On robust regression with high-dimensional predictors

Noureddine El Karoui[*], Derek Bean, Peter Bickel[†], Chingway Lim and Bin Yu[‡]

First version: July 13th, 2011
This version: November 3, 2011

Abstract

We consider the problem of understanding the properties of

$$\widehat{\beta} = \text{argmin}_\beta \sum_{i=1}^{n} \rho(y_i - X_i'\beta) \ , y_i = X_i'\beta_0 + \epsilon_i \ ,$$

where X_i is a p-dimensional vector of observed predictors, y_i is a 1-dimensional response and ρ is a given and known convex (loss) function. We are concerned with the high-dimensional case where p/n has a finite non-zero limit. This problem is central to understanding the behavior of robust regression estimators in high-dimension, something that appears to not have been studied before in statistics.

Our analysis in this paper is heuristic but grounded in rigorous methods and relies principally on the concentration of measure phenomenon. Our derivations reveal the importance of the geometry of X_i's in the behavior of $\widehat{\beta}$ - a key to understanding the robustness of our results.

In the case where X_i are i.i.d $\mathcal{N}(0, \text{Id}_p)$, $\beta_0 = 0$, and ϵ_i are i.i.d, our work leads to the following conjecture/heuristic result: $\|\widehat{\beta}\|$ is asymptotically deterministic and if $\hat{z}_\epsilon = \epsilon + \|\widehat{\beta}\|\mathcal{N}(0,1)$, where ϵ has the same distribution as ϵ_i, $\|\widehat{\beta}\|$ has the property that, asymptotically as p and n grow to infinity (while $p \leq n$ and $\limsup p/n < 1$),

$$\begin{cases} \mathbf{E}\left([\text{prox}_c(\rho)]'(\hat{z}_\epsilon)\right) & = 1 - \frac{p}{n} \ , \\ \frac{p}{n}\mathbf{E}\left(\|\widehat{\beta}\|^2\right) & = \mathbf{E}\left([\hat{z}_\epsilon - \text{prox}_c(\rho)(\hat{z}_\epsilon)]^2\right) \ , \end{cases}$$

where prox_c denotes the prox function of ρ (at c) and c is another key parameter in the problem, also determined by the above system. Our predictions are shown to match the results we observe in simulations.

The paper also covers the cases where $\beta \neq 0$ and $\text{cov}(X_i) \neq \text{Id}$. It yields predictions about the intricate behavior of the residuals and the fitted values. Various extensions are presented covering more involved models.

We also show that many well-known facts about robust regression in low-dimension are upended in high-dimension. For instance, when the errors are double-exponential, l_1 regression yields a less efficient estimator of β_0 than l_2 regression provided p/n is large enough.

Our work sheds light on a question raised by P.J Huber in his classic 1973 Annals of Statistics paper.

[*] Support from an Alfred P. Sloan Research Fellowship and NSF grant DMS-0847647 (CAREER) is gratefully acknowledged. **Contact :** `nkaroui@stat.berkeley.edu`

[†] Support from NSF grant DMS-0907362 is gratefully acknowledged.

[‡] Support from NSF Grant DMS-0906808 is gratefully acknowledged. **Key words and Phrases :** Concentration of measure, robust regression, high-dimensional statistics, quantile regression, elliptical distributions

1 Introduction

Statistics and much of data analysis is increasingly high-dimensional. As such, it is important for practitioners and theoreticians alike to understand the behavior of classic statistical methods in the high-dimensional setting. By high-dimensional setting, we mean that we will carry out our study in an asymptotic framework where n the number of observations and p the number of predictors both grow to infinity; we will assume that p/n has a finite limit which we denote by κ.

In this paper, we focus on understanding the behavior of robust regression estimators in the high-dimensional case, which we will interchangeably call regression M-estimates. More specifically, we are interested in the properties of

$$\widehat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho(y_i - X_i'\beta) \;, y_i = X_i'\beta_0 + \epsilon_i \;, X_i \in \mathbb{R}^p,$$

as n and p, grow to infinity while $p/n \to \kappa$. $\{X_i\}_{i=1}^n$ is the set of predictors here and $\{y_i\}_{i=1}^n$ are our responses. y_i are real numbers. We will assume throughout that the loss function ρ is convex.

