 - \|\widehat{\beta}_{(i)}\|^2|^2\right) \leq 2\left[\mathbf{E}\left(|\|\widehat{\beta}\|^2 - \|\widetilde{\beta}_i\|^2|^2\right) + \mathbf{E}\left(|\|\widetilde{\beta}_i\|^2 - \|\widehat{\beta}_{(i)}\|^2|^2\right)\right] .$$

Of course, using the fact that $\widehat{\beta} = \widehat{\beta} - \widetilde{\beta}_i + \widetilde{\beta}_i$ and $|\|\widehat{\beta}\|^2 - \|\widetilde{\beta}_i\|^2|^2 = [(\widehat{\beta} - \widetilde{\beta}_i)'(\widehat{\beta} + \widetilde{\beta}_i)]^2$, $|\|\widehat{\beta}\|^2 - \|\widetilde{\beta}_i\|^2|^2 = \mathrm{O}_{L_1}(\|\widehat{\beta} - \widetilde{\beta}_i\|^4) + \sqrt{\mathrm{O}_{L_1}(\|\widehat{\beta} - \widetilde{\beta}_i\|^4)}$, by the Cauchy-Schwarz inequality, since $\mathbf{E}\left(\|\widehat{\beta}\|^2\right)$ exists and is bounded by K/τ^2.

Using the results of Theorem 3.1, we see that

$$\mathbf{E}\left(|\|\widehat{\beta}\|^2 - \|\widetilde{\beta}_i\|^2|^2\right) = \mathrm{O}(\frac{\text{polyLog}(n)}{n^2}) = \mathrm{o}(n^{-1}) .$$

On the other hand, given the definition in Equation (4),

$$\|\widetilde{\beta}_i\|^2 - \|\widehat{\beta}_{(i)}\|^2 = 2\frac{1}{n}\widehat{\beta}_{(i)}'(S_i + \tau\text{Id})^{-1}X_i\psi(\text{prox}_{c_i}(\rho)(\tilde{r}_{i,(i)})) + \frac{1}{n^2}X_i'(S_i + \tau\text{Id})^{-2}X_i\psi^2(\text{prox}_{c_i}(\rho)(\tilde{r}_{i,(i)})) .$$

Since $\widehat{\beta}_{(i)}$ and S_i are independent of X_i, and $|||(S_i+\tau\text{Id})^{-1}|||_2 \leq 1/\tau$, $\widehat{\beta}_{(i)}'(S_i+\tau\text{Id})^{-1}X_i = \mathrm{O}_{L_4}(\|\widehat{\beta}_{(i)}\|/\mathsf{c}_n^{1/2})$, using our concentration assumptions on X_i (**O4**) applied to linear forms. Therefore, we see that both terms are $\mathrm{O}_{L_2}(1/n\mathsf{c}_n^{1/2})$ provided $\psi(\text{prox}_{c_i}(\rho)(\tilde{r}_{i,(i)}))$ has $4+\epsilon$ absolute moments - uniformly bounded in n - by using Hölder's inequality. Under our assumptions, given our work on $\tilde{r}_{i,(i)}$, the fact that the prox is a contractive mapping (Moreau (1965)) and that we assume that $\text{sign}(\psi(x)) = \text{sign}(x)$, it is clear that this is the case. We conclude that then

$$\mathbf{E}\left(\left|\|\widetilde{\beta}_i\|^2 - \|\widehat{\beta}_{(i)}\|^2\right|^2\right) = \mathrm{O}(\frac{\text{polyLog}(n)}{n^2}) \ .$$

Taking $Y = \|\widehat{\beta}\|^2$ and $Y_{(i)} = \|\widehat{\beta}_{(i)}\|^2$ in the Efron-Stein inequality, we clearly see that

$$\text{var}\left(\|\widehat{\beta}\|^2\right) = \mathrm{O}(\frac{\text{polyLog}(n)}{n}) = \mathrm{o}(1) \ .$$

This shows that $\|\widehat{\beta}\|^2$ has a deterministic equivalent in probability and in L_2. □

4 Leaving out a predictor

In this second main step of the proof, we do need at various points that the entries of the data vector X_i be independent, whereas as we showed before, it is not important when studying what happens when we leave out an observation.

We refer the reader to Subsubsection 1.2.2 for the definition of the various quantities that appear in the current section.

We will show later, in Subsubsection 4.2.2 that $\xi_n \geq 0$. However, we will use this information from the beginning and there are no circular arguments. Note that when $\xi_n > 0$, we have, with the definitions introduced in Subsubsection 1.2.2,

$$\mathfrak{b}_p = \frac{\frac{1}{n}\sum_{i=1}^n X_i(p)\psi(r_{i,[p]}) - \tau\mathfrak{b}_p}{\frac{1}{n}\sum_{i=1}^n X_i^2(p)\psi'(r_{i,[p]}) - u_p'(\mathfrak{S}_p + \tau\text{Id})^{-1}u_p} = \frac{n^{-1/2}N_p - \tau\mathfrak{b}_p}{\xi_n} \ .$$

The aim of our work in the second part of this proof is to establish Theorem 4.1 on p.33, which shows that $\|\widetilde{b} - \widehat{\beta}\| = \mathrm{O}(\text{polyLog}(n)/n)$ in L_k. Because the last coordinate of $\widetilde{b}$, $\mathfrak{b}_p$, has a reasonably simple probabilistic structure and our approximations are sufficiently good, we will be able to transfer our insights about this coordinate to $\widehat{\beta}_p$, the last coordinate of $\widehat{\beta}$.

Appendix D-1 provides some intuitive explanations for why $\mathfrak{b}_p$ is a natural quantity in our context.

4.1 Deterministic aspects

Proposition 4.1. *Recall the definition of $\widetilde{b}$ in Equation* (11). *We have*

$$\|\widehat{\beta} - \widetilde{b}\| \leq \frac{1}{\tau}|\mathfrak{b}_p| \sup_{1\leq i\leq n} |\mathsf{d}_{i,p}| \, |||\widehat{\Sigma}|||_2 \sqrt{\|(\mathfrak{S}_p + \tau\text{Id})^{-1}u_p\|^2 + 1} \ . \tag{24}$$

where $\mathsf{d}_{i,p} = [\psi'(\gamma_{i,p}^*) - \psi'(r_{i,[p]})]$ and $\gamma_{i,p}^*$ is in the interval $(\epsilon_i - V_i'\widehat{\gamma}, \epsilon_i - X_i'\widetilde{b})$.

Furthermore,

$$\|(\mathfrak{S}_p + \tau\text{Id})^{-1}u_p\|^2 \leq \frac{1}{n\tau}\sum_{i=1}^n X_i^2(p)\psi'(r_{i,[p]}) \ . \tag{25}$$

As we saw in Equation (16) and Lemma 2.1, we have

$$\|\widehat{\beta} - \widetilde{b}\| \leq \frac{1}{\tau}\|f(\widetilde{b})\| \ ,$$

where

$$f(\widetilde{b}) = -\frac{1}{n}\sum_{i=1}^n X_i\psi(\epsilon_i - X_i'\widetilde{b}) + \tau\widetilde{b} \ .$$

We note furthermore that

$$g(\widehat{\gamma}) \triangleq -\frac{1}{n}\sum_{i=1}^n V_i\psi(\epsilon_i - V_i'\widehat{\gamma}) + \tau\widehat{\gamma} = 0_{p-1} \ .$$

The strategy of the proof is to control $f(\widetilde{b})$ by using $g(\widehat{\gamma})$ to create good approximations and then use the fact that $g(\widehat{\gamma}) = 0_{p-1}$.

Proof. **a) Work on the first $(p-1)$ coordinates of $f(\widetilde{b})$**
We call $\mathsf{f}_{p-1}(\beta)$ the first $p-1$ coordinates of $f(\beta)$. We call $\widehat{\gamma}_{ext}$ the p-dimensional vector whose first $p-1$ coordinates are $\widehat{\gamma}$ and last coordinate is 0, i.e

$$\widehat{\gamma}_{ext} = \begin{bmatrix} \widehat{\gamma} \\ 0 \end{bmatrix} .$$

For a vector v, we use the notation $v_{comp,k}$ to denote the $p-1$ dimensional vector consisting of all the coordinates of v except the k-th.

Clearly,

$$\mathsf{f}_{p-1}(\widetilde{b}) = \mathsf{f}_{p-1}(\widetilde{b}) - g(\widehat{\gamma}) = -\frac{1}{n}\sum_{i=1}^{n} V_i \left[\psi(\epsilon_i - X_i'\widetilde{b}) - \psi(\epsilon_i - V_i'\widehat{\gamma})\right] + \tau(\widetilde{b}_{comp,p} - \widehat{\gamma}) .$$

We can write by using the mean value theorem, for $\gamma^*_{i,p}$ in the interval $(\epsilon_i - V_i'\widehat{\gamma}, \epsilon_i - X_i'\widetilde{b})$,

$$\begin{aligned} \psi(\epsilon_i - X_i'\widetilde{b}) - \psi(\epsilon_i - V_i'\widehat{\gamma}) &= \psi'(\gamma^*_{i,p})X_i'(\widehat{\gamma}_{ext} - \widetilde{b}) , \\ &= \psi'(r_{i,[p]})X_i'(\widehat{\gamma}_{ext} - \widetilde{b}) + [\psi'(\gamma^*_{i,p}) - \psi'(r_{i,[p]})]X_i'(\widehat{\gamma}_{ext} - \widetilde{b}) . \end{aligned}$$

Let us call

$$\begin{aligned} \mathsf{d}_{i,p} &= [\psi'(\gamma^*_{i,p}) - \psi'(r_{i,[p]})] , \\ \delta_{i,p} &= [\psi'(\gamma^*_{i,p}) - \psi'(r_{i,[p]})]X_i'(\widehat{\gamma}_{ext} - \widetilde{b}) , \\ \mathsf{R}_p &= -\frac{1}{n}\sum_{i=1}^{n} \mathsf{d}_{i,p} V_i X_i'(\widehat{\gamma}_{ext} - \widetilde{b}) . \end{aligned}$$

We have with this notation

$$\mathsf{f}_{p-1}(\widetilde{b}) = -\frac{1}{n}\sum_{i=1}^{n} \psi'(r_{i,[p]})V_i X_i'(\widehat{\gamma}_{ext} - \widetilde{b}) + \tau(\widetilde{b}_{comp,p} - \widehat{\gamma}) + \mathsf{R}_p \triangleq \mathsf{A}_p + \mathsf{R}_p .$$

We note that by definition,

$$\begin{aligned} \widehat{\gamma}_{ext} - \widetilde{b} &= \mathfrak{b}_p \begin{bmatrix} (\mathfrak{S}_p + \tau \mathrm{Id})^{-1} u_p \\ -1 \end{bmatrix} , \\ \widetilde{b}_{comp,p} - \widehat{\gamma} &= -\mathfrak{b}_p (\mathfrak{S}_p + \tau \mathrm{Id})^{-1} u_p . \end{aligned}$$

Therefore, $X_i'(\widehat{\gamma}_{ext} - \widetilde{b}) = \mathfrak{b}_p \left[V_i'(\mathfrak{S}_p + \tau \mathrm{Id})^{-1} u_p - X_i(p)\right]$, and

$$\mathsf{A}_p = -\mathfrak{b}_p \left(\frac{1}{n}\sum_{i=1}^{n} \psi'(r_{i,[p]}) V_i \left[V_i'(\mathfrak{S}_p + \tau \mathrm{Id})^{-1} u_p - X_i(p)\right]\right) + \tau(-\mathfrak{b}_p(\mathfrak{S}_p + \tau \mathrm{Id})^{-1} u_p) .$$

Recalling the definition of $\mathfrak{S}_p$ and u_p, we see that

$$\mathsf{A}_p = -\mathfrak{b}_p \left(\mathfrak{S}_p(\mathfrak{S}_p + \tau \mathrm{Id})^{-1} u_p - u_p + \tau(\mathfrak{S}_p + \tau \mathrm{Id})^{-1} u_p\right) = 0_{p-1} ,$$

since $\mathfrak{S}_p(\mathfrak{S}_p + \tau \mathrm{Id})^{-1} + \tau(\mathfrak{S}_p + \tau \mathrm{Id})^{-1} = \mathrm{Id}$.

We conclude that

$$\boxed{\mathsf{f}_{p-1}(\widetilde{b}) = \mathsf{R}_p .}$$

b) Work on the last coordinate of $f(\widetilde{b})$
We call $[f(\widetilde{b})]_p$ the last coordinate of $f(\widetilde{b})$. We recall the representation

$$\psi(\epsilon_i - X_i'\widetilde{b}) - \psi(\epsilon_i - V_i'\widehat{\gamma}) = \psi'(r_{i,[p]})X_i'(\widehat{\gamma}_{ext} - \widetilde{b}) + [\psi'(\gamma^*_{i,p}) - \psi'(r_{i,[p]})]X_i'(\widehat{\gamma}_{ext} - \widetilde{b})$$

and call

$$\delta_{i,p} = [\psi'(\gamma^*_{i,p}) - \psi'(r_{i,[p]})]X_i'(\widehat{\gamma}_{ext} - \widetilde{b}) \; .$$

Clearly,

$$\begin{aligned}\psi(\epsilon_i - X_i'\widetilde{b}) &= \psi(r_{i,[p]}) + \psi'(r_{i,[p]})X_i'(\widehat{\gamma}_{ext} - \widetilde{b}) + \delta_{i,p} \; ,\\ &= \psi(r_{i,[p]}) + \psi'(r_{i,[p]})\mathfrak{b}_p \left[V_i'(\mathfrak{S}_p + \tau \mathrm{Id})^{-1}u_p - X_i(p)\right] + \delta_{i,p} \; .\end{aligned}$$

We therefore see that

$$\begin{aligned}[f(\widetilde{b})]_p + \frac{1}{n}\sum_{i=1}^n X_i(p)\delta_{i,p} &= -\frac{1}{n}\sum_{i=1}^n X_i(p)\left(\psi(r_{i,[p]}) + \psi'(r_{i,[p]})\mathfrak{b}_p\left[V_i'(\mathfrak{S}_p + \tau\mathrm{Id})^{-1}u_p - X_i(p)\right]\right) + \tau\widetilde{b}_p \; ,\\ &= -\frac{1}{n}\sum_{i=1}^n X_i(p)\psi(r_{i,[p]}) - \mathfrak{b}_p u_p'(\mathfrak{S}_p + \tau\mathrm{Id})^{-1}u_p + \mathfrak{b}_p\frac{1}{n}\sum_{i=1}^n \psi'(r_{i,[p]})X_i^2(p) + \tau\mathfrak{b}_p \; ,\\ &= -\left[\frac{1}{n}\sum_{i=1}^n X_i(p)\psi(r_{i,[p]}) - \tau\mathfrak{b}_p\right] + \mathfrak{b}_p\left(\frac{1}{n}\sum_{i=1}^n \psi'(r_{i,[p]})X_i^2(p) - u_p'(\mathfrak{S}_p + \tau\mathrm{Id})^{-1}u_p\right) \; ,\\ &= -\left[\frac{1}{\sqrt{n}}N_p - \tau\mathfrak{b}_p\right] + \mathfrak{b}_p\xi_n \; ,\\ &= 0 \; .\end{aligned}$$

We conclude that

$$[f(\widetilde{b})]_p = -\frac{1}{n}\sum_{i=1}^n X_i(p)\delta_{i,p} = -\frac{1}{n}\sum_{i=1}^n \mathsf{d}_{i,p}X_i(p)X_i'(\widehat{\gamma}_{ext} - \widetilde{b}) \; .$$

Representation of $f(\widetilde{b})$

Aggregating all the results we have obtained so far, we see that

$$\begin{aligned}f(\widetilde{b}) &= \left(-\frac{1}{n}\sum_{i=1}^n \mathsf{d}_{i,p}X_iX_i'\right)(\widehat{\gamma}_{ext} - \widetilde{b}) \; ,\\ &= -\mathfrak{b}_p\left(\frac{1}{n}\sum_{i=1}^n \mathsf{d}_{i,p}X_iX_i'\right)\begin{bmatrix}(\mathfrak{S}_p + \tau\mathrm{Id})^{-1}u_p\\ -1\end{bmatrix} \; .\end{aligned}$$

We conclude immediately that

$$\|f(\widetilde{b})\| \leq |\mathfrak{b}_p| \sup_{1\leq i\leq n}|\mathsf{d}_{i,p}| \; |||\widehat{\Sigma}|||_2\sqrt{\|(\mathfrak{S}_p + \tau\mathrm{Id})^{-1}u_p\|^2 + 1} \; . \tag{26}$$

In connection with Equation (15), this gives Equation (24).
Calling $D_{\psi'(r_{\cdot,[p]})}$ the diagonal matrix with (i,i) entry $\psi'(r_{i,[p]})$, we see that

$$u_p = \frac{1}{n}V'D_{\psi'(r_{\cdot,[p]})}X(p) \text{ and } \mathfrak{S}_p = \frac{1}{n}V'D_{\psi'(r_{\cdot,[p]})}V \; . \tag{27}$$

Therefore,

$$\|(\mathfrak{S} + \tau\mathrm{Id})^{-1}u_p\|^2 = \frac{X(p)'}{\sqrt{n}}D^{1/2}_{\psi'(r_{\cdot,[p]})}\frac{D^{1/2}_{\psi'(r_{\cdot,[p]})}V}{\sqrt{n}}\left(\frac{V'D_{\psi'(r_{\cdot,[p]})}V}{n} + \tau\mathrm{Id}\right)^{-2}\frac{V'D^{1/2}_{\psi'(r_{\cdot,[p]})}}{\sqrt{n}}D^{1/2}_{\psi'(r_{\cdot,[p]})}\frac{X(p)}{\sqrt{n}} \; .$$

Clearly, using for instance the singular value decomposition of $\frac{D^{1/2}_{\psi'(r_{\cdot,[p]})}V}{\sqrt{n}}$ or Lemma V.1.5 in Bhatia (1997),

$$\frac{D^{1/2}_{\psi'(r_{\cdot,[p]})}V}{\sqrt{n}}\left(\frac{V'D_{\psi'(r_{\cdot,[p]})}V}{n}+\tau\mathrm{Id}\right)^{-1}\frac{V'D^{1/2}_{\psi'(r_{\cdot,[p]})}}{\sqrt{n}}\preceq \mathrm{Id}\;,$$

and

$$\frac{D^{1/2}_{\psi'(r_{\cdot,[p]})}V}{\sqrt{n}}\left(\frac{V'D_{\psi'(r_{\cdot,[p]})}V}{n}+\tau\mathrm{Id}\right)^{-2}\frac{V'D^{1/2}_{\psi'(r_{\cdot,[p]})}}{\sqrt{n}}\preceq \frac{\mathrm{Id}}{\tau}\;.$$

So we have

$$\|(\mathfrak{S}+\tau\mathrm{Id})^{-1}u_p\|^2\leq\frac{1}{n\tau}X(p)'D_{\psi'(r_{\cdot,[p]})}X(p)=\frac{1}{n\tau}\sum_{i=1}^n X_i^2(p)\psi'(r_{i,[p]})\;.$$

□

4.2 Probabilistic aspects

From now on, we assume that $X(p)$, the p-th column of the design matrix, is independent of $\{V_i,\epsilon_i\}_{i=1}^n$. This is consistent with Assumption **P1**.

Because $r_{i,[p]}$ are the residuals from a ridge-regularized robust regression problem with n observations and $p-1$ predictors, the analysis done above concerning the R_i - see Lemma 3.4, p. 18 - applies and will allow us to control $\max_{1\leq i\leq n}|\psi'(r_{i,[p]})|^2$. (Note that Assumption **O4** is satisfied for V_i if it is satisfied for X_i: convex 1-Lipschitz function of V_i can be trivially made to be convex 1-Lipschitz function of X_i by simply not acting on the last coordinate of X_i.)

In light of Lemma 3.4 and using independence of $X_i(p)$'s and $r_{i,[p]}$, it is clear that the upper bound in Equation (25) is $O_{L_k}(\mathrm{polyLog}(n))$ under Assumptions **O1-O7** and **P1**. In other words, at τ fixed,

$$\|(\mathfrak{S}_p+\tau\mathrm{Id})^{-1}u_p\|^2=\mathrm{O}_{L_k}(\mathrm{polyLog}(n))$$

(Note that p does not play a particular role here. If we considered the same quantity when we remove the k-th predictor instead of the p-th, and took the sup over $1\leq k\leq p$ of the corresponding random variables, the same inequality would hold, in light of our work in Section 3.)

This guarantees that

$$\left\|\begin{matrix}(\mathfrak{S}_p+\tau\mathrm{Id})^{-1}u_p\\ -1\end{matrix}\right\|^2\leq(1+\|(\mathfrak{S}_p+\tau\mathrm{Id})^{-1}u_p\|^2)=\mathrm{O}_{L_k}(\mathrm{polyLog}(n))\;.$$

We conclude, using Equation (26), that

$$\|\widehat{\beta}-\widetilde{b}\|\leq\frac{K}{\tau}\mathrm{polyLog}(n)|\mathfrak{b}_p|\sup_{1\leq i\leq n}|\mathsf{d}_{i,p}|\;|||\widehat{\Sigma}|||_2\text{ in }L_k.$$

Recall that Lemma B-4 gives a bound on $|||\widehat{\Sigma}|||_2$. At a high level, we expect $\sup_{1\leq i\leq n}|\mathsf{d}_{i,p}|$ and $\mathfrak{b}_p$ to be small, which should give us that

$$\|\widehat{\beta}-\widetilde{b}\|=\mathrm{O}_{L_k}(\mathrm{polyLog}(n)\sup_{1\leq i\leq n}|\mathsf{d}_{i,p}||\mathfrak{b}_p|)\;.$$

In fact, we will show in Proposition 4.2 that $\mathfrak{b}_p=\mathrm{O}_{L_k}(\mathrm{polyLog}(n)n^{-1/2})$ and in Proposition 4.4 that $\sup_{1\leq i\leq n}|\mathsf{d}_{i,p}|=\mathrm{O}_{L_k}(\mathrm{polyLog}(n)n^{-1/2})$.

We now show these two results.

4.2.1 On $\mathfrak{b}_p$

We recall the notations

$$N_p = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(r_{i,[p]}) X_i(p) \ ,$$

$$\xi_n = \frac{1}{n} \sum_{i=1}^{n} \psi'(r_{i,[p]}) X_i^2(p) - u_p'(\mathfrak{S}_p + \tau \mathrm{Id})^{-1} u_p \ .$$

Under our assumptions, we have $\mathbf{E}\left(X_i\right) = 0$ and $\mathrm{cov}\left(X_i\right) = \mathrm{Id}_p$ and hence $\mathbf{E}\left(X_i^2(p)\right) = 1$. Recall that since we assume that $X(p)$ is independent of $\{V_i, \epsilon_i\}_{i=1}^n$, $X(p)$ is independent of $\{r_{i,[p]}\}_{i=1}^n$.

Proposition 4.2. *We have*

$$|\mathfrak{b}_p| \leq \frac{1}{\sqrt{n}\tau} |N_p| \ .$$

*Furthermore, under assumptions **O1-O7** and **P1**, $N_p = \mathrm{O}_{L_k}(polyLog(n))$ and therefore, when τ is held fixed,*

$$\mathfrak{b}_p = \mathrm{O}_{L_k}(polyLog(n) n^{-1/2}) \ .$$

Proof. From the definition of $\mathfrak{b}_p$, we see that, when $\xi_n \neq 0$

$$\mathfrak{b}_p = \frac{1}{\sqrt{n}} \frac{N_p}{\tau + \xi_n} \ .$$

We will see later, in Subsubsection 4.2.2, that $\xi_n \geq 0$. It immediately then follows that

$$|\mathfrak{b}_p| \leq \frac{1}{\sqrt{n}\tau} |N_p| \ .$$

Using independence of $X(p)$ and $\{V_i, \epsilon_i\}_{i=1}^n$, we have for instance

$$\mathbf{E}\left(N_p^2\right) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{E}\left(X_i^2(p)\right) \mathbf{E}\left(\psi^2(r_{i,[p]})\right) \ ,$$

whether the right-hand side is finite or not.

Since $r_{i,[p]}$ are the residuals for our original problem with n observations and $p-1$ predictors, our previous analyses show that N_p has as many moments as we need and $N_p = \mathrm{O}_{L_k}(\mathrm{polyLog}(n))$. (Indeed, for higher moments, since $X_i(p)$ are independent random variables when i varies from 1 to n, it suffices to apply reasoning similar to the arguments given in Lemma 2.2 for the control of the moments of $W_n(j)$ in connections with our bounds on $r_{i,[p]}$ and therefore on $\psi(r_{i,[p]})$ under Assumption **O3**).

We therefore have

$$|\mathfrak{b}_p| \leq \frac{1}{\sqrt{n}\tau} \mathrm{O}_{L_k}(\mathrm{polyLog}(n)) \ .$$

□

4.2.2 On ξ_n

Let us write ξ_n using matrix notations: denoting by $X(p)$ the last column of the design matrix X, we have

$$\xi_n = \frac{1}{n} X(p)' D^{1/2}_{\psi'(r_{\cdot,[p]})} M D^{1/2}_{\psi'(r_{\cdot,[p]})} X(p) \ , \tag{28}$$

where

$$M = \mathrm{Id}_n - \frac{D^{1/2}_{\psi'(r_{\cdot,[p]})} V}{\sqrt{n}} \left(\frac{1}{n} V' D_{\psi'(r_{\cdot,[p]})} V + \tau \mathrm{Id}\right)^{-1} \frac{V' D^{1/2}_{\psi'(r_{\cdot,[p]})}}{\sqrt{n}} \ . \tag{29}$$

This simply comes from the representations of u_p and $\mathfrak{S}_p$ given in the proof of Proposition 4.1, specifically in Equation (27).

Lemma 4.1. *We have*

$$\xi_n \geq 0 \; .$$

*Furthermore, under Assumptions **O1-O7** and **P1**,*

$$|\xi_n - \frac{1}{n} trace\left(D^{1/2}_{\psi'(r_{\cdot,[p]})} M D^{1/2}_{\psi'(r_{\cdot,[p]})}\right)| = \mathrm{O}_{L_k}\left(\sup_{1\leq i\leq n} \psi'(r_{i,[p]})/(\sqrt{n\mathsf{c}_n})\right) \; . \tag{30}$$

Proof. Let us first focus on M, as defined in Equation (29). When $\tau > 0$, it is clear that all the eigenvalues of M are strictly positive, i.e M is positive definite. Indeed, if the singular values of $n^{-1/2} D^{1/2}_{\psi'(r_{\cdot,[p]})} V$ are denoted by σ_i, the eigenvalues of M are $\tau/(\sigma_i^2 + \tau)$.

Therefore, since $\xi_n = \frac{1}{n} v' M v$ with $v = D^{1/2}_{\psi'(r_{\cdot,[p]})} X(p)$, $\xi_n \geq 0$.

M is symmetric and has eigenvalues between 0 and 1, as we just saw. Therefore, using e.g Lemma V.1.5 in Bhatia (1997),

$$0 \preceq D^{1/2}_{\psi'(r_{\cdot,[p]})} M D^{1/2}_{\psi'(r_{\cdot,[p]})} \preceq D_{\psi'(r_{\cdot,[p]})} \; .$$

The matrix M is independent of $X(p)$ under Assumption **P1**. $D_{\psi'(r_{\cdot,[p]})}$ is also independent of $X(p)$.

Since X_p satisfy the necessary concentration assumptions under Assumption **P1**, we can now appeal to Lemma B-3 to obtain

$$\left|\frac{1}{n} X(p)' D^{1/2}_{\psi'(r_{\cdot,[p]})} M D^{1/2}_{\psi'(r_{\cdot,[p]})} X(p) - \frac{1}{n}\mathrm{trace}\left(D^{1/2}_{\psi'(r_{\cdot,[p]})} M D^{1/2}_{\psi'(r_{\cdot,[p]})}\right)\right| = \mathrm{O}_{L_k}\left(\frac{1}{\sqrt{n\mathsf{c}_n}} \sup_i \psi'(r_{i,[p]})\right) \; .$$

□

We now take a slight detour from the aim of showing that we have a very good approximation of $\widehat{\beta}$ through $\widetilde{b}$ by working on finer properties of ξ_n and $\mathfrak{b}_p$. These properties will be essential in establishing the validity of the system (3).

To get a finer understanding of ξ_n, we now focus on the properties of

$$\frac{1}{n}\mathrm{trace}\left(D^{1/2}_{\psi'(r_{\cdot,[p]})} M D^{1/2}_{\psi'(r_{\cdot,[p]})}\right) \; .$$

About $\frac{1}{n}$trace $\left(D^{1/2}_{\psi'(r_{\cdot,[p]})} M D^{1/2}_{\psi'(r_{\cdot,[p]})}\right)$

Lemma 4.2. *Let us call* $\mathfrak{S}_p = \frac{1}{n}\sum_{i=1}^n \psi'(r_{i,[p]}) V_i V_i'$ *and* $\mathfrak{S}_p(i) = \mathfrak{S}_p - \frac{1}{n}\psi'(r_{i,[p]}) V_i V_i'$. *Let us also call*

$$\mathsf{c}_{\tau,p} = \frac{1}{n} trace\left((\mathfrak{S}_p + \tau \mathrm{Id})^{-1}\right) \; ,$$
$$\zeta_i = \frac{1}{n} V_i' (\mathfrak{S}_p(i) + \tau\mathrm{Id})^{-1} V_i - \mathsf{c}_{\tau,p} \; .$$

*Then we have under Assumptions **O1-O7** and **P1**, if M is the matrix defined in Equation* (29),

$$\left|\frac{1}{n} trace\left(\mathrm{Id}_n - M\right) - \left(\frac{1}{n} trace\left(D^{1/2}_{\psi'(r_{\cdot,[p]})} M D^{1/2}_{\psi'(r_{\cdot,[p]})}\right)\right)\mathsf{c}_{\tau,p}\right| \leq \left[\sup_i |\zeta_i|\right] \frac{1}{n}\sum_i \psi'(r_{i,[p]}) \; . \tag{31}$$

We also have

$$\frac{1}{n} trace\left(\mathrm{Id}_n - M\right) = \frac{p-1}{n} - \tau \mathsf{c}_{\tau,p} \; .$$

Proof. We call $d_{i,i} = \psi'(r_{i,[p]})/n$. Of course, by using the Sherman-Morrison-Woodbury formula (see e.g Horn and Johnson (1990), p.19),

$$\begin{aligned} M_{i,i} &= 1 - d_{i,i} V_i' (V' D_{\psi'(r_{\cdot,[p]})} V/n + \tau \mathrm{Id})^{-1} V_i \; , \\ &= 1 - d_{i,i} \frac{V_i'(\mathfrak{S}_p(i) + \tau\mathrm{Id})^{-1} V_i}{1 + d_{i,i} V_i' (\mathfrak{S}_p(i) + \tau\mathrm{Id})^{-1} V_i} \; , \\ &= \frac{1}{1 + d_{i,i} V_i'(\mathfrak{S}_p(i) + \tau\mathrm{Id})^{-1} V_i} \; . \end{aligned}$$

Recall that we are interested in $\frac{1}{n}\sum_i \psi'(r_{i,[p]})M_{i,i} = \frac{1}{n}\text{trace}\left(D^{1/2}_{\psi'(r_{\cdot,[p]})} M D^{1/2}_{\psi'(r_{\cdot,[p]})}\right)$. Note that, since $\text{trace}\,(AB) = \text{trace}\,(BA)$,

$$\text{trace}\,(\text{Id}_n - M) = \text{trace}\left((\mathfrak{S}_p + \tau\text{Id})^{-1}\mathfrak{S}_p\right) = p - 1 - \tau\text{trace}\left((\mathfrak{S}_p+\tau\text{Id})^{-1}\right) = p - 1 - n\tau\mathsf{c}_{\tau,p}\ .$$

On the other hand,

$$\text{trace}\,(\text{Id}_n - M) = \sum_i (1 - M_{i,i}) = \sum_i d_{i,i}\frac{V_i'(\mathfrak{S}_p(i)+\tau\text{Id})^{-1}V_i}{1 + d_{i,i}V_i'(\mathfrak{S}_p(i)+\tau\text{Id})^{-1}V_i}\ . \tag{32}$$

With our definitions, we have

$$\frac{1}{n}\text{trace}\,(\text{Id}_n - M) = \left(\frac{1}{n}\sum_i \psi'(r_{i,[p]})M_{i,i}\right)\mathsf{c}_{\tau,p} + \frac{1}{n}\sum_i \psi'(r_{i,[p]})\frac{\zeta_i}{1 + d_{i,i}V_i'(\mathfrak{S}_p(i)+\tau\text{Id})^{-1}V_i}\ .$$

It immediately follows that

$$\left|\frac{1}{n}\text{trace}\,(\text{Id}_n - M) - \left(\frac{1}{n}\sum_i \psi'(r_{i,[p]})M_{i,i}\right)\mathsf{c}_{\tau,p}\right| \leq \left[\sup_i |\zeta_i|\right]\frac{1}{n}\sum_i \psi'(r_{i,[p]})\ ,$$

as announced. □

Controlling ζ_i

Lemma 4.3. *Suppose we can find $\{\mathsf{r}^{(i)}_{j,[p]}\}_{j\neq i}$ independent of V_i such that $\sup_{j\neq i}|\mathsf{r}^{(i)}_{j,[p]} - r_{j,[p]}| \leq \delta_n(i)$.*
Suppose further that we can find K_n such that

$$\sup_i \sup_{j\neq i} |\psi'(\mathsf{r}^{(i)}_{j,[p]}) - \psi'(r_{j,[p]})| \leq K_n$$

Then

$$\sup_i |\zeta_i| = \text{O}_{L_k}\left(\frac{1}{\tau^2}K_n|||\widehat{\Sigma}|||_2 + \frac{\mathit{polyLog}(n)}{\tau\sqrt{n\mathsf{c}_n}} + \frac{1}{n\tau}\right)\ , \tag{33}$$

provided K_n has $3k$ uniformly bounded moments.

Proof. We call

$$AM_{i,p} = \frac{1}{n}\sum_{j\neq i}\psi'(\mathsf{r}^{(i)}_{j,[p]})V_jV_j'\ .$$

Then, using for instance the first resolvent identity, i.e $A^{-1} - B^{-1} = A^{-1}(B-A)B^{-1}$, we see that

$$|||(\mathfrak{S}_p(i)+\tau\text{Id})^{-1} - (AM_{i,p}+\tau\text{Id})^{-1}|||_2 \leq \frac{1}{\tau^2}K_n|||\widehat{\Sigma}|||_2\ ,$$

since $|||\frac{1}{n}\sum_i V_iV_i'|||_2 \leq |||\widehat{\Sigma}|||_2$. In particular,

$$\left|\frac{1}{n}V_i'(\mathfrak{S}_p(i)+\tau\text{Id})^{-1}V_i - \frac{1}{n}V_i'(AM_{i,p}+\tau\text{Id})^{-1}V_i\right| \leq \frac{\|V_i\|^2}{n}\frac{1}{\tau^2}K_n|||\widehat{\Sigma}|||_2\ .$$

However, since $AM_{i,p}$ is independent of V_i, we can use Lemma B-3 and see that

$$\sup_{1\leq i\leq n}\left|\frac{1}{n}V_i'(AM_{i,p}+\tau\text{Id})^{-1}V_i - \frac{1}{n}\text{trace}\left((AM_{i,p}+\tau\text{Id})^{-1}\right)\right| = \text{O}_{L_k}(\frac{\text{polyLog}(n)}{\tau\sqrt{n\mathsf{c}_n}})\ ,$$

by using the fact that $\lambda_{\max}((AM_{i,p}+\tau\text{Id})^{-1}) \leq \frac{1}{\tau}$.
However, by the argument we gave above,

$$\left|\frac{1}{n}\text{trace}\left((AM_{i,p}+\tau\text{Id})^{-1}\right) - \frac{1}{n}\text{trace}\left((\mathfrak{S}_p(i)+\tau\text{Id})^{-1}\right)\right| \leq \frac{1}{\tau^2}K_n|||\widehat{\Sigma}|||_2\frac{p}{n}\ .$$

We conclude that

$$\sup_{1\leq i\leq n}\left|\frac{1}{n}V_i'(\mathfrak{S}_p(i)+\tau\text{Id})^{-1}V_i-\frac{1}{n}\text{trace}\left((\mathfrak{S}_p(i)+\tau\text{Id})^{-1}\right)\right|\leq\frac{1}{\tau^2}K_n|||\widehat{\Sigma}|||_2\sup_{1\leq i\leq n}\left[\frac{p}{n}+\frac{\|V_i\|^2}{n}\right]+\frac{\text{polyLog}(n)}{\tau\sqrt{n\mathsf{c}_n}}\ ,$$

in L_k.

Now, it is clear that under **O1** and **O4**, $\sup_{1\leq i\leq n}\|V_i\|^2/n=\text{O}_{L_k}(1)$ and finally

$$\sup_{1\leq i\leq n}\left|\frac{1}{n}V_i'(\mathfrak{S}_p(i)+\tau\text{Id})^{-1}V_i-\frac{1}{n}\text{trace}\left((\mathfrak{S}_p(i)+\tau\text{Id})^{-1}\right)\right|=\text{O}_{L_k}(\frac{1}{\tau^2}K_n|||\widehat{\Sigma}|||_2+\frac{\text{polyLog}(n)}{\tau\sqrt{n\mathsf{c}_n}})\ .$$

Control of $\frac{1}{n}$trace $\left((\mathfrak{S}_p(i)+\tau\text{Id})^{-1}\right)-\frac{1}{n}$trace $\left((\mathfrak{S}_p+\tau\text{Id})^{-1}\right)$
Using the Sherman-Woodbury-Morrison formula, we have

$$(\mathfrak{S}_p(i)+\tau\text{Id})^{-1}-(\mathfrak{S}_p+\tau\text{Id})^{-1}=\frac{\psi'(r_{i,[p]})}{n}\frac{(\mathfrak{S}_p(i)+\tau\text{Id})^{-1}V_iV_i'(\mathfrak{S}_p(i)+\tau\text{Id})^{-1}}{1+\frac{\psi'(r_{i,[p]})}{n}V_i'(\mathfrak{S}_p(i)+\tau\text{Id})^{-1}V_i}\ .$$

After taking traces, we see that

$$0\leq\text{trace}\left((\mathfrak{S}_p(i)+\tau\text{Id})^{-1}\right)-\text{trace}\left((\mathfrak{S}_p+\tau\text{Id})^{-1}\right)\leq\frac{1}{\tau}\ ,$$

since $V_i'(\mathfrak{S}_p(i)+\tau\text{Id})^{-2}V_i\leq\frac{1}{\tau}V_i'(\mathfrak{S}_p(i)+\tau\text{Id})^{-1}V_i$.

Therefore,

$$0\leq\frac{1}{n}\text{trace}\left((\mathfrak{S}_p(i)+\tau\text{Id})^{-1}\right)-\frac{1}{n}\text{trace}\left((\mathfrak{S}_p+\tau\text{Id})^{-1}\right)\leq\frac{1}{n\tau}\ .$$

We conclude that

$$\sup_{1\leq i\leq n}|\zeta_i|=\text{O}_{L_k}\left(\frac{1}{\tau^2}K_n|||\widehat{\Sigma}|||_2+\frac{\text{polyLog}(n)}{\tau\sqrt{n\mathsf{c}_n}}+\frac{1}{n\tau}\right)\ ,$$

provided we can use Holder's inequality. In effect, this requires K_n to have $3k$ uniformly bounded moments. □

4.2.3 Control of K_n

A natural choice for $\mathsf{r}_{j,[p]}^{(i)}$ defined in Lemma 4.3 is to use a leave one out estimator of $\widehat{\gamma}$. Hence, all the work done in Theorem 3.1 becomes immediately relevant.

Lemma 4.4. *Suppose we use for $\{\mathsf{r}_{j,[p]}^{(i)}\}_{j\neq i}$ the residuals we would get by using a leave-one-out estimator of $\widehat{\gamma}$, i.e excluding (V_i,ϵ_i) from problem* (7).

With the notations of Lemma 4.3, we have

$$\sup_i(\delta_n(i))=\text{O}_{L_k}\left(\frac{\textit{polyLog}(n)}{n^{1/2}}\right)\ .$$

Therefore,

$$K_n=\text{O}_{L_k}\left(n^{-1/2}\textit{polyLog}(n)\right)$$

Proof. The first statement of the Lemma is an application of Theorem 3.1 with $R_j=r_{j,[p]}$ and $\tilde{r}_{j,(i)}=\mathsf{r}_{j,[p]}^{(i)}$.