Robust regression estimators have a long history in statistics (see for instance Huber (1964), Anscombe (1967), Relles (1968), Huber (1972), Andrews et al. (1972), Huber (1973), Yohai (1974), Bickel (1975), Bickel (1981), Bickel (1984), Portnoy (1984), Portnoy (1985), Portnoy (1987), Mammen (1989) and for book-length treatments Huber and Ronchetti (2009), Maronna et al. (2006)) and are widely used in many fields, including econometrics (Koenker (2005)). We shall note in passing that the interest of studying the case p/n not close to zero was already noted in Huber (1973) (case (e) p.802 there) who also pointed out that it would be a nontrivial problem to get a solution in the case of interest here. Our paper can be seen as shedding some light on this now old problem.

Another interesting aspect of the problem is that it is - at its mathematical heart - a question about the behavior of solutions of optimization problems involving random data (i.e M-estimation) in high-dimension. For considerations close in spirit to the ones presented in this paper see El Karoui (2010) and El Karoui (2009b) (relevant for the analysis of the least-squares problem beyond Gaussianity assumptions, among many other things), and El Karoui and Koesters (2011) (relevant for problems related to ridge regression and regularized discriminant analysis and is done under weak distributional assumptions of independent interest for random matrix theory and, we believe, high-dimensional statistics).

Though there has been a lot of classical work on asymptotics for robust regression estimators in the classical setting i.e p/n tends to 0 (though $p \to \infty$ much slower than n was allowed), we are not aware of work that deals with the case where p/n has a non-zero limit. Even simple questions such as the impact of the loss function, or that of the distribution of the errors are not easy to answer. One of the key measures we are interested in is expected squared prediction error. In other words, for a new set of predictors X_{new} and a new response $y_{\text{new}} = \epsilon_{\text{new}} + X_{\text{new}}'\beta_0$, we are interested in $EPE = \mathbf{E}\left((y_{\text{new}} - X_{\text{new}}'\widehat{\beta})^2\right)$. It is easy to see that under our model

$$EPE = \sigma_\epsilon^2 + (\widehat{\beta} - \beta_0)'\text{cov}\left(X_{\text{new}}\right)(\widehat{\beta} - \beta_0) \;.$$

After some manipulations, it turns out that to understand this problem in its general form it is sufficient to understand the case where $\text{cov}\left(X_i\right) = \text{Id}_p$ and $\beta_0 = 0$, in which case $\|\widehat{\beta}\|^2$ is the key quantity of interest. It is also the case that, as we explain in detail later, when X_i's are i.i.d $\mathcal{N}(0,1)$, understanding $\|\widehat{\beta}\|$ is essentially equivalent to understanding the distribution of $\widehat{\beta}$, which justifies further our interest in $\|\widehat{\beta}\|$.

Here is a brief summary of our findings. **1.** When $\{X_i\}_{i=1}^n$ are i.i.d $\mathcal{N}(0,\mathrm{Id}_p)$, $\beta_0 = 0$, and ϵ_i are i.i.d, our work leads to the following conjecture: $\|\widehat{\beta}\|$ is asymptotically deterministic and if $\hat{z}_\epsilon = \epsilon + \|\widehat{\beta}\|\mathcal{N}(0,1)$, where ϵ has the same distribution as ϵ_i and is independent of the normal component of $\hat{z}_\epsilon$, $\|\widehat{\beta}\|$ has the property that, asymptotically as p and n grow to infinity (while $p \leq n$ and $\limsup p/n < 1$),

$$\begin{cases} \mathbf{E}\left([\mathrm{prox}_c(\rho)]'(\hat{z}_\epsilon)\right) & = 1 - \frac{p}{n}\ , \\ \frac{p}{n}\mathbf{E}\left(\|\widehat{\beta}\|^2\right) & = \mathbf{E}\left([\hat{z}_\epsilon - \mathrm{prox}_c(\rho)(\hat{z}_\epsilon)]^2\right)\ , \end{cases}$$