The control of K_n follows immediately by using our assumptions on ψ' and on the growth of $\mathcal{B}_n(i)$ and $L(\mathcal{B}_n(i))$ we had before, now applied to the situation with $p-1$ predictors. □

Important remark: the previous remark has important consequences for c_i defined in Equation (5): we just showed that $\sup_i|\frac{1}{n}V_i'(\mathfrak{S}_p(i)+\tau\text{Id})^{-1}V_i-\mathsf{c}_{\tau,p}|=\text{O}_{L_k}(\text{polyLog}(n)/\sqrt{n})$. Recalling the notation

$$c_\tau=\frac{1}{n}\text{trace}\left(\left[\frac{1}{n}\sum_{i=1}^n\psi'(R_i)X_iX_i'+\tau\text{Id}_p\right]^{-1}\right)\ ,$$

which is the analog of $\mathsf{c}_{\tau,p}$ when we use all the predictors and not only $(p-1)$, we see that $\sup_i |c_i - c_\tau| = \mathrm{O}(n^{-1/2}\text{polyLog}(n))$. Indeed, c_i in Equation (5) is defined, in the notation of the proof of Lemma 4.3 as $\frac{1}{n}V_i'(AM_{i,p}+\tau\text{Id})^{-1}V_i$, with the role of $\{\mathsf{r}_{j,[p]}^{(i)}\}_{j\neq i}$ being played by the residuals obtained from the leave-one-out estimate of $\widehat{\beta}$, excluding (X_i,ϵ_i) from the problem. Lemma 4.3 in connection with Theorem 4.1 shows that $\sup_i |\frac{1}{n}V_i'(AM_{i,p}+\tau\text{Id})^{-1}V_i - \mathsf{c}_{\tau,p}| = \mathrm{O}_{L_k}(\text{polyLog}(n)/\sqrt{n})$ under our assumptions. Passing from the $p-1$ dimensional version of this result, i.e Lemma 4.3, to the p-dimensional version gives the approximation

$$\sup_i |c_i - c_\tau| = \mathrm{O}_{L_k}(n^{-1/2}\text{polyLog}(n)) \ . \tag{34}$$

4.2.4 Further results on ξ_n and $\mathfrak{b}_p$

We can combine all the results we have obtained so far in the following proposition.

Proposition 4.3. *We have, under Assumptions **O1-O7** and **P1**,*

$$\left|\mathsf{c}_{\tau,p}(\xi_n+\tau) - \frac{p-1}{n}\right| \leq \mathrm{O}_{L_k}\left((\sup_i \psi'(r_{i,[p]}))\left(\frac{polyLog(n)}{\sqrt{n\mathsf{c}_n}} + \frac{1}{\tau^2}K_n|||\widehat{\Sigma}|||_2 + \frac{1}{n\tau}\right)\right) = \mathrm{O}_{L_k}\left(\frac{polyLog(n)}{\sqrt{n}}\right) \ . \tag{35}$$

Furthermore, under our assumptions,

$$\left(\frac{p}{n}\right)^2 n\mathbf{E}\left(\mathfrak{b}_p^2\right) = \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left((\mathsf{c}_{\tau,p}\psi(r_{i,[p]})^2\right) + \mathrm{o}(1) \ . \tag{36}$$

Proof. The proof of Equation (35) consists just in aggregating all the previous results and noticing that $\mathsf{c}_{\tau,p} \leq (p-1)/(n\tau)$ and therefore remains bounded. Indeed, we have

$$\frac{p-1}{n} - \tau\mathsf{c}_{\tau,p} = \frac{1}{n}\text{trace}\left(\text{Id} - M\right) \geq 0 \ .$$

This latter quantity was approximated in Lemma 4.2 by

$$\left(\frac{1}{n}\text{trace}\left(D^{1/2}_{\psi'(r_{\cdot,[p]})}MD^{1/2}_{\psi'(r_{\cdot,[p]})}\right)\right)\mathsf{c}_{\tau,p} \ .$$

And in Lemma 4.1, we approximated ξ_n by $\left(\frac{1}{n}\text{trace}\left(D^{1/2}_{\psi'(r_{\cdot,[p]})}MD^{1/2}_{\psi'(r_{\cdot,[p]})}\right)\right)$.

We recall that

$$(\tau+\xi_n)\sqrt{n}\mathfrak{b}_p|\{V_i,\epsilon_i\} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(r_{i,[p]})X_i(p) \ .$$

Therefore,

$$\mathsf{c}_{\tau,p}(\tau+\xi_n)\sqrt{n}\mathfrak{b}_p|\{V_i,\epsilon_i\} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathsf{c}_{\tau,p}\psi(r_{i,[p]})X_i(p)$$

Now, $\mathsf{c}_{\tau,p}\psi(r_{i,[p]})$, which depends only on $\{V_i,\epsilon_i\}_{i=1}^n$, is independent of $\{X_i(p)\}_{i=1}^n$.

Since $X_i(p)$'s are independent with mean 0 and variance 1, we conclude that

$$\mathbf{E}\left((\mathsf{c}_{\tau,p}(\tau+\xi_n)\sqrt{n}\mathfrak{b}_p)^2\right) = \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left((\mathsf{c}_{\tau,p}\psi(r_{i,[p]})^2\right) \ .$$

Given the result in Equation (35) and our bound on $\sqrt{n}\mathfrak{b}_p$ in Proposition 4.2, this means that

$$\left(\frac{p}{n}\right)^2 n\mathbf{E}\left(\mathfrak{b}_p^2\right) = \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left((\mathsf{c}_{\tau,p}\psi(r_{i,[p]})^2\right) + \mathrm{o}(1) \ .$$

□

We now need to control $\mathsf{d}_{i,p}$ to show that our approximation of $\widehat{\beta}$ by $\widetilde{b}$ in Proposition 3.1 will yield sufficiently good results that they can be used to prove Theorem 1.1.

4.2.5 On $\mathsf{d}_{i,p}$

Recall the definition

$$\mathsf{d}_{i,p}=[\psi'(\gamma^*_{i,p})-\psi'(r_{i,[p]})] ,$$

where $\gamma^*_{i,p}\in(r_{i,[p]},r_{i,[p]}+\nu_i)$, with

$$\nu_i=\mathfrak{b}_pX_i'\begin{bmatrix}(\mathfrak{S}_p+\tau\mathrm{Id})^{-1}u_p\\-1\end{bmatrix}=\mathfrak{b}_p\pi_i .$$

We call $\widetilde{B}_n(i)=\sup_i|r_{i,[p]}|+\sup_i|\pi_i|$.

We have the following result.

Proposition 4.4. *We have, under Assumptions **O1-O7** and **P1**, at fixed τ,*

$$\sup_i|\mathsf{d}_{i,p}|=\mathrm{O}_{L_k}\left(\frac{polyLog(n)}{\sqrt{n}\mathsf{c}_n^{1/2}}L(\widetilde{B}_n(i))\left[\psi'(-\widetilde{B}_n(i))\vee\psi'(\widetilde{B}_n(i))\right]\right) .$$

Hence,

$$\sup_i|\mathsf{d}_{i,p}|=\mathrm{O}_{L_k}\left(\frac{polyLog(n)}{\sqrt{n}}\right) .$$

Proof. Recall the definition

$$\mathsf{d}_{i,p}=[\psi'(\gamma^*_{i,p})-\psi'(r_{i,[p]})] ,$$

where $\gamma^*_{i,p}\in(r_{i,[p]},r_{i,[p]}+\nu_i)$, with

$$\nu_i=\mathfrak{b}_pX_i'\begin{bmatrix}(\mathfrak{S}_p+\tau\mathrm{Id})^{-1}u_p\\-1\end{bmatrix}=\mathfrak{b}_p\pi_i .$$

Therefore,

$$\pi_i=V_i'(\mathfrak{S}_p+\tau\mathrm{Id})^{-1}u_p-X_i(p) .$$

Recall that $u_p=\frac{1}{n}V'D_{\psi'(r_{\cdot,[p]})}X(p)$. Using independence of $X(p)$ with $\{(V_i,\epsilon_i)\}_{i=1}^n$, and our concentration assumptions on $X(p)$ formulated in **P1**, we see that according to Lemma B-2, we have

$$\sup_i|V_i'(\mathfrak{S}_p+\tau\mathrm{Id})^{-1}u_p|=\mathrm{O}_{L_k}\left(\frac{\mathrm{polyLog}(n)}{\mathsf{c}_n^{1/2}}\sup_i\|\frac{1}{n}D_{\psi'(r_{\cdot,[p]})}V(\mathfrak{S}_p+\tau\mathrm{Id})^{-1}V_i\|\right) ,$$

where we look at $V_i'(\mathfrak{S}_p+\tau\mathrm{Id})^{-1}u_p$ as a linear form in $X(p)$.

Now,

$$\|\frac{1}{n}D_{\psi'(r_{\cdot,[p]})}V(\mathfrak{S}_p+\tau\mathrm{Id})^{-1}V_i\|^2=\frac{1}{n}V_i'(\mathfrak{S}_p+\tau\mathrm{Id})^{-1}\frac{V'D^2_{\psi'(r_{\cdot,[p]})}V}{n}(\mathfrak{S}_p+\tau\mathrm{Id})^{-1}V_i .$$

Since $\mathfrak{S}_p=\frac{V'D_{\psi'(r_{\cdot,[p]})}V}{n}$, we have $\frac{V'D^2_{\psi'(r_{\cdot,[p]})}V}{n}\preceq|||D_{\psi'(r_{\cdot,[p]})}|||_2\mathfrak{S}_p$ and we conclude that

$$\frac{1}{n}V_i'(\mathfrak{S}_p+\tau\mathrm{Id})^{-1}\frac{V'D^2_{\psi'(r_{\cdot,[p]})}V}{n}(\mathfrak{S}_p+\tau\mathrm{Id})^{-1}V_i\leq\frac{\|V_i\|^2}{n\tau}|||D_{\psi'(r_{\cdot,[p]})}|||_2=\frac{\|V_i\|^2}{n\tau}\sup_i\psi'(r_{i,[p]}) .$$

We also note that $\sup_i X_i(p)=\mathrm{O}_{L_k}(\mathrm{polyLog}(n)/\sqrt{\mathsf{c}_n})$ and conclude that

$$\begin{aligned}\sup_i|\pi_i|&=\mathrm{O}_{L_k}\left(\frac{\mathrm{polyLog}(n)}{\mathsf{c}_n^{1/2}}\left[1+\sqrt{\sup_i\psi'(r_{i,[p]})\sup_i\frac{\|V_i\|^2}{n\tau}}\right]\right) ,\\&=\mathrm{O}_{L_k}\left(\frac{\mathrm{polyLog}(n)}{\mathsf{c}_n^{1/2}}\left[1+\sqrt{\sup_i\psi'(r_{i,[p]})}\right]\right) .\end{aligned}$$

Recalling that $\mathfrak{b}_p = \mathrm{O}_{L_k}(n^{-1/2}\text{polyLog}(n))$, we finally see that

$$\sup_i \nu_i = \mathrm{O}_{L_k}\left(\frac{\text{polyLog}(n)}{\sqrt{n}\mathsf{c}_n^{1/2}}\left[1+\sqrt{\sup_i \psi'(r_{i,[p]})}\right]\right)$$

As before, we can control $\sup_i \psi'(r_{i,[p]})$ by using the work done in Proposition 3.3, since $r_{i,[p]}$ are the residuals when we work with $p-1$ predictors and n observations. The growth conditions we have imposed on ψ' and $\mathcal{E}_n$ therefore guarantee control of $\left[\sup_i \psi'(r_{i,[p]})\right]$ as polyLog(n) in L_k. Recall that $\widetilde{B}_n(i) = \sup_i |r_{i,[p]}| + \sup_i |\pi_i|$.

So we have shown that under our assumptions,

$$\sup_i |\mathsf{d}_{i,p}| = \mathrm{O}_{L_k}\left(\frac{\text{polyLog}(n)}{\sqrt{n}\mathsf{c}_n^{1/2}} L(\widetilde{B}_n(i))\left[\psi'(-\widetilde{B}_n(i)) \vee \psi'(\widetilde{B}_n(i))\right]\right) \ .$$

Proposition 3.2 then allows us to conclude, by giving us polyLog bounds on $\widetilde{B}_n(i)$. □

4.3 Final conclusions

We can now gather together our approximation results in the following Theorem.

Theorem 4.1. *Under Assumptions **O1-O7** and **P1**, we have, for any fixed* $\tau > 0$,

$$\|\widehat{\beta} - \widetilde{b}\| \leq \mathrm{O}_{L_k}\left(\frac{\textit{polyLog}(n)}{n}\right)$$

In particular,

$$\sqrt{n}(\widehat{\beta}_p - \mathfrak{b}_p) = \mathrm{O}_{L_k}(\textit{polyLog}(n)/\sqrt{n}) \ ,$$
$$\sup_i |X_i'(\widehat{\beta} - \widetilde{b})| = \mathrm{O}_{L_k}\left(\frac{\textit{polyLog}(n)}{\sqrt{n}}\right) \ ,$$
$$\sup_i |R_i - r_{i,[p]}| = \mathrm{O}_{L_k}\left(\frac{\textit{polyLog}(n)}{\sqrt{n}}\right) \ .$$

Proof. The theorem is just the aggregation of all of our results, using the key bound on $\|\widehat{\beta}-\widetilde{b}\|$ in Proposition 4.1.

The last statement is the only one that might need an explanation. With the notations of the proof of Proposition 4.4, we have $R_i - r_{i,[p]} = X_i'(\widetilde{b} - \widehat{\beta}) + \nu_i$. The results in the proof of Proposition 4.4 as well as the bound on $\|\widetilde{b} - \widehat{\beta}\|$ give us the announced result. □

We note that when the vectors X_i's are i.i.d with i.i.d entries, all the coordinates play a symmetric role. In particular,

$$\mathbf{E}\left(\|\widehat{\beta}\|^2\right) = p\mathbf{E}\left(\widehat{\beta}_p^2\right) \ .$$

We now recall that $\widetilde{b}_p = \mathfrak{b}_p$ and remind the reader that $\mathfrak{b}_p = \mathrm{O}_{L_k}(\text{polyLog}(n)/\sqrt{n})$. So using the results of the previous theorem, Equation (36) and summing over all the coordinates, we have, asymptotically,

$$\frac{p}{n}\mathbf{E}\left(\|\widehat{\beta}\|^2\right) = \frac{p^2}{n}\mathbf{E}\left(\widehat{\beta}_p^2\right) = \frac{p^2}{n}\mathbf{E}\left(\mathfrak{b}_p^2\right) + \mathrm{o}(1) = \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left((\mathsf{c}_{\tau,p}\psi(r_{i,[p]})^2\right) + \mathrm{o}(1) \ . \quad (37)$$

Furthermore, when $\{(V_i, \epsilon_i)\}_{i=1}^n$ are exchangeable, it is clear that $r_{i,[p]} \overset{\mathcal{L}}{=} r_{j,[p]}$, by symmetry. Since $\mathsf{c}_{\tau,p}$ does not depend on i, we therefore see that in this case which corresponds to Assumption **F1**,

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left((\mathsf{c}_{\tau,p}\psi(r_{i,[p]})^2\right) = \mathbf{E}\left((\mathsf{c}_{\tau,p}\psi(r_{1,[p]})^2\right) \ .$$

4.3.1 On $\mathsf{c}_{\tau,p}$ and c_τ

Proposition 4.5. *We have*

$$|c_\tau - \mathsf{c}_{\tau,p}| = \mathrm{O}_{L_k}(n^{-1/2} polyLog(n)) \;.$$

Proof. Let us recall the notation

$$S = \frac{1}{n}\sum_{i=1}^n \psi'(R_i) X_i X_i' \;, \text{ and } c_\tau = \frac{1}{n}\mathrm{trace}\left((S+\tau \mathrm{Id}_p)^{-1}\right) \;.$$

If we call $\Gamma = \frac{1}{n}\sum_{i=1}^n \psi'(R_i) V_i V_i'$ and $a = \frac{1}{n}\sum_{i=1}^n \psi'(R_i) X_i^2(p)$, we see that

$$S = \begin{pmatrix} \Gamma & \mathsf{v} \\ \mathsf{v} & a \end{pmatrix} \;.$$

According to Lemma C-2, we have

$$|c_\tau - \frac{1}{n}\mathrm{trace}\left((\Gamma+\tau\mathrm{Id})^{-1}\right)| \leq \frac{1}{n}\frac{1+a/\tau}{\tau} \;.$$

It is clear that under our assumptions, $a = \mathrm{O}_{L_k}(\mathrm{polyLog}(n))$ (using e.g Lemma 3.4). It is also clear that

$$\sup_i |\psi'(R_i) - \psi'(r_{i,[p]})| = \mathrm{O}_{L_k}(\mathrm{polyLog}(n)/\sqrt{n}) \;.$$

Hence, using arguments similar to the ones we have used in the proof of Lemma 4.3 (i.e first resolvent identity, etc...), we see that

$$\left|\frac{1}{n}\mathrm{trace}\left((\Gamma+\tau\mathrm{Id})^{-1}\right) - \frac{1}{n}\mathrm{trace}\left((\mathfrak{S}_p+\tau\mathrm{Id})^{-1}\right)\right| = \mathrm{O}_{L_k}(\mathrm{polyLog}(n)/\sqrt{n}) \;.$$

Since $\mathsf{c}_{\tau,p} = \frac{1}{n}\mathrm{trace}\left((\mathfrak{S}_p+\tau\mathrm{Id})^{-1}\right)$, the result we announced follows immediately. □

In light of this result, we see, using Theorems 3.1 and 4.1 that Equation (37) can be re-written

$$\frac{p}{n}\mathbf{E}\left(\|\widehat{\beta}\|^2\right) = \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left((c_\tau \psi(R_i))^2\right) + \mathrm{o}(1) = \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left((c_i \psi(\mathrm{prox}_{c_i}(\rho)(\tilde{r}_{i,(i)})))^2\right) + \mathrm{o}(1) \;,$$

where we have used the remark we made after Lemma 4.4 that showed that $\sup_i |c_i - c_\tau| = \mathrm{O}_{L_k}(n^{-1/2}\mathrm{polyLog}(n))$. (See also Lemma A-2 and its proof where we compute the derivative of $\mathrm{prox}_c(\rho)(x)$ with respect to c.)

So we finally have:

Proposition 4.6. *Under Assumptions* ***O1-O7*** *and* ***P1****,*

$$\frac{p}{n}\mathbf{E}\left(\|\widehat{\beta}\|^2\right) = \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left((c_\tau \psi(prox_{c_\tau}(\rho)(\tilde{r}_{i,(i)})))^2\right) + \mathrm{o}(1) \;. \tag{38}$$

This will give us the second equation of our system. We also note that for any x, $c_\tau \psi[\mathrm{prox}_{c_\tau}(\rho)(x)] = x - \mathrm{prox}_{c_\tau}(\rho)(x) = \mathrm{prox}_1((c_\tau \rho)^*)(x)$ - see e.g Moreau (1965). In Bean et al. (2013), we found that this formulation was nicer when further analytic manipulations where needed.

If we further assume that (X_i, ϵ_i) are exchangeable and hence play a symmetric role - which is for instance the case when (X_i, ϵ_i)'s are i.i.d - we see that $\tilde{r}_{i,(i)} \overset{\mathcal{L}}{=} \tilde{r}_{j,(j)}$ and hence

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left((c_\tau \psi(\mathrm{prox}_{c_\tau}(\rho)(\tilde{r}_{i,(i)})))^2\right) = \mathbf{E}\left((c_\tau \psi(\mathrm{prox}_{c_\tau}(\rho)(\tilde{r}_{1,(1)})))^2\right)$$

5 Putting things together

5.1 On the asymptotic distribution of $\tilde{r}_{i,(i)}$

We have the following lemma.

Lemma 5.1. *Under Assumptions **O1-O7** and **P1**, as n and p tend to infinity, $\tilde{r}_{i,(i)}$ behaves like $\epsilon_i + \sqrt{\mathbf{E}\left(\|\widehat{\beta}\|^2\right)}Z$, where $Z \sim \mathcal{N}(0,1)$ is independent of ϵ_i, in the sense of weak convergence.*

Furthermore, if $i \neq j$, $\tilde{r}_{i,(i)}$ and $\tilde{r}_{j,(j)}$ are asymptotically independent.

Proof. The only problem is of course showing that $\widehat{\beta}'_{(i)}X_i$ is approximately $\mathcal{N}(0, \mathbf{E}\left(\|\widehat{\beta}\|^2\right))$. Recall that $\widehat{\beta}_{(i)}$ is independent of X_i and that X_i has mean 0, variance 1 and that the third absolute moment of its entries are assumed to be bounded uniformly in n.

We recall that in the proof of Proposition 3.4, we showed that $\mathbf{E}\left(\|\widehat{\beta}\|^2 - \|\widehat{\beta}_{(i)}\|^2\right) \to 0$. Recall that we have also shown that $\text{var}\left(\|\widehat{\beta}\|^2\right) \to 0$ and $\text{var}\left(\|\widehat{\beta}_{(i)}\|^2\right) \to 0$. Note also that our earlier bounds guarantee that $\mathbf{E}\left(\|\widehat{\beta}\|^2\right)$ and $\mathbf{E}\left(\|\widehat{\beta}_{(i)}\|^2\right)$ remain bounded.

The first part of the Lemma will be shown - by appealing to Slutsky's lemma (Lehmann and Casella (1998), Theorem I.8.10) - if we can show that $\widehat{\beta}'_{(i)}X_i$ behaves like $\mathcal{N}(0, \mathbf{E}\left(\|\widehat{\beta}_{(i)}\|^2\right))$.

This follows from a simple generalization of the standard Lindeberg-Feller theorem (see e.g Stroock (1993)). Indeed, if $a_{n,p}(k)$ are random variables with $\sqrt{\sum_{k=1}^p a_{n,p}(k)^2} = A_n$, $\mathbf{E}\left(A_n^2\right)$ remains bounded in n, and $a_{n,p}(k)'s$ are independent of X_i, we see that: a) if $\mathsf{Z} \sim \mathcal{N}(0, \text{Id}_p)$, independent of $a_{n,p}(k)$, then $a'_{n,p}\mathsf{Z} \sim A_n\mathsf{N}$ where $\mathsf{N} \sim \mathcal{N}(0,1)$ and independent of A_n (conditionally and unconditionally on $a_{n,p}$); b) Theorem 2.1.5 and its proof in Stroock (1993) hold provided $\sum_{i=1}^n \mathbf{E}\left(|a_{n,p}(k)|^3\right) = \text{o}(1)$. The proof simply needs to be started conditionally on $a_{n,p}$, and the final moment bounds are then taken unconditionally. This very mild generalization gives, if ϕ is a $\mathcal{C}^3$ function, with bounded 2nd and third derivatives,

$$\forall \epsilon > 0 \,, \left|\mathbf{E}\left(\phi(a'_{n,p}X_i)\right) - \mathbf{E}\left(\phi(A_n\mathsf{N})\right)\right| \leq K\left(\epsilon\|\phi^{(3)}\|_\infty \mathbf{E}\left(\sum_{k=1}^p a_{n,p}(k)^2\right) + \frac{\|\phi^{(2)}\|_\infty}{\epsilon}\sum_{k=1}^p \mathbf{E}\left(|a_{n,p}(k)|^3\right)\right) \,,$$

where K is a constant that depend on the second and third absolute moments of the entries of X_i. It is therefore independent of n and p under our assumptions on X_i.

In our setting, $a_{n,p}(k) = \widehat{\beta}_{(i)}(k)$. Recall that we have shown that

$$\widehat{\beta}_p = \text{O}_{L_k}\left(\frac{\text{polyLog}(n)}{\sqrt{n}\tau}\right) \,.$$

The same arguments we used apply also to $(\widehat{\beta}_{(i)})_p$, the p-th coordinate of the leave-one-out estimate $\widehat{\beta}_{(i)}$. So it is clear that

$$\mathbf{E}\left(|(\widehat{\beta}_{(i)})_p|^3\right) = \text{O}(\text{polyLog}(n)n^{-3/2}) \,.$$

We conclude that $\mathbf{E}\left(\sum_{k=1}^p |(\widehat{\beta}_{(i)})_k|^3\right) = \text{O}(\text{polyLog}(n)n^{-1/2}) = \text{o}(1)$. This, in connection with Corollary 2.1.9 in Stroock (1993), shows that $\widehat{\beta}'_{(i)}X_i$ behaves asymptotically like $\|\widehat{\beta}_{(i)}\|\mathsf{N}$ in the sense of weak convergence.

Since $\|\widehat{\beta}_{(i)}\| - \mathbf{E}\left(\|\widehat{\beta}_{(i)}\|\right) \to 0$ in probability and $\mathbf{E}\left(\|\widehat{\beta}_{(i)}\|\right)$ remains bounded, Slutsky's lemma guarantees that

$$\widehat{\beta}'_{(i)}X_i \text{ behaves like } \mathbf{E}\left(\|\widehat{\beta}_{(i)}\|\right)\mathsf{N}$$

asymptotically, in the sense of weak convergence. (In other words, the difference of the characteristic functions of the random variables on the two sides of the statement above goes to 0 pointwise.)

This shows the first part of the lemma.

Second part For the second part, we use a leave-two-out approach, namely we use the approximation $\tilde{r}_{i,(i)} = \epsilon_i - \widehat{\beta}'_{(i)} X_i = \epsilon_i - \widehat{\beta}'_{(ij)} X_i + \mathrm{O}_{L_k}(\text{polyLog}(n)/(\sqrt{n}\mathsf{c}_n))$ and similarly for $\tilde{r}_{j,(j)}$ (this is clear from Theorem 3.1; $\widehat{\beta}_{(ij)}$ is computed by solving Problem (2) without (X_i, ϵ_i) nor (X_j, ϵ_j)). Let us call $t_i = \epsilon_i - \widehat{\beta}'_{(ij)} X_i$ and $t_j = \epsilon_j - \widehat{\beta}'_{(ij)} X_j$. It is clear that t_i and t_j are independent conditional on $(X_{(ij)})$ and $\{\epsilon_k\}_{k \neq (i,j)} \triangleq \epsilon_{(ij)}$. All we have to do to complete our proof is to show that $\alpha_i = \widehat{\beta}'_{(ij)} X_j$ and $\alpha_j = \widehat{\beta}'_{(ij)} X_i$ are asymptotically independent (the arguments above establish their asymptotic normality). Essentially because their dependence on $X_{(ij)}$ is asymptotically only through $\|\widehat{\beta}_{(ij)}\|$, which is asymptotically deterministic by arguments similar to those used in the proof of Proposition 3.4, we see that t_i and t_j are asymptotically independent. Let us now give a formal proof.

The arguments we gave above apply to $\widehat{\beta}_{ij}$ as they did to $\widehat{\beta}_{(i)}$. In particular, since

$$\mathbf{E}\left(\sum_{k=1}^{p} |(\widehat{\beta}_{(ij)})_k|^3\right) = \mathrm{O}(\text{polyLog}(n) n^{-1/2}) = \mathrm{o}(1) ,$$

we also have

$$\sum_{k=1}^{p} |(\widehat{\beta}_{(ij)})_k|^3 = \mathrm{o}_P(1).$$

Of course, $\widehat{\beta}_{(ij)}$ depends only on $\{X_{(ij)}, \epsilon_{(ij)}\}$. We call $P_{(ij)}$ the joint probability measure $P_{(ij)} = \prod_{k \neq (i,j)} P_{X_k, \epsilon_k}$, i.e probability computed with respect to all our random variables except (X_i, ϵ_i) and (X_j, ϵ_j) (we slightly abuse notation and do not index this probability measure by n for the sake of clarity).

So we have found $E^n_{(ij)}$, depending only on $(X_{(ij)}, \epsilon_{(ij)})$, such that $P_{(ij)}(E^n_{(ij)}) \to 1$ and $\sum_{k=1}^{p} |(\widehat{\beta}_{(ij)})_k|^3 = \mathrm{o}(1)$ when $(X_{(ij)}, \{\epsilon_k\}_{k \neq (i,j)}) \in E^n_{(ij)}$. The arguments we gave above (treating $a_{n,p}$'s as deterministic quantities) then imply that, when $(X_{(ij)}, \epsilon_{(i,j)}) \in E^n_{(ij)}$,

$$\widehat{\beta}'_{(ij)} X_i | (X_{(ij)}, \epsilon_{(ij)}) \text{ behaves like } \|\widehat{\beta}_{(ij)}\| \mathsf{N} .$$

Let us now use characteristic function arguments. Let $(w_i, w_j) \in \mathbb{R}^2$ be fixed and

$$\chi(w_i, w_j) = \mathbf{E}\left(e^{\imath(w_1 \alpha_i + w_2 \alpha_j)}\right) = \mathbf{E}\left(e^{\imath(w_1 \alpha_i + w_2 \alpha_j)} \left[1_{E^n_{(ij)}} + 1_{[E^n_{(ij)}]^c}\right]\right) .$$

Since $P([E^n_{(ij)}]^c) = P_{(ij)}([E^n_{(ij)}]^c) \to 0$, we can just focus on $\mathbf{E}\left(e^{\imath(w_1 \alpha_i + w_2 \alpha_j)} 1_{E^n_{(ij)}}\right)$, since the modulus of the functions we are integrating is bounded by 1.

Now

$$\mathbf{E}\left(e^{\imath(w_1 \alpha_i + w_2 \alpha_j)} 1_{E^n_{(ij)}}\right) = \mathbf{E}\left(1_{E^n_{(ij)}} \mathbf{E}\left(e^{\imath(w_1 \alpha_i + w_2 \alpha_j)} | X_{(ij)}, \epsilon_{(ij)}\right)\right) ,$$

since $1_{E^n_{(ij)}}$ is a deterministic function of $(X_{(ij)}, \epsilon_{(ij)}$.

Now independence of X_i and X_j implies that

$$\mathbf{E}\left(e^{\imath(w_1 \alpha_i + w_2 \alpha_j)} | X_{(ij)}, \epsilon_{(ij)}\right) = \mathbf{E}\left(e^{\imath w_1 \alpha_i} | X_{(ij)}, \epsilon_{(ij)}\right) \mathbf{E}\left(e^{\imath w_2 \alpha_j} | X_{(ij)}, \epsilon_{(ij)}\right) .$$

Also, our conditional asymptotic normality arguments above imply that

$$1_{E^n_{(ij)}} \left[\mathbf{E}\left(e^{\imath w_1 \alpha_i} | X_{(ij)}, \epsilon_{(ij)}\right) - e^{-w_1^2/2 \|\widehat{\beta}_{(ij)}\|^2}\right] \to 0$$

in $P_{(ij)}$-probability. We therefore have

$$1_{E^n_{(ij)}} \left[\mathbf{E}\left(e^{\imath(w_1 \alpha_i + w_2 \alpha_j)} | X_{(ij)}, \epsilon_{(ij)}\right) - e^{-(w_1^2/2 + w_2^2/2) \|\widehat{\beta}_{(ij)}\|^2}\right] \to 0$$

in $P_{(ij)}$-probability.

So we conclude that

$$\mathbf{E}\left(1_{E^n_{(ij)}} e^{\imath(w_1 \alpha_i + w_2 \alpha_j)}\right) - \mathbf{E}\left(1_{E^n_{(ij)}} e^{-(w_1^2/2 + w_2^2/2) \|\widehat{\beta}_{(ij)}\|^2}\right) \to 0 .$$

Since $P(1_{E^n_{(ij)}}) \to 1$ and $\|\widehat{\beta}_{(ij)}\|^2$ is asymptotically deterministic by arguments similar to those used in the proof of Proposition 3.4, we see that

$$\mathbf{E}\left(1_{E^n_{(ij)}} \mathrm{e}^{-(w_1^2/2+w_2^2/2)\|\widehat{\beta}_{(ij)}\|^2}\right) - \mathrm{e}^{-\left[(w_1^2/2+w_2^2/2)\mathbf{E}\left(\|\widehat{\beta}_{(ij)}\|^2\right)\right]} \to 0 \;.$$

Therefore,

$$\mathbf{E}\left(\mathrm{e}^{\imath(w_1\alpha_i+w_2\alpha_j)}\right) - \mathbf{E}\left(\mathrm{e}^{\imath w_1\alpha_i}\right)\mathbf{E}\left(\mathrm{e}^{\imath w_2\alpha_j}\right) \to 0 \;.$$

This proves that α_i and α_j are asymptotically independent. This implies that t_i and t_j are asymptotically independent and so are $\tilde{r}_{i,(i)}$ and $\tilde{r}_{j,(j)}$ using e.g Slutsky's lemma. The lemma is shown.

□

We are now in position to show that $c_\tau = \frac{1}{n}\text{trace}\left((S+\tau \text{Id}_p)^{-1}\right)$ is asymptotically deterministic. We however need the following preliminary result.

Lemma 5.2. *We work under Assumptions **O1-O7**, **P1** and **F2**.*

Consider the random function

$$g_n(x) = \frac{1}{n}\sum_{i=1}^n \frac{1}{1+x\psi'(prox_x(\rho)(\tilde{r}_{i,(i)}))} \;, \textit{ defined for } x \geq 0.$$

Let $B>0$ be in $\mathbb{R}_+$. Call $F_{\rho,B}(u) = ([\psi'(0)+L(|u|)|u|] + BL(|u|)[|\psi(u)|+|\psi(-u)|])$, where $L(|u|)$ is the Lipschitz constant of ψ' on $[-|u|,|u|]$. We have, for any $(x,y)\in\mathbb{R}_+^2$, and $x\leq B$, $y\leq B$

$$\sup_{(x,y):|x-y|\leq\eta, x\leq B, y\leq B} |g_n(x)) - g_n(y)| \leq \eta\frac{1}{n}\sum_{i=1}^n F_{\rho,B}(\tilde{r}_{i,(i)}) \;.$$

In particular, we have

$$P^*\left(\sup_{(x,y):|x-y|\leq\eta, x\leq B, y\leq B} |g_n(x)) - g_n(y)| > \delta\right) \leq \frac{\eta}{\delta}\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left(F_{\rho,B}(\tilde{r}_{i,(i)})\right) \;.$$

Hence, g_n is stochastically equicontinuous on $[0,B]$ for any $B>0$ given, since under our assumptions $\mathbf{E}\left(F_{\rho,B}(\tilde{r}_{i,(i)})\right)$ is uniformly bounded in n,

We used the notation P^* above to denote outer probability and avoid a discussion of potential measure theoretic issues associated with taking a supremum over a non-countable collection of random variables (see e.g van der Vaart (1998), Section 18.2). We refer the reader to e.g Pollard (1984) for more details on stochastic equicontinuity. We note that relying on outer measure arguments to avoid potential measurability issues is standard in the empirical process theory literature (see e.g van der Vaart (1998), Chapter 18).

Proof. Let us consider the function

$$h_u(x) = \frac{1}{1+x\psi'(\text{prox}_x(\rho)(u))} = \frac{\partial}{\partial u}\text{prox}_x(\rho)(u) \;.$$

The last equality comes from Lemma A-3.

We have, since ψ' is non-negative,

$$|h_u(x) - h_u(y)| \leq |x\psi'(\text{prox}_x(\rho)(u)) - y\psi'(\text{prox}_y(\rho)(u))| \wedge 1 \;.$$

Therefore, since $x,y\geq 0$,

$$|h_u(x)-h_u(y)| \leq |x-y|\psi'(\text{prox}_x(\rho)(u)) + y|\psi'(\text{prox}_x(\rho)(u)) - \psi'(\text{prox}_y(\rho)(u))| \;.$$

In particular, if $|x-y|\leq \eta$, and $x\vee y\leq B$

$$\sup_{y:|x-y|\leq \eta;x\vee y\leq B}|h_u(x)-h_u(y)|\leq \eta\psi'(\text{prox}_x(\rho)(u))+B\sup_{y:|x-y|\leq\eta,x\vee y\leq B}|\psi'(\text{prox}_x(\rho)(u))-\psi'(\text{prox}_y(\rho)(u))|\;.$$

Under our assumptions, Lemma A-1 implies that, for $y\geq 0$, $\sup_y|\text{prox}_y(\rho)(u)|\leq |u|$. One of our assumptions is that ψ' is Lipschitz on any $[-t,t]$ with Lipschitz constant $L(t)$. Therefore,

$$|\psi'(\text{prox}_x(\rho)(u))-\psi'(\text{prox}_y(\rho)(u))|\leq L(|u|)|\text{prox}_x(\rho)(u)-\text{prox}_y(\rho)(u)|\;.$$

We recall that, according to Lemma A-2,

$$\frac{\partial}{\partial x}\text{prox}_x(\rho)(u)=-\frac{\psi(\text{prox}_x(\rho)(u))}{1+x\psi'(\text{prox}_x(\rho)(u))}\;.$$

Furthermore, since ψ is non-decreasing and changes sign at 0, we also have

$$\sup_x|\frac{\partial}{\partial x}\text{prox}_x(\rho)(u)|\leq|\psi(u)|\vee|\psi(-u)|\;.$$

This naturally gives us a bound on the Lipschitz constant of the function $x\rightarrow\text{prox}_x(\rho)(u)$. We finally conclude that

$$|\psi'(\text{prox}_x(\rho)(u))-\psi'(\text{prox}_y(\rho)(u))|\leq L(|u|)[|\psi(u)|\vee|\psi(-u)|]|x-y|\;.$$

We therefore have, when $x\vee y\leq B$

$$\sup_{y:|x-y|\leq\eta}|h_u(x)-h_u(y)|\leq\eta\psi'(\text{prox}_x(\rho)(u))+BL(|u|)[|\psi(u)|\vee|\psi(-u)|]\eta\;.$$

Of course, $\psi'(\text{prox}_x(\rho)(u))\leq\psi'(0)+L(|u|)|u|$, by using again $|\text{prox}_x(\rho)(u)|\leq|u|$, $\text{prox}_x(\rho)(0)=0$ and the fact that the Lipschitz constant of ψ' on $[-|\text{prox}_x(\rho)(u)|,|\text{prox}_x(\rho)(u)|]$ is less than $L(u)$.

Therefore, if when $x\vee y\leq B$ we have

$$\sup_{y:|x-y|\leq\eta}|h_u(x)-h_u(y)|\leq\eta\left([\psi'(0)+L(|u|)|u|]+BL(|u|)[|\psi(u)|+|\psi(-u)|]\right)\;.$$

Therefore, we also have

$$\sup_{(x,y):|x-y|\leq\eta,x\vee y\leq B}|h_u(x)-h_u(y)|\leq\eta\left([\psi'(0)+L(|u|)|u|]+BL(|u|)[|\psi(u)|+|\psi(-u)|]\right)\;.$$

We denote by $F_{\rho,B}(u)=([\psi'(0)+L(|u|)|u|]+BL(|u|)[|\psi(u)|+|\psi(-u)|])$. This analysis shows that for x given, if $|x-y|\leq\eta$ and $x\vee y\leq B$, we have

$$\sup_{(x,y):|x-y|\leq\eta,x\leq B,y\leq B}|g_n(x))-g_n(y)|\leq\eta\frac{1}{n}\sum_{i=1}^n F_{\rho,B}(\tilde{r}_{i,(i)})\;.$$

We can now take expectations, and get the result in L_1 provided $\mathbf{E}\left(F_{\rho,B}(\tilde{r}_{i,(i)})\right)$ is finite and remains bounded in n. However, this holds since $F_{\rho,B}$ grows at most polynomially at ∞, and ϵ_i, $\|\widehat{\beta}_{(i)}\|$ and X_i have bounded moments for any given order, by Assumptions **O4**, **F2** and our work on $\|\widehat{\beta}\|$.

We have established stochastic equicontinuity of $g_n(x)$ on $[0,B]$. □

Lemma 5.3. *Let us call $G_n(x)=\mathbf{E}(g_n(x))$. Let $B>0$ be given. For any given $x_0\leq B$,*

$$g_n(x_0)-G_n(x_0)=\text{o}_{L_2}(1)\;.$$

*Under our assumptions **O1-O7**, **P1** and **F2**, we also have*

$$\mathbf{E}^*\left(\sup_{0\leq x\leq B}|g_n(x)-G_n(x)|\right)\rightarrow 0\;.$$

Note that when (X_i, ϵ_i)'s are further assumed to be i.i.d, $G_n(x)$ can be written as the expectation of a bounded continuous function of $\tilde{r}_{1,(1)}$, by symmetry between the $\tilde{r}_{i,(i)}$'s.

Proof. Asymptotic pairwise independence of $\tilde{r}_{i,(i)}$ implies that

$$\mathrm{var}\left(g_n(x_0)\right) \to 0$$

and therefore gives the first result.