where prox_c denotes the prox function of ρ at c (see Appendix B in case definitions are needed) and c is another key parameter in the problem, also determined by the above system. Our predictions are shown to match the results we observe in simulations. The heuristic analysis we carry out in this paper reveals several other unexpected features for quantities that are well understood in low-dimension. **2.** If $\widehat{\beta}_{(i)}$ is the solution of the robust regression problem with the i-th observation removed, it is false that $X_i'(\widehat{\beta} - \widehat{\beta}_{(i)}) \simeq 0$. **3.** The fitted values are in general non-Gaussian. **4.** We can predict the highly non-Gaussian behavior of the residuals. **5.** In general, both ρ (the loss function) and the distribution of the errors ϵ_i have a significant impact on the behavior of $\|\widehat{\beta}\|$. **6.** Last but not least, classical intuition about the connection between loss function and error distribution is upended: it is sometimes the case that the loss function that appears natural in low-dimension, for instance for maximum likelihood reasons (and indeed can be shown in that setting to lead to more efficient estimators), sees its performance degrade in high-dimension and becomes dominated by another loss function. We illustrate this phenomenon with l_1 loss (median regression) and double exponential errors. It is known that in low-dimension (i.e p/n close to 0) using l_1 loss instead of l_2 loss leads to an estimator of β_0 that is twice as efficient. However, as Figure 6 shows, when p/n is large enough (roughly greater than 0.3 from the graph), it becomes more efficient to use least squares than to use median regression, in sharp contrast with low-dimensional results and intuition.

We propose a heuristic approach in this paper to derive such results - a rigorous mathematical study is underway. We should note that our heuristic itself is grounded in (rigorous) theory. We think that the method we use here is potentially widely useful and reasonably easy to work with for a variety of problems. We also think that it will help non-specialists derive similar results in cases of interest to them. Also, this exposition of “large p, large n” asymptotics allows us to point out some of the key tools we think are useful and hence should have a pedagogical value for many non-specialists. Finally, our heuristics are very well confirmed by simulations, and hence we think they offer a lot of insight in the behavior of robust regression estimators. We should further note that taking the high-dimensional point of view has often been in our experience a good way to get accurate predictions for the behavior of statistical methods even when p and n are not very large. This modern point of view tends to have value even in reasonably classical situations.

In Section 2, we present our approach and a general form of the results. A brief summary of the most important statistical issues follows our statement of the conjecture, on p. 14. In Section 3, we present some applications to various robust regression problems. In high-dimensional statistics problems, it is key to understand the impact of the geometric features of the design matrix. For that purpose we present in Section 4 an analysis pertaining to the robustness of our results and its sensitivity to the geometry of the design matrix. We conclude in Section 5.

2 Main results

As explained above, our analysis here is heuristic, though it is grounded in rigorous use of concentration of measure arguments (some random matrix theory - itself intimately linked with concentration of measure (see e.g El Karoui (2009a)) - is lurking in the background (a fact that will be clear to experts) but our approach manages to avoid relying to it). To derive our heuristics we use a double leave-one-out approach, which we detail below.

Our discussion here will concern the problem:

$$\widehat{\beta}=\text{argmin}_{\beta}\sum_{i=1}^{n}\rho(y_i-X_i'\beta)\;,y_i=X_i'\beta_0+\epsilon_i\;, \tag{1}$$

where X_i are independent $\mathcal{N}(0,\Sigma)$, and ϵ_i are i.i.d with mean 0 and independent of X_i. Here ρ is a convex function. Our end result is a system of two equations that characterizes $\|\widehat{\beta}\|$ (see Conjecture 1 on p. 14). In the case of Gaussian predictors, knowledge of $\|\widehat{\beta}\|$ is sufficient to understand the joint distribution of $\widehat{\beta}$. We also get predictions for the behavior of the residuals.

We will focus only on the case $p<n$ because when $p>n$, in the situations we will be looking at, the problem is under-determined: it is possible to find an infinite number of solutions to this problem, since the null space of $X'X$ is not going to be reduced to zero. We will get back to this point later.

2.1 Preliminaries

We remind the reader of some useful facts that we will use repeatedly in the derivation of our heuristics.