Let us pick $\epsilon > 0$. By the stochastic equicontinuity of g_n and our L_1 bound, we can find $x_1, \ldots, x_K$, independent of n, such that for all $x \in [0, B]$, there exists l such that, when n is large enough,

$$\mathbf{E}\left(|g_n(x) - g_n(x_l)|\right) \leq \epsilon \ .$$

Note that

$$|g_n(x) - G_n(x)| \leq |g_n(x) - g_n(x_l)| + |g_n(x_l) - G_n(x_l)| + |G_n(x_l) - G_n(x)| \ .$$

We immediately get

$$\mathbf{E}^*\left(\sup_{0 \leq x \leq B} |g_n(x) - G_n(x)|\right) \leq 2\epsilon + \mathbf{E}\left(\sup_{1 \leq l \leq K} |g_n(x_l) - G_n(x_l)|\right) \ .$$

Because K is finite, the fact that for all l, $|g_n(x_l) - G_n(x_l)| \to 0$ in L_2 implies that $\sup_{1 \leq l \leq K} |g_n(x_l) - G_n(x_l)| \to 0$ in L_2. In particular, if n is sufficiently large,

$$\mathbf{E}\left(\sup_{1 \leq l \leq K} |g_n(x_l) - G_n(x_l)|\right) \leq \epsilon \ .$$

The lemma is shown. □

Lemma 5.4. *Call $c_\tau = \frac{1}{n}\mathrm{trace}\left((S + \tau \mathrm{Id}_p)^{-1}\right)$. Call as before*

$$g_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1 + x\psi'(\mathrm{prox}_x(\rho)(\tilde{r}_{i,(i)}))}$$

Then c_τ is a near solution of

$$\begin{aligned} &\frac{p}{n} - \tau x - 1 + g_n(x) = 0 \ , \ \textit{i. e} \\ &\frac{p}{n} - \tau c_\tau - 1 + g_n(c_\tau) = \mathrm{o}_{L_k}(1) \ . \end{aligned}$$

Asymptotically, near solutions of

$$\delta_n(x) \triangleq \frac{p}{n} - \tau x - 1 + g_n(x) = 0 \ ,$$

are close to solutions of

$$\Delta_n(x) = \frac{p}{n} - \tau x - 1 + \mathbf{E}\left(g_n(x)\right) = 0 \ .$$

More precisely, call $T_{n,\epsilon} = \{x : |\Delta_n(x)| \leq \epsilon\}$. Note that $T_{n,\epsilon} \subseteq (0, p/(n\tau) + \epsilon/\tau)$. For any given ϵ, as $n \to \infty$, near solutions of $\delta_n(x_n) = 0$ belong to $T_{n,\epsilon}$ with high-probability.

*Our assumptions concerning the distribution of ϵ_i's, specifically **F1**, guarantee that as $n \to \infty$, there is a unique solution to $\Delta_n(x) = 0$.*

Hence c_τ is asymptotically deterministic.

We note that the deterministic equivalent of c_τ is $c_\rho(\kappa)$ in Theorem 1.1.

Proof. Let δ_n be the function

$$\delta_n(x) = \frac{p}{n} - \tau x - 1 + g_n(x) \ ,$$

and $\Delta_n(x) = \mathbf{E}\left(\delta_n(x)\right)$. Call x_n a solution $\delta_n(x_n) = 0$ and $x_{n,0}$ a solution of $\Delta_n(x_{n,0}) = 0$.

These solutions exist, since δ_n is continuous, $\delta_n(0) = p/n > 0$ and $\delta_n(p/(n\tau)) \leq 0$. The same arguments apply for Δ_n.

Since $0 \leq g_n \leq 1$, we see that $x_n \leq p/(n\tau)$, for otherwise, $\delta_n(x) < 0$. The same argument shows that if $x > (p/n + \epsilon)/\tau$, $\Delta_n(x) < -\epsilon$ and $x \notin T_{n,\epsilon}$. Similarly, near solutions of $\delta_n(x) = 0$ must be less or equal to $(p/n + \epsilon)/\tau$.

- **Proof of the fact that c_τ is such that $\delta_n(c_\tau) = \mathrm{o}_P(1)$**

An important remark is that c_τ is a near solution of $\delta_n(x) = 0$. This follows most clearly for arguments we have developed for $\mathsf{c}_{\tau,p}$ so we start by giving details through arguments for this random variable. Recall that in the notation of Lemma 4.2, we had

$$\frac{p-1}{n} - \tau\mathsf{c}_{\tau,p} = \frac{1}{n}\mathrm{trace}\left(\mathrm{Id}_n - M\right).$$

Now, according to Equation (32),

$$\frac{1}{n}\mathrm{trace}\left(\mathrm{Id}_n - M\right) = 1 - \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1 + \psi'(r_{i,[p]})\frac{1}{n}V_i'(\mathfrak{S}_p(i) + \tau\mathrm{Id})^{-1}V_i}.$$

According to Lemmas 4.3 and 4.4, we have

$$\sup_i \left|\frac{1}{n}V_i'(\mathfrak{S}_p(i) + \tau\mathrm{Id})^{-1}V_i - \mathsf{c}_{\tau,p}\right| = \mathrm{O}_{L_k}(\mathrm{polyLog}(n)n^{-1/2}).$$

Of course, when $x \geq 0$ and $y \geq 0$, $|1/(1+x) - 1/(1+y)| \leq |x-y| \wedge 1$. Using our bounds on $\psi'(r_{i,[p]})$, we easily see that,

$$p/n - \tau\mathsf{c}_{\tau,p} - 1 + \frac{1}{n}\sum_{i=1}^{n}\frac{1}{1 + \mathsf{c}_{\tau,p}\psi'(r_{i,[p]})} = \mathrm{O}_{L_k}(n^{-1/2}\mathrm{polyLog}(n)) \ .$$

Exactly the same computations can be made with c_τ, so we have established that

$$p/n - \tau c_\tau - 1 + \frac{1}{n}\sum_{i=1}^{n}\frac{1}{1 + c_\tau\psi'(R_i)} = \mathrm{O}_{L_k}(n^{-1/2}\mathrm{polyLog}(n)) \ . \tag{39}$$

Now we have seen in Theorem 3.1 that

$$\sup_i |R_i - \mathrm{prox}_{c_i}(\rho)(\tilde{r}_{i,i})| = \mathrm{O}_{L_k}(n^{-1/2}\mathrm{polyLog}(n)) \ .$$

Through our assumptions on ψ', this of course implies that

$$\sup_i |\psi'(R_i) - \psi'[\mathrm{prox}_{c_i}(\rho)(\tilde{r}_{i,i})]| = \mathrm{O}_{L_k}(n^{-1/2}\mathrm{polyLog}(n))$$

We have furthermore noted that $\sup_i |c_i - c_\tau| = \mathrm{O}_{L_k}(n^{-1/2}\mathrm{polyLog}(n))$ after Lemma 4.4. Using the proof of Lemma A-2, this implies that

$$\sup_i \left|\psi'[\mathrm{prox}_{c_i}(\rho)(\tilde{r}_{i,i})] - \psi'[\mathrm{prox}_{c_\tau}(\rho)(\tilde{r}_{i,i})]\right| = \mathrm{O}_{L_k}(n^{-1/2}\mathrm{polyLog}(n))$$

and therefore

$$\sup_i \left|\psi'[R_i] - \psi'[\mathrm{prox}_{c_\tau}(\rho)(\tilde{r}_{i,i})]\right| = \mathrm{O}_{L_k}(n^{-1/2}\mathrm{polyLog}(n))$$

So we have established that $\delta_n(c_\tau) = \mathrm{O}_{L_k}(n^{-1/2}\mathrm{polyLog}(n))$.

- **Final details**

Note that for any given x, $\delta_n(x) - \Delta_n(x) = \mathrm{o}_P(1)$ by using Lemma 5.3. In our case, with the notation of this lemma, $B = p/(n\tau) + \eta/\tau$, for $\eta > 0$ given.

This implies that, for any given $\epsilon > 0$

$$\sup_{x \in (0, p/(n\tau)+\eta/\tau]} |\delta_n(x) - \Delta_n(x)| < \epsilon \ ,$$

with high-probability when n is large. Therefore, for any $\epsilon > 0$

$$|\Delta_n(x_n)| \leq \epsilon$$

with high-probability. This exactly means that $x_n \in T_{n,\epsilon}$ with high-probability. The same argument applies for near solutions of $\delta_n(x) = 0$, which, for any $\epsilon > 0$ must belong to $T_{n,\epsilon}$ as $n \to \infty$ with high-probability. Of course, there is nothing random about $T_{n,\epsilon}$ which is a deterministic set. Note that $T_{n,\epsilon}$ is compact because it is bounded and closed, using the fact that g_n and $\mathbf{E}(g_n)$ are continuous.

If $T_{n,0}$ were reduced to a single point, we would have established the asymptotically deterministic character of c_τ.

Given our work concerning the limiting behavior of $\tilde{r}_{i,(i)}$ and our assumptions about ϵ_i's, we see that Lemma C-1 applies to $\lim_{n\to\infty} \Delta_n(x)$ under assumption **F1**. Therefore, as $n \to \infty$, $T_{n,0}$ is reduced to a point and c_τ is asymptotically non-random. □

Proof of Theorem 1.1

As we had noted in El Karoui et al. (2011),

$$\frac{\partial}{\partial t}\mathrm{prox}_c(\rho)(t) = \mathrm{prox}_c(\rho)'(t) = \frac{1}{1 + c\psi'(\mathrm{prox}_c(\rho)(t))} \ .$$

So Δ_n can be interpreted as

$$\Delta_n(x) = \frac{p}{n} - \tau x - 1 + \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left(\mathrm{prox}_x(\rho)'(\tilde{r}_{i,(i)})\right) \ .$$

The fact that c_τ is asymptotically arbitrarily close to the root of $\Delta_n(x) = 0$ gives us the first equation in the system appearing in Theorem 1.1.

The second equation of the system comes from Equation (38). Theorem 1.1 is shown under Assumptions **O1-O7**, **P1** and **F1-F2**.

6 Extensions

6.1 From the $\tau > 0$ case to the case $\tau = 0$

Our original motivation in El Karoui et al. (2011) and El Karoui et al. (2013) was to study the "unpenalized" problem, namely $\widehat{\beta}(\{Y_i, X_i\})$ was defined as

$$\widehat{\beta}(\{Y_i, X_i\}) = \mathrm{argmin}_\beta \frac{1}{n}\sum_{i=1}^n \rho(Y_i - X_i'\beta) \ .$$

We now explain how we can derive results in the unpenalized case from the ones we have obtained in the penalized case, i.e $\tau > 0$, when $p/n < 1$ and $Y_i = X_i'\beta_0 + \epsilon_i$.

We first note that when $p < n$, and when X_i's are such that $\mathrm{span}\{X_i\}_{i=1}^n = \mathbb{R}^p$, if $Y_i = X_i'\beta_0 + \epsilon_i$,

$$\widehat{\beta}(\{Y_i, X_i\}) - \beta_0 = \mathrm{argmin}_\beta \frac{1}{n}\sum_{i=1}^n \rho(\epsilon_i - X_i'\beta) \triangleq \widehat{\beta} \ ,$$

essentially by a change of variables (see El Karoui et al. (2011) and El Karoui et al. (2012) for details if needed). So to understand the error we make when using regression M-estimates, i.e the vector $\widehat{\beta}(Y_i, X_i) -$

β_0, it is enough to study the properties of the estimator $\widehat{\beta}$. In other words, we simply need to understand the null case of the problem, i.e $\beta_0 = 0$. Of course, we have previously studied the penalized version of this particular problem.

(Note that under Assumptions **O4** and **P1**, the classic result of Bai and Yin (1993) applies, which guarantees that span$\{X_i\}_{i=1}^n = \mathbb{R}^p$ with probability going to 1. If X_i's have e.g continuous distributions, this is guaranteed non-asymptotically with probability 1.)

The next result requires the notion of modulus of convexity of a function. We refer the reader to Proposition 1.1.2 on p. 73 in Hiriart-Urruty and Lemaréchal (2001) for a definition. When ρ is twice differentiable, the modulus of convexity is a lower bound on its second derivative (see Theorem 4.3.1 on p. 115 in Hiriart-Urruty and Lemaréchal (2001)).

We have the following theorem.

Theorem 6.1. *Suppose our assumptions **O1-O7**, **P1** and **F1-F2** hold. Call, for $\tau > 0$,*

$$\widehat{\beta}_\tau = \text{argmin}_\beta \frac{1}{n}\sum_{i=1}^n \rho(\epsilon_i - X_i'\beta) + \tau\frac{\|\beta\|^2}{2} .$$

When $\tau_1, \tau_2 > 0$, we have

$$\|\widehat{\beta}_{\tau_1} - \widehat{\beta}_{\tau_2}\| \leq \frac{\sqrt{2}|\tau_2 - \tau_1|}{\tau_2\sqrt{\tau_1}}\sqrt{\frac{1}{n}\sum_{i=1}^n \rho(\epsilon_i)} .$$

Suppose further that $\limsup p/n < 1$ *and ρ is strongly convex with modulus of convexity C. We have, if*

$$\widehat{\beta} = \text{argmin}_\beta \frac{1}{n}\sum_{i=1}^n \rho(\epsilon_i - X_i'\beta) ,$$

$$\|\widehat{\beta}_\tau - \widehat{\beta}\| \leq \frac{\sqrt{2\tau}}{C\lambda_{\min}(\widehat{\Sigma})}\sqrt{\frac{1}{n}\sum_{i=1}^n \rho(\epsilon_i)} .$$

Hence, under our assumptions, as n and p tend to infinity,

$$\lim_{\tau\to 0}\|\widehat{\beta}_\tau - \widehat{\beta}\| = \mathrm{o}_P(1) .$$

Recall that under our assumptions, if τ is fixed and $p/n \to \kappa$, $0 < \kappa < \infty$, $\lim_{n,p\to\infty} |\|\widehat{\beta}_\tau\| - r_\rho(\kappa;\tau)| = 0$, where $r_\rho(\kappa;\tau)$ is deterministic and characterized by System (3)*.*

Hence, when $\kappa < 1$, $\|\widehat{\beta}\|$ is asymptotically deterministic and we have, if $r_\rho(\kappa; 0) = \lim_{\tau\to 0} r_\rho(\kappa;\tau)$,

$$\lim_{n,p\to\infty} |\|\widehat{\beta}\| - r_\rho(\kappa; 0)| \to 0 \textit{ in probability} .$$

Proof. We call $f_\tau(\beta) = -\frac{1}{n}\sum_{i=1}^n X_i\psi(\epsilon_i - X_i'\beta) + \tau\beta$. Note that for any $\tau_1, \tau_2 \geq 0$,

$$f_{\tau_2}(\beta) = f_{\tau_1}(\beta) + (\tau_2 - \tau_1)\beta .$$

Hence, $f_{\tau_2}(\widehat{\beta}_{\tau_1}) = (\tau_2 - \tau_1)\widehat{\beta}_{\tau_1}$. Using Proposition 2.1 with f_{τ_2} playing the role of f, we have

$$\|\widehat{\beta}_{\tau_2} - \widehat{\beta}_{\tau_1}\| \leq \frac{|\tau_2 - \tau_1|}{\tau_2}\|\widehat{\beta}_{\tau_1}\| .$$

We now turn to approximations when ρ is strongly convex. Since by definition, $\widehat{\beta}$ is such that

$$\sum_{i=1}^n X_i\psi(\epsilon_i - X_i'\widehat{\beta}) = 0 ,$$

we see that $f_\tau(\widehat{\beta}) = \tau\widehat{\beta}$. By a similar token, we see that $f_0(\widehat{\beta}_\tau) = -\tau\widehat{\beta}_\tau$.

If ρ is strongly convex with modulus of convexity C, we see, using Proposition 2.1 that by working with f_0 - along the same lines as in the proof of Proposition 2.1 - we get

$$\|\widehat{\beta}_\tau - \widehat{\beta}\| \leq \frac{1}{C\lambda_{\min}(\widehat{\Sigma})}\|f_0(\widehat{\beta}_\tau)\| = \frac{\tau}{C\lambda_{\min}(\widehat{\Sigma})}\|\widehat{\beta}_\tau\| \ .$$

Recall that we showed in Equation (18) that

$$\|\widehat{\beta}_\tau\| \leq \sqrt{\frac{2}{\tau}}\sqrt{\frac{1}{n}\sum_{i=1}^n \rho(\epsilon_i)} \ .$$

This shows that

$$\|\widehat{\beta}_{\tau_2} - \widehat{\beta}_{\tau_1}\| \leq \frac{|\tau_2 - \tau_1|}{\tau_2\sqrt{\tau_1}}\sqrt{\frac{1}{n}\sum_{i=1}^n \rho(\epsilon_i)} \ , \text{ and}$$

$$\|\widehat{\beta}_\tau - \widehat{\beta}\| \leq \frac{\sqrt{2\tau}}{C\lambda_{\min}(\widehat{\Sigma})}\sqrt{\frac{1}{n}\sum_{i=1}^n \rho(\epsilon_i)} \ .$$

Under our assumptions, $\frac{1}{n}\sum_{i=1}^n \rho(\epsilon_i) = \mathrm{O}_P(1)$. Under the assumptions that, for instance, the entries of X_i's are i.i.d mean 0, variance 1, with 4+ϵ moments (which is always the case under our assumptions), it is well known that $\lambda_{\min}(\widehat{\Sigma}) \to (1 - \sqrt{\frac{p}{n}})^2$ in probability and a.s (Bai (1999)).

We conclude that $\|\widehat{\beta}_\tau - \widehat{\beta}\| \to 0$ in probability as $\tau \to 0$ under the assumptions stated in the theorem. It is also clear that the mapping $\tau \mapsto \|\widehat{\beta}_\tau\|$ is continuous on $[0, \infty)$ with probability going to 1 as $n, p \to \infty$, while $p/n \to \kappa < 1$. Furthermore, $\|\widehat{\beta}_\tau\|$ is bounded in probability on $[0, \infty)$.

The other statements in the theorem follow easily. For instance, to show that $r_\rho(\kappa; \tau)$ is right-continuous at 0 when $\kappa < 1$, we can use the bound

$$\begin{aligned}|r_\rho(\kappa;\tau_1) - r_\rho(\kappa;\tau_2)| \leq{}& |r_\rho(\kappa;\tau_1) - \|\widehat{\beta}_{\tau_1}\|| + ||\widehat{\beta}_{\tau_1}\| - \|\widehat{\beta}\|| \\ &+ |r_\rho(\kappa;\tau_2) - \|\widehat{\beta}_{\tau_2}\|| + ||\widehat{\beta}_{\tau_2}\| - \|\widehat{\beta}\|| \ .\end{aligned}$$

Given our previous results and bounds, it is clear that for any given $\epsilon > 0$, we can find $\delta > 0$, such that if τ_1 and $\tau_2 > 0$ are less than δ, the right hand side is less than ϵ with probability going to 1 as $n, p \to \infty$ while $p/n \to \kappa < 1$. So the left-hand side, which is deterministic, is smaller than a random variable which is less than ϵ with probability going to 1. The left-hand side must therefore be less than ϵ. This shows that $\tau \mapsto r_\rho(\kappa; \tau)$ is right-continuous at 0. Therefore $r_\rho(\kappa; 0) = \lim_{\tau\to 0} r_\rho(\kappa; \tau)$ is well defined.

The fact that $||\widehat{\beta}\| - r_\rho(\kappa; 0)| = \mathrm{o}_P(1)$ follows by similar bounds and arguments. □

Under for instance Gaussian design assumptions (i.e X_i's have distribution $\mathcal{N}(0, \mathrm{Id}_p)$), it is possible to bound $\mathbf{E}\left(1/\lambda_{\min}(\widehat{\Sigma})\right)$ using essentially results in Silverstein (1985) as well as elementary but non-trivial linear algebra (see the appendix of Halko et al. (2011) for instance). This would give an approximation in L_2, provided the random variable $\rho(\epsilon_i)$ has enough moment.

It seems possible with quite a bit of extra work to dispense with the assumption of strong convexity - see Appendix D-3 for a brief discussion.

We note that convergence in probability of $\widehat{\beta}$ is enough for our confidence interval statements from Bean et al. (2013) (details in the supplementary material of that paper) to go through. This is quite important from the standpoint of statistical applications.

Finally, now that the probabilistic properties of $\widehat{\beta}$ and in particular its norm are well-understood through regularization techniques, the simplest way to get an analog of Theorem 3.1 under Assumptions **O1-O7**, **P1**, **F1-F2**, is to go through Section 3 without regularization, i.e using $\tau = 0$, and hence relying on the second bound in Proposition 2.1. All the approximations now hold with high-probability (since they involve the smallest eigenvalue of a sample covariance matrix), but this is good enough to describe the marginal behavior of the residuals.

6.2 Other extensions

Heteroskedastic errors ϵ_i's

In El Karoui et al. (2013), we considered many extensions of the basic problem, including that of ϵ_i's with different distributions.

Our approximation results make essentially no use of the assumption that ϵ_i's have the same distribution (Assumption **F1** is only used in Lemma 5.4 at the very end of the proof). So all of our approximations will go through for this more general case.

A case of particular interest is when ϵ_i's are chosen from one distribution with probability $1-\alpha$ and from another one with probability α - i.e there is "α-contamination" (Huber and Ronchetti (2009), Section 4.5). (Various symmetry arguments that appear in the proof will go through in this case, by applying them to each subgroup of observations.)

Clearly the System (3) is changed then. And indeed, one key situation where we assume that ϵ_i's have the same distribution is in Lemma 5.4, when we show that $\Delta_n(x)$ has asymptotically a unique zero. We now explain briefly how to take care of this problem in the heteroskedastic case, i.e ϵ_i's with different distribution.

If ϵ_i's are independent and such that $W_{i,r}=\epsilon_i+Z_r$, where $Z_r\sim\mathcal{N}(0,r^2)$ is independent of ϵ_i, each satisfy the conditions of Lemma C-1, it is clear that the same Lemma applies to

$$F_n(x)=\frac{p}{n}-\tau x-1+\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left((\text{prox}_x(\rho))'(W_{i,r})\right) ,$$

since this function turns out to be decreasing as an average of decreasing functions.

Provided ϵ_i's are such that $F_n(x)$ above has a limit (which is decreasing), all of our arguments concerning the first equation of the system in Theorem 1.1 will go through.

The key function in the second equation of the system becomes in the heteroskedastic case

$$G_n(x)=\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left([W_{i,r}-\text{prox}_x(\rho)(W_{i,r})]^2\right) .$$

If this function has a limit as $n\to\infty$, our arguments concerning the second equation of the system will go through.

The functions F_n and G_n in the α-contamination discussed above are very well-behaved and so our arguments can be adapted to this interesting situation.

More general assumptions on the X_i's

Our proof makes very strong use of the independence of the predictor vectors X_i's in Section 3 and of the independence of the entries of X_i's in Section 4. We also use strongly the fact that we assume that the entries of each X_i have mean 0 and variance 1.

But we do not use strongly the assumption that the entries of X_i have the same distribution, and it seems that most arguments go through without this assumption. The passage of Equation (36) to Equation (37) appeals to the assumption of i.i.d-ness of the entries of X_i through a symmetry argument, but this could be avoided at the cost of a little bit more technical work. So it is clear that, with a bit more work, our approach could handle the case where X_i's have independent but not identically entries.

Moving from random vectors X_i's like the ones we have studied to vectors of the form $\tilde{X}_i=\lambda_i X_i$, where λ_i are independent, mean 0, variance 1 random variables (i.e scalar) independent of X_i does not offer any new technical difficulties. Indeed, our work in El Karoui et al. (2011) and El Karoui et al. (2013) handled - heuristically - that case, so the arguments we gave here would be fairly easy to modify.

This extended class of models - which is akin to elliptical distributions in multivariate statistics (see Anderson (2003)) - is interesting because it includes distributions that do not share the geometric properties that "concentrated" random vectors have in common. In particular, elliptical distributions show clearly that there is no hope to have universality results that are meaningful from a statistical point of view in the problems we have studied. We refer the interested reader to El Karoui et al. (2013) and e.g Diaconis and Freedman (1984), Hall et al. (2005), El Karoui (2009), El Karoui (2010) and El Karoui and Koesters (2011) for discussion of these matters in various contexts.

We do not solve the elliptical problem here in complete details because of the extra notational burden involved. We just note that it appears that if λ_i's are bounded, and so are $1/\lambda_i$'s, the results of Section 3 go through. In Section 4, a number of small adjustments are needed and easy to make in light of the arguments in El Karoui et al. (2013). For instance, in Lemma 4.2, instead of subtracting $\mathsf{c}_{\tau,p}$ in the definition of η_i, one would have to subtract $\lambda_i^2 \mathsf{c}_{\tau,p}$. The system (3) would also change, as indicated in El Karoui et al. (2013).

Other extensions

Another natural extension of the work presented here is to study the weighted regression case, i.e for weights $\{w_i\}_{i=1}^n$, $\widehat{\beta}$ is defined as

$$\widehat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n w_i \rho(\epsilon_i - X_i' \beta) + \frac{\tau}{2} \|\beta\|^2 \;.$$

Once again, only minor modifications seem needed to our proof - the conceptual difficulties were dealt with in El Karoui et al. (2013). More generally, working on the problem of understanding

$$\widehat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \rho_i(\epsilon_i - X_i' \beta) + \frac{\tau}{2} \|\beta\|^2 \;,$$

where ρ_i are potentially different functions and X_i's are "elliptical" (as defined above) seems to be within relatively easy reach of the method developed and presented here.

As in the case of errors with different distributions, the main issue appears to be to make sure that the functions that appear in the limiting system have the properties we require.

Finally, we see that when $Y_i = X_i'\beta_0 + \epsilon_i$, it would be interesting to understand better, for p and n large, but p possibly larger than n,

$$\widehat{\beta}_\tau = \text{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \rho_i(Y_i - X_i' \beta) + \frac{\tau}{2} \|\beta\|^2 \;,$$

i.e the "ridge-regularized robust regression" problem when the responses Y_i are not independent of the predictors X_i. This problem - a variant of the one we have studied here - should be amenable to analysis with the method we used here. Some difficulties, both conceptual and technical arise however, owing to the fact that when $p/n > 1$, $\text{span}\{X_i\}_{i=1}^n \neq \mathbb{R}^p$, which prevents the use of some of the reduction to the null case ideas we employed in Section 6.1. We are currently working on these extensions.

APPENDIX

A Notes on the proximal mapping

Lemma A-1. *Let ρ be differentiable and such that ψ changes sign at 0, i.e $sign(\psi(x)) = sign(x)$ for $x \neq 0$. Then,*

$$prox_c(\rho)(0) = 0 \ .$$

Furthermore,

$$|\psi(prox_c(\rho)(x))| \leq |\psi(x)| \ .$$

Also,

$$|\psi(prox_c(\rho)(x))| \leq |x|/c \ .$$

Proof. By definition, we have

$$\text{prox}_c(\rho)(x) + c\psi(\text{prox}_c(\rho)(x)) = x \ .$$

Therefore,

$$\text{prox}_c(\rho)(0) = -c\psi(\text{prox}_c(\rho)(0)) \ .$$

Hence, if we call $y = \text{prox}_c(\rho)(0)$, we have $\text{sign}(y) = -\text{sign}(\psi(y))$. The assumptions on ψ therefore guarantee that $y = 0$, for otherwise we would have a contradiction.

We also note that $\text{sign}(\text{prox}_c(\rho)(x)) = \text{sign}(x)$, since $\text{sign}[\psi(\text{prox}_c(\rho)(x))] = \text{sign}(\text{prox}_c(\rho)(x)))$.

Using contractivity of the prox (see Moreau (1965)), we see that

$$|\text{prox}_c(\rho)(x)| = |\text{prox}_c(\rho)(x) - \text{prox}_c(\rho)(0)| \leq |x| \ .$$

Since ρ is convex, we see that ψ is non-decreasing. If $x > 0$, $\text{prox}_c(\rho)(x) > 0$, and therefore,

$$0 \leq \psi(\text{prox}_c(\rho)(x)) \leq \psi(x) \ .$$

Similarly, if $x < 0$, $x \leq \text{prox}_c(\rho)(x) < 0$ and therefore, $\psi(x) \leq \psi(\text{prox}_c(\rho)(x)) \leq 0$. The second statement of the lemma is shown.

The last statement is a simple consequence of the fact that $c\psi(\text{prox}_c(\rho)(x)) = x - \text{prox}_c(\rho)(x)$, from which it immediately follows that

$$|\psi(\text{prox}_c(\rho)(x))| \leq \frac{|x|}{c} \ .$$

□

We will also need the following simple result.

Lemma A-2. *Suppose x is a real and ρ is twice differentiable and convex. Then, for $c > 0$, we have*

$$\frac{\partial}{\partial c} prox_c(\rho)(x) = -\frac{\psi(prox_c(\rho)(x))}{1 + c\psi'(prox_c(\rho)(x))} \ .$$

and

$$\frac{\partial}{\partial c} \rho(prox_c(\rho)(x)) = -\frac{\psi^2(prox_c(\rho)(x))}{1 + c\psi'(prox_c(\rho)(x))} \ .$$

In particular, at x given $c \to \rho(prox_c(\rho)(x))$ is decreasing in c.

Proof. Using the fact that

$$\operatorname{prox}_c(\rho)(x) + c\psi(\operatorname{prox}_c(\rho)(x)) = x,$$

we easily see that

$$\frac{\partial}{\partial c}\operatorname{prox}_c(\rho)(x) = -\frac{\psi(\operatorname{prox}_c(\rho)(x))}{1 + c\psi'(\operatorname{prox}_c(\rho)(x))} .$$

It then follows immediately that

$$\frac{\partial}{\partial c}\rho(\operatorname{prox}_c(\rho)(x)) = -\frac{\psi^2(\operatorname{prox}_c(\rho)(x))}{1 + c\psi'(\operatorname{prox}_c(\rho)(x))} .$$

The denominator is positive, from which we immediately deduce that $c \to \rho(\operatorname{prox}_c(\rho)(x))$ is decreasing in c. □

We also make the following observation, which was essential to finding the system of equations (3) in El Karoui et al. (2013)

Lemma A-3. *We have*

$$\frac{\partial}{\partial x}\mathit{prox}_c(\rho)(x) = \frac{1}{1 + c\psi'(\mathit{prox}_c(\rho)(x))} .$$

Moreover, at c fixed, when ψ' is continuous, $x \to \frac{1}{1+c\psi'(\mathit{prox}_c(\rho)(x))}$ is a bounded, continuous function of x.

A proof of the first fact follows immediately from the well-known representation (see Moreau (1965))

$$\operatorname{prox}_c(\rho)(x) = (\mathrm{Id} + c\psi)^{-1}(x) .$$

The second result is also immediate, since $\psi' \geq 0$.

We finally make notice of the following simple fact.

Lemma A-4. *The function $c \to [c\psi(\mathit{prox}_c(\rho)(x))]^2$ (defined on $\mathbb{R}_+$) is increasing, for any x.*

Proof. Let us consider $f_x(c) = c\psi(\operatorname{prox}_c(\rho)(x))$. Note that $f_x(c) = x - \operatorname{prox}_c(\rho)(x)$. So

$$\frac{\partial}{\partial c}f_x(c) = \frac{\psi(\operatorname{prox}_c(\rho)(x))}{1 + c\psi'(\operatorname{prox}_c(\rho(x)))} .$$

Hence,

$$\frac{\partial}{\partial c}f_x^2(c) = 2c\frac{\psi^2(\operatorname{prox}_c(\rho)(x))}{1 + c\psi'(\operatorname{prox}_c(\rho(x)))} \geq 0$$

since $c \geq 0$ and $\psi' \geq 0$. □

Examples : for the sake of concreteness, we now give a couple examples of proximal mappings.

1. if $\rho(x) = x^2/2$, $\operatorname{prox}_c(\rho)[x] = \frac{x}{1+c}$.
2. if $\rho(x) = |x|$, $\operatorname{prox}_c(\rho)[x] = sgn(x)(|x| - c)_+$, i.e the "soft-thresholding" function.

B On convex Lipschitz functions of random variables

In this section, we provide a brief reminder concerning convex Lipschitz functions of random variables.

Lemma B-1. *Suppose that $\{X_i\}_{i=1}^n \in \mathbb{R}^p$ satisfy the following concentration property: $\exists C_n, c_n$ such that for any G_i, a convex, 1-Lipschitz (with respect to Euclidean norm) function of X_i,*

$$P(|G_i(X_i) - m_i| \geq t) \leq C_n \exp(-c_n t^2) ,$$

where m_i is deterministic.

Let us now fix $\{F_i\}_{i=1}^n$, n functions which are convex and 1-Lipschitz in X_i. Then if $\mathcal{F}_n = \sup_i |F_i(X_i) - m_i|$, we have, even when the X_i's are dependent:

1. *if $u_n = \sqrt{\log(n)/c_n}$, $\mathbf{E}(\mathcal{F}_n) \leq u_n + C_n/(2\sqrt{c_n}\sqrt{\log n}) = \frac{\sqrt{\log n}}{\sqrt{c_n}}(1 + C_n/(2\log n))$. Similar bounds hold in L_k for any finite given k.*

2. *when $C_n \leq C$, where C is independent of n, there exists K, independent of n such that $\mathcal{F}_n/u_n \leq K$ with overwhelming probability, i.e probability asymptotically smaller than any power of $1/n$.*

3. *m_i can be chosen to be the mean or the median of $F_i(X_i)$.*

In particular,

$$\mathcal{F}_n = \mathrm{O}(\mathit{polyLog}(n)/\sqrt{c_n})$$

in probability and any L_k, k fixed and given.

We note that similar techniques can be used to extend the result to situations where we have $P(|F_i(X_i) - m_i| \geq t) \leq C_n \exp(-c_n t^\alpha)$, with $\alpha \neq 2$. Of course, the order of magnitudes of the bounds then change.

Proof. Item 3 of the previous Lemma follows from Proposition 1.8 in Ledoux (2001) - the impact of such choice is simply to possibly change C_n and c_n by constants (independent of p and n).

Clearly, by a simple union bound,

$$P(\mathcal{F}_n \geq t) \leq 1 \wedge nC_n \exp(-c_n t^2) .$$

Hence, for any $u \geq 0$,

$$\mathbf{E}\left(\mathcal{F}_n^k\right) \leq u^k + \int_u^\infty kt^{k-1} nC_n \exp(-c_n t^2) ,$$

since $\mathbf{E}\left(\mathcal{F}_n^k\right) = \int_0^\infty kt^{k-1} P(\mathcal{F}_n \geq t) dt$. Standard computations show that when $u^2 c_n$ is large, and $k \geq 1$,

$$\int_u^\infty t^{k-1} \exp(-c_n t^2) dt = \mathrm{O}(\frac{u^k}{2c_n u^2} \exp(-c_n u^2)) .$$

So we see that in that case, for a constant K_k that depends only on k,

$$\mathbf{E}\left(\mathcal{F}_n^k\right) \leq u^k (1 + K_k \frac{nC_n}{c_n u^2} \exp(-c_n u^2)) .$$

Taking $u_n = \sqrt{\log n / c_n}$, we see that

$$\mathbf{E}\left(\mathcal{F}_n^k\right) \leq u_n^k (1 + K_k \frac{C_n}{\log n}) .$$

We conclude that when $C_n / \log n$ remains bounded, $\mathbf{E}\left(\mathcal{F}_n^k\right) / u_n^k$ remains bounded. In the case $k = 1$, it is easy to see that $K_k = 1/2$ and we do not require $\sqrt{c_n} u$ to be large for our arguments to go through. This gives the bound announced in the Lemma.

The probabilistic bound comes simply from the fact that, by a simple union bound,

$$P(\mathcal{F}_n \geq t u_n) \leq nC_n \exp(-(\log n) t^2) \leq C_n \exp(-(\log n)(t^2 - 1)) .$$

Hence, when C_n remains bounded in n,

$$P\left(\frac{\mathcal{F}_n}{u_n} \geq K\right) \leq n^{-d} ,$$

for any given d, if K is large enough. If we allow K to grow like a power of $\log n$, we also see that the right hand side above can be made even smaller.

□

We recall that we denote by $X_{(i)} = \{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\}$. If I is a subset of $\{1, \ldots, n\}$ of size $n - 1$, we call X_I the collection of the corresponding X_i random variables. We call X_{I^c} the remaining random variable.

Lemma B-2. *Suppose X_i's are independent and satisfy the concentration inequalities as above. Consider the situation where $F_{I_k}(\cdot)$ is a convex Lipschitz function of 1 variable; $F_{I_k}(\xi)$ depends on X through X_{I_k} only and we call $\mathcal{L}_{I_k}$ the Lipschitz constant of $F_{I_k}(\cdot)$ (at X_{I_k} given). $\mathcal{L}_{I_k}$ is assumed to be random, since X_{I_k} is. Call $m_{F_{I_k}} = m_{F_{I_k}(X_{I_k^c})}|X_{I_k}$, m being the mean or the median. As before, call $\mathcal{F}_n = \sup_{j=1,\ldots,n} |F_{I_j}(X_{I_j^c}) - m_{F_{I_j}}|$ Then $\mathcal{F}_n = \mathrm{O}(\sqrt{\log n / c_n} \sup_{1\leq j\leq n} \mathcal{L}_{I_j})$ in probability and in L_k, i.e there exists $K > 0$, independent of n, such that*

$$\mathbf{E}\left({\mathcal{F}_n}^k\right) \leq K(\sqrt{\log n / c_n})^k \sqrt{\mathbf{E}\left(\sup_{1\leq j\leq n} \mathcal{L}^{2k}_{I_j}\right)} \, .$$

Hence, $\mathcal{F}_n$ is polyLog$(n)/c_n^{1/2} \sup_{1\leq j\leq n} \mathcal{L}_{I_j}$ in L_k (provided, of course that $\sqrt{\mathbf{E}\left(\sup_{1\leq j\leq n} \mathcal{L}^{2k}_{I_j}\right)}$ is finite)

Remark : the previous lemma also applies when replacing $\mathcal{F}_n = \sup_{j=1,\ldots,n} |F_{I_j}(X_{I_j^c}) - m_{F_{I_j}}|$ by $\mathcal{F}_n^j = \sup_{k=1,\ldots,n} |F_{I_j,k}(X_{I_j^c}) - m_{F_{I_j,k}}|$, i.e considering n (random) functions of the same random variable $X_{I_j^c}$, as the proof makes clear. Here the random functions $F_{I_j,k}(\xi)$ depend only on X_{I_j}.

Proof. We call $\mathcal{L} = \sup_j \mathcal{L}_{I_j}$. By Holder's inequality, we have

$$\mathbf{E}\left({\mathcal{F}_n}^k\right) = \mathbf{E}\left(({\mathcal{F}_n}^k/\mathcal{L}^k)\mathcal{L}^k\right) \leq \sqrt{\mathbf{E}\left({\mathcal{F}_n}^{2k}/\mathcal{L}^{2k}\right)}\sqrt{\mathbf{E}\left(\mathcal{L}^{2k}\right)} \, .$$

Let us call $\widetilde{\mathcal{F}}_n = \mathcal{F}_n/\mathcal{L}$. Using the same idea as in the proof of the previous Lemma,

$$\begin{aligned}\mathbf{E}\left(\widetilde{\mathcal{F}}^k_n\right) &\leq u^k + \sum_{j=1}^n \int_u^\infty kx^{k-1} P(|F_{I_j}(X_{I_j^c}) - m_{F_{I_j}}| \geq \mathcal{L}x)dx \, , \\ &\leq u^k + \sum_{j=1}^n \int_u^\infty kx^{k-1} P(|F_{I_j}(X_{I_j^c}) - m_{F_{I_j}}| \geq L_{I_j}x)dx \, , \\ &= u^k + \sum_{j=1}^n \int_u^\infty kx^{k-1} \mathbf{E}\left(P(|F_{I_j}(X_{I_j^c}) - m_{F_{I_j}}| \geq L_{I_j}x|X_{I_j})\right) dx \, .\end{aligned}$$

Now our assumptions guarantee that

$$P(|F_{I_k}(X_{I_k^c}) - m_{F_{I_k}}| \geq L_{I_k}x|X_{I_k}) \leq C_n \exp(-c_n x^2) \, ,$$

since F_{I_k}/L_{I_k} is 1-Lipschitz (and X_{I_k} is independent of $X_{I_k^c}$). We conclude that

$$\mathbf{E}\left(\widetilde{\mathcal{F}}^k_n\right) \leq u^k + nC_n \int_u^\infty kx^{k-1} \exp(-c_n x^2) \, .$$

This is exactly the same situation as we had before and the conclusion follows. □

Lemma B-3. *Suppose the assumptions of the previous Lemma are satisfied. Consider $Q_{I_j} = \frac{1}{n} X'_{I_j^c} M_{I_j} X_{I_j^c}$, where M_{I_j} is a random positive-semidefinite matrix depending only on X_{I_j} whose largest eigenvalue is λ_{max,I_j}. Assume that $\mathbf{E}(X_i) = 0$, $\mathrm{cov}(X_i) = \mathrm{Id}_p$ and $nc_n \to \infty$. Then, we have in L_k,*

$$\sup_{1\leq j\leq n} \left| Q_{I_j} - \frac{1}{n} trace\left(M_{I_j}\right)\right| = \mathrm{O}_{L_k}(\frac{polyLog(n)}{\sqrt{nc_n}} \sup_{1\leq j\leq n} \lambda_{max,I_j}) \, .$$

The same bound holds when considering a single Q_{I_j} without the polyLog(n) term.