2.1.1 Invariance remarks for Gaussian predictors

Reduction of the problem to 0 signal and Id-covariance situation
We can rewrite the objective function

$$\sum_{i=1}^{n}\rho(y_i-X_i'\beta)$$

as

$$\sum_{i=1}^{n}\rho(\epsilon_i-X_i'(\beta-\beta_0))=\sum_{i=1}^{n}\rho(\epsilon_i-(\beta-\beta_0)'\Sigma^{1/2}X_i^{\text{Id}})$$

where X_i^{Id} is $\mathcal{N}(0,\text{Id}_p)$. When $p\leq n$, the vector space spanned by $\{X_i\}_{i=1}^n$ is $\mathbb{R}^p$ (with probability 1) and hence understanding the properties of

$$\widehat{\beta}_{\text{original}}=\text{argmin}_{\beta\in\mathbb{R}^p}\sum_{i=1}^{n}\rho(y_i-X_i'\beta)\;,y_i=X_i'\beta_0+\epsilon_i\;,$$

is equivalent to understanding the properties of

$$\widehat{\beta}_{\text{simple}}=\text{argmin}_{\beta}\sum_{i=1}^{n}\rho(\epsilon_i-X_i^{\text{Id}}{}'\beta)\;. \tag{2}$$

As a matter of fact, we have (when there is a unique minimizer to our problem)

$$\widehat{\beta}_{\text{original}}=\beta_0+\Sigma^{-1/2}\widehat{\beta}_{\text{simple}}\;.$$

Hence, in the rest of the paper, we study only the properties of $\widehat{\beta}_{\text{simple}}$.

Rotational invariance of the law of $\widehat{\beta}$ when predictors are i.i.d $\boldsymbol{\mathcal{N}(0, \mathrm{Id}_p)}$: Let us focus for a moment on this form of the problem. When X_i is $\mathcal{N}(0, \mathrm{Id}_p)$, if O is an orthogonal matrix, $X_i \overset{\mathcal{L}}{=} OX_i$. Now clearly, if

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

and O is an orthogonal matrix, $\widehat{\beta}(\{OX_i\}; \{\epsilon_i\}) = O\widehat{\beta}(\{X_i\}; \{\epsilon_i\})$. We conclude that when ϵ_i is independent of X_i, for any O, and when there is a unique minimizer to our problem,

$$\widehat{\beta} \overset{\mathcal{L}}{=} O\widehat{\beta} \ .$$

The distribution of $\widehat{\beta}$ is therefore invariant by rotation. It is standard (see Eaton (2007), pp.235-237, Proposition 7.3 and comments) to conclude that

$$\|\widehat{\beta}\| \text{ and } \frac{\widehat{\beta}}{\|\widehat{\beta}\|} \text{ are independent } \ .$$

Also,

$$\frac{\widehat{\beta}}{\|\widehat{\beta}\|} \text{ is uniformly distributed on the sphere of radius 1 in dimension } p \ .$$

The same is true when X_i are of the form $X_i = \lambda_i X_i^{\mathrm{Id}}$, where λ_i are random variables independent of X_i (provided that $\mathrm{Card}\,\{i : |\lambda_i| > 0\} \geq p$), by the same arguments. (In particular, λ_i could be deterministic.)

2.1.2 A quick reminder on concentration of measure

The concentration of measure phenomenon (Ledoux (2001)) is a key building block in our modern understanding of high-dimensional probability and statistics. For the purpose of this paper, we will mostly need the following approximate equality: under regularity condition on the deterministic symmetric matrix A, if X is $\mathcal{N}(0, \mathrm{Id})$, we have

$$\boxed{\frac{X'AX}{p} \simeq \frac{\mathrm{trace}\,(A)}{p} \ .} \tag{3}$$

As a matter of fact, if $\lambda_i(A)$ are the eigenvalues of A,

$$X'AX \overset{\mathcal{L}}{=} \sum_{i=1}^{p} \lambda_i(A) Z_i^2$$

where Z_i are i.i.d $\mathcal{N}(0,1)$. Hence, $\mathrm{var}\,(X'AX) = 2\mathrm{trace}\,(A^2)$. And therefore, $\mathrm{var}\,(X'AX)\,/p^2 \to 0$ whenever $\mathrm{trace}\,(A^2)\,/p^2 \to 0$. This is the type of regularity conditions we have in mind. We also note that if A is stochastic but independent of X, similar statements can be made (by conditioning first on A), though they require a bit more care to be carried out rigorously. (The careful reader will have noticed that A should be indexed by p and the limiting statements we just made are really about sequences of matrices. However, we did not make this precise to avoid cumbersome notations.)