Proof. Lemma B-2 applies to $\sqrt{Q_{I_j}}$ and $\sup_{1\leq j\leq n} |\sqrt{Q_{I_j}} - m_{\sqrt{Q_{I_j}}}|$. The corresponding Lipschitz constant if of course $\sqrt{\lambda_{max,I_j}/n}$. Indeed, for v a vector and M a positive definite matrix, $v \to \sqrt{v'Mv} = \|M^{1/2}v\|$ is convex and $\sqrt{|||M|||_2}$-Lipschitz.

So all we need to do is show that we can go from this control to the control of $\sup_{1\leq j\leq n} |Q_{I_j} - \frac{1}{n}\text{trace}\left(M_{I_j}\right)|$.

Of course,

$$|Q_{I_j} - \frac{1}{n}\text{trace}\left(M_{I_j}\right)| \leq |Q_{I_j} - m^2_{\sqrt{Q_{I_j}}}| + |m^2_{\sqrt{Q_{I_j}}} - \frac{1}{n}\text{trace}\left(M_{I_j}\right)| \ .$$

The idea from there is simply to use the fact that for a and b non-negative, $(a+b)^k \leq 2^{k-1}(a^k+b^k)$. Using Proposition 1.9 in Ledoux (2001) and specifically the variance bound there, we know that, conditional on X_{I_j},

$$|m^2_{\sqrt{Q_{I_j}}} - \frac{1}{n}\text{trace}\left(M_{I_j}\right)| \leq \frac{C_n}{nc_n}\lambda_{max}(M_{I_j}) \ .$$

Here, we have used the fact

$$\mathbf{E}\left(\left(\sqrt{Q_{I_j}}\right)^2 |X_{I_j}\right) = \frac{1}{n}\text{trace}\left(M_{I_j}\right) \ .$$

On the other hand,

$$|Q_{I_j} - m^2_{\sqrt{Q_{I_j}}}| = \left|\sqrt{Q_{I_j}} - m_{\sqrt{Q_{I_j}}}\right| \left|\sqrt{Q_{I_j}} + m_{\sqrt{Q_{I_j}}}\right| \leq \left|\sqrt{Q_{I_j}} - m_{\sqrt{Q_{I_j}}}\right|^2 + 2\left|\sqrt{Q_{I_j}} - m_{\sqrt{Q_{I_j}}}\right| m_{\sqrt{Q_{I_j}}} \ ,$$

since $m_{\sqrt{Q_{I_j}}} \geq 0$.

Therefore,

$$\sup_{1\leq j\leq n} |Q_{I_j} - m^2_{\sqrt{Q_{I_j}}}| \leq \sup_{1\leq j\leq n} \left|\sqrt{Q_{I_j}} - m_{\sqrt{Q_{I_j}}}\right|^2 + 2\left[\sup_{1\leq j\leq n} \left|\sqrt{Q_{I_j}} - m_{\sqrt{Q_{I_j}}}\right|\right] \left[\sup_{1\leq j\leq n} m_{\sqrt{Q_{I_j}}}\right]$$

Lemma B-2 gives us control of $\sup_{1\leq j\leq n} \left|\sqrt{Q_{I_j}} - m_{\sqrt{Q_{I_j}}}\right|$ in L_{2k} and therefore control of

$$\sup_{1\leq j\leq n} \left|\sqrt{Q_{I_j}} - m_{\sqrt{Q_{I_j}}}\right|^2$$

in L_{2k} with a bound of the form $\frac{\text{polyLog}(n)}{(nc_n)} \sup_{1\leq j\leq n} \lambda_{max}(M_{I_j})$.

The result will therefore be shown provided we control $\left[\sup_{1\leq j\leq n} \left|\sqrt{Q_{I_j}} - m_{\sqrt{Q_{I_j}}}\right|\right] \left[\sup_{1\leq j\leq n} m_{\sqrt{Q_{I_j}}}\right]$. By using Holder's inequality and our control of $\left[\sup_{1\leq j\leq n} \left|\sqrt{Q_{I_j}} - m_{\sqrt{Q_{I_j}}}\right|\right]$ in L_{2k}, it is clear that the only issue remaining is control of $\left[\sup_{1\leq j\leq n} m_{\sqrt{Q_{I_j}}}\right]$ in L_{2k}.

Since $m_{\sqrt{Q_{I_j}}} = \mathbf{E}_{X_{I_j^c}}\left(\sqrt{X'_{I_j^c}M_{I_j}X_{I_j^c}/n}\right) \leq \sqrt{\mathbf{E}_{X_{I_j}}\left(X'_{I_j^c}M_{I_j}X_{I_j^c}/n\right)} = \sqrt{\text{trace}\left(M_{I_j}\right)/n}$, since $\text{cov}\left(X_{I_j}\right) = \text{Id}_p$, we see that

$$\left[\sup_{1\leq j\leq n} m_{\sqrt{Q_{I_j}}}\right] \leq \sqrt{p/n} \sup_{1\leq j\leq n} \sqrt{\lambda_{\max,I_j}} \ .$$

Therefore,

$$\left[\sup_{1\leq j\leq n} \left|\sqrt{Q_{I_j}} - m_{\sqrt{Q_{I_j}}}\right|\right] \left[\sup_{1\leq j\leq n} m_{\sqrt{Q_{I_j}}}\right] \leq K\frac{\text{polyLog}(n)\sqrt{p/n}}{\sqrt{nc_n}} \sup_{1\leq j\leq n} \lambda_{\max,I_j} \text{ in } L_k \ ,$$

provided all the random quantities we work with have $2k$ moments.

The conclusions of the Lemma follow by recalling our assumption that p/n remains bounded and using the fact that $1/c_n \geq K/\sqrt{c_n}$ in the situations we are considering, i.e c_n bounded but possibly going to zero. □

On the spectral norm of covariance matrices

In this subsection, we show that under our initial concentration assumptions, we can control $|||\widehat{\Sigma}|||_2$. These results are very likely known but we did not find a reference covering precisely the same question we consider. The proof is a simple adaption of the well-known ϵ-net argument explained e.g in Talagrand (2003), Appendix A.4.

Lemma B-4. *Suppose X_i's are independent random vectors in $\mathbb{R}^p$, satisfying our concentration assumptions in **O4**, and having mean 0 and covariance Id_p. Let $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i'$. Then,*

$$|||\widehat{\Sigma}|||_2 = \mathrm{O}_P(\mathsf{c}_n^{-1}) .$$

The results hold also in L_k.

Proof. We study the largest singular value, σ_1 of the matrix $X/\sqrt{n}$, where the i-th row of X is X_i. Of course,

$$\sigma_1(X/\sqrt{n}) = \sup_{u,v,\|u\|=1,\|v\|=1} \frac{1}{\sqrt{n}} u'Xv .$$

Note that

$$u'Xv = \sum_{i=1}^n u_i(X_i'v) .$$

Consider first the case where $\mathsf{c}_n = 1$. Under our assumptions, $X_i'v$ are independent subGaussian random vectors, with mean 0. Note that $\mathrm{var}\,(X_i'v) = 1$ if $\mathrm{cov}\,(X_i) = 1$ and $\|v\| = 1$. Computing the moment generating function of $u'Xv$, we see that this random variable is itself subGaussian and has variance 1. Therefore, we have for all t, and constants c_1 and c_2,

$$P(|u'Xv| > t) \leq c_1 \exp(-c_2 t^2) .$$

The ϵ-net argument given in the proof of Lemma A.4.1 in Talagrand (2003) then can be applied and the conclusions of that Lemma reached. (A slight adaption is needed to handle the fact that $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^p$ but it is completely trivial and omitted). The fact that the results hold in L_k is a simple consequence of the proof.

In the case where $\mathsf{c}_n \neq 1$, we just need to note that the moment generating function of $u'Xv$ is smaller than that of a Gaussian random variable with variance $1/\mathsf{c}_n$. The result follows immediately. □

C Miscellaneous results

C-1 An analytic remark

One of our assumptions concerns the existence and uniqueness of a solution of the equation $F(x) = 0$, where

$$F(x) = \frac{p}{n} - \tau x - 1 + \mathbf{E}\left((\mathrm{prox}_x(\rho))'(W)\right)$$

where W is a random variable and $(\mathrm{prox}_x(\rho))'(t) = \frac{\partial}{\partial t}\mathrm{prox}_x(\rho)(t) = \frac{1}{1+x\psi'(\mathrm{prox}_x(\rho)(t))}$.

We now show that under mild conditions on W this equation has a unique solution. This guarantees that our assumptions are not terribly strong and in particular apply to problems of interest to statisticians.

Lemma C-1. *Suppose that W has a smooth density f with sign$(f'(x)) = -$sign(x). Suppose further that $\lim_{|t|\to\infty} tf(t) = 0$ and that sign$(\psi(x))$ = sign(x). Then, if*

$$F(x) = \frac{p}{n} - \tau x - 1 + \mathbf{E}\left((prox_x(\rho))'(W)\right) ,$$

the equation $F(x) = 0$ has a unique solution.

Proof. We call

$$G(x) \triangleq \mathbf{E}\left((\text{prox}_x(\rho))'(W)\right) .$$

Of course,

$$\mathbf{E}\left((\text{prox}_x(\rho))'(W)\right) = \int (\text{prox}_x(\rho))'(t) f(t) dt.$$

Using contractivity of the proximal mapping (see Moreau (1965)) we see that $\lim_{|t|\to\infty} \text{prox}_x(\rho)(t) f(t) = 0$ under our assumptions.

Integrating the previous equation by parts, we see that

$$\mathbf{E}\left((\text{prox}_x(\rho))'(W)\right) = -\int (\text{prox}_x(\rho))(t) f'(t) dt .$$

To compute $G'(x)$, we differentiate under the integral sign (under our assumptions the conditions of Theorem 9.1 in Durrett (1996) are satisfied) to get

$$G'(x) = \int \frac{\psi(\text{prox}_x(\rho)(t)) f'(t)}{1 + x\psi'(\text{prox}_x(\rho)(t))} dt .$$

Under our assumptions, $\text{sign}(\psi(\text{prox}_x(\rho)(t))) = \text{sign}(t)$ and $\text{sign}(f'(t)) = -\text{sign}(t)$, so that

$$\forall t \neq 0, \text{sign}(\psi(\text{prox}_x(\rho)(t)) f'(t)) = -1 .$$

Since the denominator of the function we integrate is positive, we conclude that

$$G'(x) \leq 0 .$$

Since $F'(x) = -\tau + G'(x)$, we see that $F'(x) < 0$. Therefore F is a decreasing function on $\mathbb{R}_+$. Of course, $F(0) = p/n$ and $\lim_{x\to\infty} F(x) = -\infty$. So we conclude that the equation $F(x) = 0$ has a unique root. □

Remark: the conditions on the density of W are satisfied in many situations. For instance if $W = \epsilon + rZ$, where ϵ is symmetric about 0 and log-concave, Z is $\mathcal{N}(0,1)$ and $r > 0$, it is clear that the density of W satisfies the conditions of our lemma. Similar results hold under weaker assumptions on ϵ of course but since the paper is already a bit long, we do not dwell on these issues which are well-known in the theory of log-concave functions (see e.g Karlin (1968), Prékopa (1973) and Ibragimov (1956)).

C-2 A linear algebraic remark

We have the following lemma.

Lemma C-2. *Suppose the $p \times p$ matrix A is positive semi-definite and*

$$A = \begin{pmatrix} \Gamma & v \\ v' & a \end{pmatrix} .$$

Here $a \in \mathbb{R}$. Let τ be a strictly positive real. Call $\Gamma_\tau = \Gamma + \tau \text{Id}_{p-1}$. Then we have

$$\textit{trace}\left((A + \tau \text{Id}_p)^{-1}\right) = \textit{trace}\left(\Gamma_\tau^{-1}\right) + \frac{1 + v'\Gamma_\tau^{-2} v}{a + \tau - v'\Gamma_\tau^{-1} v} .$$

In particular,

$$\left|\textit{trace}\left((A + \tau)^{-1}\right) - \textit{trace}\left(\Gamma_\tau^{-1}\right)\right| \leq \frac{1 + a/\tau}{\tau} .$$

Proof. The first equation is simply an application of the block inversion formula for matrices (see Horn and Johnson (1990), p.18) and the Sherman-Morrison-Woodbury formula (Horn and Johnson (1990), p.19). Suppose temporarily that A is positive definite. Then the Schur complement formula guarantees that $a \geq v'\Gamma^{-1} v \geq v'\Gamma_\tau^{-1} v$. The fact that $a \geq v'\Gamma_\tau^{-1} v$ in general is obtained by a continuity argument (change A to $A_\epsilon = A + \epsilon \text{Id}_p$ and let ϵ tend to 0). This implies that

$$\frac{1}{a + \tau - v'\Gamma_\tau^{-1} v} \leq \frac{1}{\tau} .$$

Since $v'\Gamma_\tau^{-2} v \leq \frac{1}{\tau} v'\Gamma_\tau^{-1} v \leq a/\tau$, we get the second equation. □

D Sketch of proof and discussion of statistical issues

D-1 Sketch of proof and explanations

We give some explanations about our proof in case it is helpful for the reader. For simplicity of notations, we consider the unpenalized problem ($\tau = 0$) in the case $\beta_0 = 0$ and hence consider $\widehat{\beta}$ defined as

$$\widehat{\beta} : \sum_{i=1}^{n} -X_i \psi(\epsilon_i - X_i' \widehat{\beta}) = 0_p \; . \tag{D-1}$$

We recall that the probabilistic heuristics developed in El Karoui et al. (2013) suggested the following:

$$\text{let } \tilde{r}_{j,(i)} = \epsilon_j - X_j' \widehat{\beta}_{(i)} \; , \forall j, 1 \leq j \leq n \; .$$

Call $S_i = \frac{1}{n} \sum_{j \neq i} \psi'(\tilde{r}_{j,(i)}) X_j X_j'$. Then, first order perturbation arguments suggest that we "should have"

$$\widehat{\beta} - \widehat{\beta}_{(i)} \simeq \frac{1}{n} S_i^{-1} X_i \psi(\epsilon_i - X_i' \widehat{\beta}) \; . \tag{D-2}$$

Measure concentration for quadratic forms in X_i's would then imply that $X_i'(\widehat{\beta} - \widehat{\beta}_{(i)}) \simeq c_i \psi(\epsilon_i - X_i' \widehat{\beta})$, where $c_i \simeq \frac{1}{n} \text{trace}\left(S_i^{-1}\right)$. Furthermore it is plausible that $c_i \simeq c$, where c does not depend on i. This would yield

$$\tilde{r}_{i,(i)} - R_i \simeq c \psi(R_i) \; ,$$

and hence for i-th residual,

$$R_i = \epsilon_i - X_i' \widehat{\beta} \simeq \text{prox}(c\rho)(\tilde{r}_{i,(i)}) \; .$$

Section 3 makes all this precise, though it does not address the question of whether c is asymptotically deterministic (i.e whether it can be approximated by a deterministic quantity). Two main issues arise: a key one is, of course, how to verify an approximation like Equation (D-2). The key tools for this task are developed in Section 2. Equation (D-2) suggests an approximation of $\widehat{\beta}$ by a function of $\widehat{\beta}_{(i)}$. However, it turns out that this approximation is too coarse to carry out rigorously all the steps needed in the proof; other issues arise when trying to approximate c_i by trace $\left(S_i^{-1}\right)$. So we came up with much more precise approximations than the ones we just discussed that allow us to rigorously prove all the results we need to carry the proof out. Technically, working with $\tau > 0$ simplifies a number of arguments. This is also quite natural statistically and analytically, since it makes the problem strongly convex. In particular, this leads to the proof of the fact var $\left(\|\widehat{\beta}\|_2^2\right) \to 0$ (through the Efron-Stein inequality, which is nothing else than a version of Burkholder's inequality) and hence $\|\widehat{\beta}\|_2^2$ can be approximated by a deterministic quantity. This part of the proof does not require the X_i's to have independent entries (this partly motivated our decision to work under assumptions we chose for X_i's, namely **O4**).

The second part of the proof, Section 4 can be understood in part in the following light. Call $r_{i,[p]}$ the residuals based on first $p-1$ predictors, called $\{V_i\}_{i=1}^n$. Call

$$\mathfrak{S}_p = \sum_i \psi'(r_{i,[p]}) V_i V_i' \; , \text{ and } u_p = \sum_i \psi'(r_{i,[p]}) V_i X_i(p) \; ,$$

Intuitively it is reasonable to think that in many situations $r_{i,[p]} \simeq R_i$, when p is large. Then, first order approximations to Equation (D-1) suggest that

$$\widehat{\beta}_p \simeq \frac{\sum X_i(p) \psi(r_{i,[p]})}{\sum X_i^2(p) \psi'(r_{i,[p]}) - u_p' \mathfrak{S}_p^{-1} u_p} \; .$$

After somewhat fine manipulations, the denominator appears to be approximately equal to $\frac{p-1}{c}$, where c is defined as above.

After further work, this suggests the second equation of the system - ignoring for a moment the issue of whether $\widehat{\beta}_{(i)}' X_i$ is approximately normal. Section 4 justifies all the approximations we just discussed.

Furthermore, parts of Section 4 lay the ground work for proving the first equation of the system with Lemma 4.2 - which is closest technically to techniques used to prove Marchenko-Pastur style results in random matrix theory by the method of Stieltjes transforms.

Section 5 proves the first equation of the system and addresses two further questions: asymptotic normality of $\widehat{\beta}'_{(i)}X_i$ as well as asymptotically deterministic character of c.

One potential difficulty with this proof is that many quantities of interest cannot be studied independently of one another - the system characterizing the quantity of main interest to us illustrates this point clearly. We hope this short discussion sheds some light on the proof strategy.

The random matrix point of view is of course essential to both understanding the problem and carrying out of the proof. While we borrowed ideas and tools from this area of probability, our work is not a straight application of existing results: the main idea of our work is that even though Equation (D-1) does not look like a random matrix question, it can be cast as one (and therefore allows us to treat the problem of interest in great generality when it comes to X_i's, for instance). Furthermore the random matrices appearing in our work are non-standard: some of them are weighted covariance matrices, i.e of the form $\frac{1}{n}\sum_{i=1}^n w_i X_i X_i'$. This makes them look somewhat classical except that w_i's here are basically $\psi'(R_i)$, and hence they depend on X_i's and furthermore it is not clear at the beginning of the proof how the empirical distribution of those w_i's behaves. Understanding this latter question is one of the main problems here. Hence, our work is far from being a straight application of standard random matrix results; but it shows how versatile tools developed in this area of probability are and how they can be brought to bear on a new class of problems.

D-2 Statistical issues

D-2.1 Reminder and setup explanation

A classic result (Huber (1973)) in low dimension (i.e p fixed, $n \to \infty$) is the fact that, under mild technical conditions and when ϵ_i's are i.i.d, the performance of methods similar to the ones we studied here when $\tau = 0$ is governed by the quantity

$$\mathsf{r}(\epsilon, \psi) = \frac{\mathbf{E}\left(\psi^2(\epsilon)\right)}{[\mathbf{E}\left(\psi'(\epsilon)\right)]^2}\ , \epsilon \text{ having the same distribution as } \epsilon_i's\ , \psi = \rho'\ .$$

Further analysis (see e.g Huber and Ronchetti (2009)) shows that if f_ϵ is the density of ϵ, the previous quantity is minimized for $\rho = -\log(f_\epsilon)$, once again under regularity conditions. Note that almost by definition this function ρ is convex only if ϵ has a log-concave density. Since our analysis was partly motivated by similar considerations, it is therefore important that we allow ψ's and ρ's that grow relatively fast at infinity so that $\rho = -\log(f_\epsilon)$ be part of the functions covered by our theorem for many log-concave densities. While some statisticians will not think of this setup as standard "robust" statistics, it is nonetheless central to the classical (i.e low dimensional) understanding of optimality in statistics. This latter topic was part of our statistical motivation here.

This is partly why working with log-concave errors and hence dealing with ρ's that are allowed to grow polynomially at infinity is, on top of its probabilistic interest, part of the setup we chose.

We further note that the so-called Huber function (see Huber and Ronchetti (2009), p. 84) which often comes to mind when discussing robust statistics arose from an analysis of the quantity $\mathsf{r}(\epsilon, \psi)$ described above, in the so-called Gaussian contamination problem (see Huber and Ronchetti (2009), p. 83). This is not the type of questions we are concerned with in this paper - we are mostly concerned with the probabilistic analysis of $\widehat{\beta}$ defined in Equation (2). Even in the case of Huber's work, the corresponding probabilistic analysis was a necessary first step towards the statistical question of understanding the Gaussian contamination problem. This "first step" is what we undertook in the current paper, and it is much more difficult than in the low-dimensional setting. In the type of optimality questions in statistics that were a motivation for us for first considering this problem, where f_ϵ is assumed to be known, the Huber function is *not* a natural choice, since it has no optimality property in that setting in general. Hence, the fact that in the setup considered here, our theorem does not cover the Huber function is *not* much of an issue because, among other things : a) of the Huber function lack of naturalness in high-dimension in light of the results of Bean et al. (2013) b) it seems possible that with quite a bit of further approximation

theoretic work, as discussed in the main text and below, we might be able to get results for the Huber function by viewing it as a limit of smooth (and strongly convex) functions, at least for the kind of ϵ_i's we work with here (e.g having a log-concave density, or, in any case having many moments).
To summarize, our aim with this paper was the development of a probabilistic understanding of regression M-estimates when the errors ϵ_i's are for instance log-concave, since this type of errors where the ones used in the analytic results of the paper Bean et al. (2013) and give rise to convex optimal loss functions in the classical low-dimensional setting. Understanding this type of questions is a necessary first step to develop a theory of robust statistics (as seen from the work of Huber) - but the development of that theory in high-dimension is not the aim of the current paper, whose aims are once again probabilistic. This explains our choice of probabilistic setup.

D-2.2 About $\|\widehat{\beta}-\beta_0\|$

The fact that $\|\widehat{\beta}-\beta_0\|$ in Lemma (A) (i.e in the setup $\tau=0$) does not converge to zero has been a source of confusion for some statisticians working on sparse modeling (i.e assuming that β_0 has finitely many non-zero coordinates, even as $p\to\infty$). Of course, this very fact is what renders the problem interesting probabilistically (and in the author's opinion also statistically: these estimators perform very well as explained in the paper when projected on most directions, but there are data-dependent directions where they encounter "problems"; so they are good for the most important statistical tasks, but not uniformly good.).

One line of argument is that because $\|\widehat{\beta}-\beta_0\|$ does not go to zero, $\widehat{\beta}$ is of no interest statistically (!). However, this objection is simply overcome by the following points : **1)** consider for the sake of simplicity the case where $X_{i,j}$ are i.i.d $\mathcal{N}(0,1)$. As shown in El Karoui et al. (2011), we then have

$$\widehat{\beta}-\beta_0 \overset{\mathcal{L}}{=} \|\widehat{\beta}-\beta_0\|_2 u\;,$$

where u is uniform on the unit sphere in $\mathbb{R}^p$ and independent of $\|\widehat{\beta}-\beta_0\|_2$. Hence, using the fact that $u\overset{\mathcal{L}}{=}Z/\|Z\|$, where $Z\sim\mathcal{N}(0,\mathrm{Id}_p)$, we see that

$$\sup_{1\leq k\leq p}\sqrt{p}|\widehat{\beta}(k)-\beta_0(k)| = \mathrm{O}_P(\sqrt{\log(p)})\;,$$

since $\max_{1\leq k\leq p}|Z_k| = \mathrm{O}_P(\sqrt{\log(p)})$ where Z_k's are i.i.d $\mathcal{N}(0,1)$. In other words, for each i, $\widehat{\beta}(i)$ is $\sqrt{p}$-consistent for $\beta_0(i)$ and the coordinates of $\widehat{\beta}$ contains a lot of information about those of β_0. In the case where $X_{i,j}$ are simply i.i.d and satisfy our assumptions, similar arguments can be made in the case of strongly convex ρ, based on Theorem (C) via permutation-symmetry arguments and approximations as $\tau\to0$ but they require a bit more care. We leave them to the interested reader, since that has very little to do with the main efforts of this paper.
We also note that the stochastic representation above makes it clear that $\|\widehat{\beta}-\beta\|_{2+\epsilon}\to0$ for any $\epsilon>0$. And we just explained, other norms may be more important for certain statistical tasks than the Euclidean norm.
2) Furthermore, if β_0 is *not sparse*, and for instance all its coordinates of size roughly $p^{-1/2}$ (concretely $\beta_0(i)=u_i/\sqrt{p}$ with $\sum_{i=1}^p u_i^2=K$, $|u_i|>\eta$ for some η for all i's, K fixed independently of p, for instance), sparse methods (returning an estimate s of β_0 with $\mathsf{s}(i)=0$ for all but finitely many i's even as $p\to\infty$) would be such that $\|\mathsf{s}-\beta_0\|_2$ does not converge to zero. Hence, sparse estimators fall themselves under the scope of the criticism levied by their advocates. Here we *do not require sparsity* of β_0- for instance in the context of Lemma (A) - and still understand the probabilistic behavior of the quantities of interest to us. This is what allowed us (as explained in e.g Bean et al. (2013)) to then propose new statistical methods, using non-standard functions ρ. Without a probabilistic understanding of the problem, this would not have been possible.
3) In the context of Lemma (A), and using the previous remarks for $X_{i,j}$ e.g i.i.d $\mathcal{N}(0,1)$, if β_0 is sparse, i.e it is supported on finitely many coordinates with relatively large entries on those coordinates, we can threshold $\widehat{\beta}$ at level $C=(\log(p))^{1/2+\eta}/\sqrt{p}$ for some $\eta>0$, i.e apply the function f_C such that $f_C(x)=x1_{|x|\geq C}$ to

each coordinate of $\widehat{\beta}$, to create a new estimator $\mathsf{T}_{\widehat{\beta}}$ such that

$$\|\mathsf{T}_{\widehat{\beta}} - \beta_0\|_2 \to 0 \ .$$

This is a simple application of the stochastic representation result we mentioned a few lines above. Understanding $\|\widehat{\beta} - \beta_0\|_2$, as we do in this paper, is a way to pick better thresholds C than the conservative one just described. Furthermore, the methods we just described may be numerically more efficient than many sparse methods than rely on solving other optimization problems than the ones we are concerned with here.

D-2.3 About prediction error

In a number of machine learning/statistics setting, a quantity of interest is the prediction error, i.e the error made when we use $\widehat{\beta}$ as a proxy for β_0 on a new data: more concretely, one considers

$$\text{PredError} = Y_{new} - X_{new}'\widehat{\beta} \ .$$

Here (X_{new}, Y_{new}) are not part of the initial dataset $\{(X_i, Y_i)\}_{i=1}^n$. In general, the practitioner gets to observe only X_{new} and wants to predict the corresponding Y_{new} which is assumed to not be known. If the linear model holds, i.e $Y_{new} = X_{new}'\beta_0 + \epsilon_{new}$, we have $\text{PredError} = \epsilon_{new} - X_{new}'(\widehat{\beta} - \beta_0)$, so that if $\text{var}\,(\epsilon_{new}) = \sigma^2_{\epsilon_{new}}$ and X_{new} has mean 0 and covariance Id_p, conditional on the observed data $\{(X_i, Y_i)\}_{i=1}^n$,

$$\mathbf{E}_{X_{new},\epsilon_{new}}(\text{PredError}) = \sigma^2_{\epsilon_{new}} + \|\widehat{\beta} - \beta_0\|_2^2 \ .$$

Hence, understanding the probabilistic properties of $\|\widehat{\beta} - \beta_0\|_2^2$ is key in assessing how accurate our method is at predicting Y_{new}. This is another motivation for our study, beside its probabilistic interest.

D-3 Non-smooth ρ and ψ, strong convexity question

It seems that based on the results of the current paper, in particular Theorem 1.1, one could handle certain non-smooth functions of potential interest through some further approximation-theoretic work - essentially showing that approximating the non-smooth function ρ by a family of smooth functions does not change very much $\|\widehat{\beta}\|$ as a function of ρ.

For instance the Huber function is differentiable, but not twice differentiable at exactly two points. For the sake of concreteness, let us take ψ_H to be such that $\psi_H(x) = x 1_{-1\leq x \leq 1} + \text{sign}(x) 1_{|x|>1}$. At these points, i.e -1 and 1, we could smooth ψ_H to create a family of functions $\psi_{\text{smooth},\eta}$, $\eta > 0$ such that $\psi_{\text{smooth},\eta}$ is close to ψ_H (arbitrarily so as $\eta \to 0$) and our results apply to $\psi_{\text{smooth},\eta}$. A natural example would be to use $\psi_{\text{smooth},\eta}$ such that $\psi'_{\text{smooth},\eta}(x) = 1_{|x|\leq 1-\eta} + \frac{1-|x|}{\eta} 1_{1-\eta<|x|\leq 1}$. (Note that $\psi'_{\text{smooth},\eta}$ is Lipschitz with Lipschitz constant $1/\eta$.) Approximation theoretic work would then be required to show that we can transfer our results concerning $\widehat{\beta}_{\psi_{\text{smooth},\eta}}$ to $\widehat{\beta}_{\psi}$. This is the strategy we used in Section 6.1 to go from the ℓ_2-regularized to the un-regularized ($\tau = 0$) problem. Because these arguments have not much to do with probability theory and because they are, as we explained above, really of secondary importance for us, we leave them for future work. (We note that our current results apply for any $\eta > 0$, such as $\eta = 10^{-9}$.)

We believe that similar ideas should allow us to extend Lemma (A) to certain functions that are not strongly convex, by approximating them by a family of strongly convex functions, such as the ones described in the discussion on p. 10: we could try approximating ρ by $\rho_\eta = \rho + \eta p_2$, where $p_2 = x^2/2$ - at least when working with errors ϵ_i's that have two moments. Another approach to handle the situation where ρ is not strongly convex could be to refine the second part of Proposition 2.1 to functions that are strongly convex but only on a subset of $\mathbb{R}$ where, in the notation of Proposition 2.1 "most" of the intervals $(\epsilon_i - X_i'\beta_1, \epsilon_i - X_i'\beta_2)$ fall, for "well-chosen" β_1 and β_2 . This would naturally entail to refine quite considerably a number of results of Sections 3 and 4.

D-4 Examples of distributions satisfying Assumptions O4 and P1

- The fact that assumption **O4** is satisfied when X_i has independent entries that are supported on an interval of width $\sqrt{\mathsf{c}_n}/2$ is a simple application of Corollary 4.10 in Ledoux (2001) - which is itself a consequence of results in Talagrand (1995). Our statement concerning the situation where X_i's has independent entries with strongly log-concave density - following Theorem 1.1 - is a consequence of Theorem 2.7 in Ledoux (2001). The same results justify the fact that **P1** holds for the situation where the entries of the $n \times p$ matrix X are i.i.d with the distributions we just considered.

• For a broader discussion of distributions satisfying Assumption **O4**, we refer the reader to El Karoui (2009), p.. 2386-2387.

• Geometric consequences of these concentration assumptions. Using in **O4**, $G(x) = \|x\|_2/\sqrt{p}$ shows under our assumptions that $\sup_{1\leq i\leq n} |\|X_i\|/\sqrt{p} - m_p| = \mathrm{o}_P(1)$, where m_p is a median of the random variable $\|X_1\|/\sqrt{p}$. Our Lemma B-3 shows furthermore that $\sup_{1\leq i\leq n} |\|X_i\|^2/p - 1| = \mathrm{o}_P(1)$ under the assumptions we work with (simply use $M_{I_j} = \mathrm{Id}_p$ in Lemma B-3 and the proof goes through). This means that the data vector $X_i/\sqrt{p}$'s all have essentially unit norm, and hence are located near the unit sphere in $\mathbb{R}^p$. Similarly, one can establish the fact that the vector X_i is nearly orthogonal to any $\{X_j\}_{j\neq i}$ under these conditions. Hence working under the assumptions like **O4** and **P1** allow us to show what is the gist of the arguments, while also understanding some of the key geometric assumptions that are made about the vectors $\{X_i\}_{i=1}^n$.

References

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis.* Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, third edition.

Anscombe, F. J. (1967). Topics in the investigation of linear relations fitted by the method of least squares. (With discussion.). *J. Roy. Statist. Soc. Ser. B* **29**, 1–52.

Bai, Z. D. (1999). Methodologies in spectral analysis of large-dimensional random matrices, a review. *Statist. Sinica* **9**, 611–677. With comments by G. J. Rodgers and Jack W. Silverstein; and a rejoinder by the author.

Bai, Z. D. and Yin, Y. Q. (1993). Limit of the smallest eigenvalue of a large-dimensional sample covariance matrix. *Ann. Probab.* **21**, 1275–1294.

Baranchik, A. J. (1973). Inadmissibility of maximum likelihood estimators in some multiple regression problems with three or more independent variables. *Ann. Statist.* **1**, 312–321.

Bayati, M. and Montanari, A. (2012). The LASSO risk for Gaussian matrices. *IEEE Trans. Inform. Theory* **58**, 1997–2017. URL http://dx.doi.org/10.1109/TIT.2011.2174612.

Bean, D., Bickel, P. J., El Karoui, N., and Yu, B. (2013). Optimal m-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences* **110**, 14563–14568. URL http://www.pnas.org/content/110/36/14563.abstract.

Beck, A. and Teboulle, M. (2010). *Convex Optimization in Signal Processing and Communications*, chapter Gradient-Based Algorithms with Applications in Signal Recovery Problems, pp. 33–88. Cambridge University Press.

Bhatia, R. (1997). *Matrix analysis*, volume 169 of *Graduate Texts in Mathematics.* Springer-Verlag, New York.

Diaconis, P. and Freedman, D. (1984). Asymptotics of graphical projection pursuit. *Ann. Statist.* **12**, 793–815. URL http://dx.doi.org/10.1214/aos/1176346703.

Donoho, D. and Montanari, A. (2013). High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *arXiv:1310.7320* .

Donoho, D. L., Maleki, A., and Montanari, A. (2009). Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences* **106**, 18914–18919. URL http://www.pnas.org/content/106/45/18914.abstract.

Durrett, R. (1996). *Probability: theory and examples.* Duxbury Press, Belmont, CA, second edition.

Efron, B. and Stein, C. (1981). The jackknife estimate of variance. *Ann. Statist.* **9**, 586–596.

El Karoui, N. (2009). Concentration of measure and spectra of random matrices: Applications to correlation matrices, elliptical distributions and beyond. *The Annals of Applied Probability* **19**, 2362–2405.

El Karoui, N. (2010). High-dimensionality effects in the Markowitz problem and other quadratic programs with linear constraints: risk underestimation. *Ann. Statist.* **38**, 3487–3566. URL http://dx.doi.org/10.1214/10-AOS795.

El Karoui, N., Bean, D., Bickel, P., Lim, C., and Yu, B. (2011). On robust regression with high-dimensional predictors. Technical Report 811, UC, Berkeley, Department of Statistics. Originally submitted as manuscript AoS1111-009. Not under consideration anymore.

El Karoui, N., Bean, D., Bickel, P., Lim, C., and Yu, B. (2012). On robust regression with high-dimensional predictors. *PNAS* .

El Karoui, N., Bean, D., Bickel, P. J., Lim, C., and Yu, B. (2013). On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences* URL http://www.pnas.org/content/early/2013/08/15/1307842110.abstract.

El Karoui, N. and Koesters, H. (2011). Geometric sensitivity of random matrix results: consequences for shrinkage estimators of covariance and related statistical methods. *Submitted to Bernoulli* Available at arXiv:1105.1404 (68 pages).

El Karoui, N. and Purdom, E. (2015). Can we trust the bootstrap in high-dimension? Technical Report 824, UC Berkeley, Department of Statistics. Submitted to JASA.

Halko, N., Martinsson, P. G., and Tropp, J. A. (2011). Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.* **53**, 217–288. URL http://dx.doi.org/10.1137/090771806.

Hall, P., Marron, J. S., and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67**, 427–444.

Hiriart-Urruty, J.-B. and Lemaréchal, C. (2001). *Fundamentals of convex analysis.* Grundlehren Text Editions. Springer-Verlag, Berlin. Abridged version of ıt Convex analysis and minimization algorithms. I [Springer, Berlin, 1993; MR1261420 (95m:90001)] and ıt II [ibid.; MR1295240 (95m:90002)].

Horn, R. A. and Johnson, C. R. (1990). *Matrix analysis.* Cambridge University Press, Cambridge. Corrected reprint of the 1985 original.

Huber, P. J. (1972). The 1972 Wald lecture. Robust statistics: A review. *Ann. Math. Statist.* **43**, 1041–1067.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *Ann. Statist.* **1**, 799–821.

Huber, P. J. and Ronchetti, E. M. (2009). *Robust statistics.* Wiley Series in Probability and Statistics. John Wiley & Sons Inc., Hoboken, NJ, second edition. URL http://dx.doi.org/10.1002/9780470434697.

Ibragimov, I. A. (1956). On the composition of unimodal distributions. *Teor. Veroyatnost. i Primenen.* **1**, 283–288.

Johnstone, I. (2001). On the distribution of the largest eigenvalue in principal component analysis. *Ann. Statist.* **29**, 295–327.

KARLIN, S. (1968). *Total positivity. Vol. I.* Stanford University Press, Stanford, Calif.

LEDOUX, M. (2001). *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs.* American Mathematical Society, Providence, RI.

LEHMANN, E. L. and CASELLA, G. (1998). *Theory of point estimation.* Springer Texts in Statistics. Springer-Verlag, New York, second edition.

MAMMEN, E. (1989). Asymptotics with increasing dimension for robust regression with applications to the bootstrap. *Ann. Statist.* **17**, 382–400. URL `http://dx.doi.org/10.1214/aos/1176347023`.

MARČENKO, V. A. and PASTUR, L. A. (1967). Distribution of eigenvalues in certain sets of random matrices. *Mat. Sb. (N.S.)* **72 (114)**, 507–536.

MOREAU, J.-J. (1965). Proximité et dualité dans un espace hilbertien. *Bull. Soc. Math. France* **93**, 273–299.

POLLARD, D. (1984). *Convergence of stochastic processes.* Springer Series in Statistics. Springer-Verlag, New York.

PORTNOY, S. (1984). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency. *Ann. Statist.* **12**, 1298–1309. URL `http://dx.doi.org/10.1214/aos/1176346793`.

PORTNOY, S. (1985). Asymptotic behavior of M estimators of p regression parameters when p^2/n is large. II. Normal approximation. *Ann. Statist.* **13**, 1403–1417. URL `http://dx.doi.org/10.1214/aos/1176349744`.

PORTNOY, S. (1987). A central limit theorem applicable to robust regression estimators. *J. Multivariate Anal.* **22**, 24–50. URL `http://dx.doi.org/10.1016/0047-259X(87)90073-X`.

PRÉKOPA, A. (1973). On logarithmic concave measures and functions. *Acta Sci. Math. (Szeged)* **34**, 335–343.

RANGAN, S. (2011). Generalized approximate message passing for estimation with random linear mixing. In *IEEE International Symp. On Information Theory (St. Petersburg).*

RELLES, D. (1968). *Robust Regression by Modified Least Squares.* Ph.D. thesis, Yale University.

RUSZCZYŃSKI, A. (2006). *Nonlinear optimization.* Princeton University Press, Princeton, NJ.

SCHIROTZEK, W. (2007). *Nonsmooth analysis.* Universitext. Springer, Berlin. URL `http://dx.doi.org/10.1007/978-3-540-71333-3`.

SHCHERBINA, M. and TIROZZI, B. (2003). Rigorous solution of the Gardner problem. *Comm. Math. Phys.* **234**, 383–422. URL `http://dx.doi.org/10.1007/s00220-002-0783-3`.

SILVERSTEIN, J. W. (1985). The smallest eigenvalue of a large-dimensional Wishart matrix. *Ann. Probab.* **13**, 1364–1368.

SILVERSTEIN, J. W. (1995). Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices. *J. Multivariate Anal.* **55**, 331–339.

STEIN, C. (1960). Multiple regression. In *Contributions to probability and statistics*, pp. 424–443. Stanford Univ. Press, Stanford, Calif.

STROOCK, D. W. (1993). *Probability theory, an analytic view.* Cambridge University Press, Cambridge.

TALAGRAND, M. (1995). Concentration of measure and isoperimetric inequalities in product spaces. *Inst. Hautes Études Sci. Publ. Math.* pp. 73–205.

Talagrand, M. (2003). *Spin glasses: a challenge for mathematicians*, volume 46 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics]*. Springer-Verlag, Berlin. Cavity and mean field models.

van der Vaart, A. W. (1998). *Asymptotic statistics.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

Wachter, K. W. (1978). The strong limits of random matrix spectra for sample matrices of independent elements. *Annals of Probability* **6**, 1–18.

Yin, Y. Q., Bai, Z. D., and Krishnaiah, P. R. (1988). On the limit of the largest eigenvalue of the large-dimensional sample covariance matrix. *Probab. Theory Related Fields* **78**, 509–521.