It should be noted that the approximate equality in Equation (3) holds when X has a much more general distribution than just a Gaussian one. It is true as long as the distribution of X satisfies concentration inequalities for convex 1-Lipschitz (with respect to Euclidian norm) functions of X. We refer to Ledoux (2001) for examples of such distributions - see also El Karoui (2009a) for a collection of examples essentially extracted from various parts of Ledoux's monograph. Finally, we note that the concentration of measure phenomenon (often) entails the stronger approximation

$$\sup_{i=1,\ldots,n} \left| \frac{X_i' A X_i}{p} - \frac{\text{trace}\,(A)}{p} \right| \to 0 \;,$$

in probability when $\frac{p}{n} \to \kappa \in (0,\infty)$ (again under various regularity conditions satisfied by the Gaussian distribution). When we approximate quadratic forms in X_i below by their trace without worrying about doing this over ever increasing collections of examples, we essentially appeal to this stronger version of the approximation.

2.1.3 A linear algebraic identity

We will make repeated use in the heuristic derivation that follows of the Sherman-Morrison-Woodbury formula, in its simplest, rank-1 version (see Horn and Johnson (1990), p.19). This formula gives the following identity: if A is an invertible matrix, and u is a vector, we have the rank-1 update formula:

$$(A+uu')^{-1} = A^{-1} - \frac{A^{-1}uu'A^{-1}}{1+u'A^{-1}u} \;.$$

In particular,

$$(A+uu')^{-1}u = \frac{A^{-1}u}{1+u'A^{-1}u} \;\text{ and }\; u'(A+uu')^{-1}u = 1 - \frac{1}{1+u'A^{-1}u} \;.$$

• Consequences for forms in (mildly) dependent random vectors and matrices

This formula will be helpful when we encounter quadratic forms involving a Gaussian random vector and a matrix that depends on this vector. Specifically, if A is of the form $A = \sum_{j=1}^n w_j X_j X_j' = A_i + w_i X_i X_i'$ (with say X_i i.i.d $\mathcal{N}(0, \text{Id}_p)$ and $w_i \geq 0$ deterministic and "well-behaved" in the sense that they do not yield an A with small singular values - these conditions can all be made precise by appealing to random matrix theory), we will be able to make sense of $X_i' A^{-1} X_i$ by using the fact that the previous formula gives us

$$w_i\, X_i' A^{-1} X_i = 1 - \frac{1}{1 + w_i X_i' A_i^{-1} X_i} \;.$$

When A_i is independent of X_i (and its smallest singular value is not too small), our previous discussion of concentration arguments will essentially yield

$$w_i X_i' A^{-1} X_i \simeq 1 - \frac{1}{1 + w_i \text{trace}\left(A_i^{-1}\right)} \;.$$

Though often times it will be the case that trace $\left(A^{-1}\right) \simeq$ trace $\left(A_i^{-1}\right)$, one of the take-away messages of these calculations should be that the dependence between A and X_i (which is often mild, for instance when $w_i = \frac{1}{n}$ for all i) makes the tempting "$X_i' A^{-1} X_i \simeq$ trace $\left(A^{-1}\right)$" (which amounts to using concentration and ignoring dependence between X_i and A) completely false. This is one of the many subtleties we need to keep in mind when working with high-dimensional data (or models).

2.2 Leave-one out approaches

We present a double leave-one out approach which will allow us to conjecture the behavior of $\|\widehat{\beta}\|$ in Equation (2) (see p.14 for a statement of the conjecture). The idea of systematically using leave-one-out methods is tied to our experience in random matrix theory, which is very much hidden in the background of all the derivations that follow.

In the derivation, we will assume that we can take derivatives as we wish, so ρ is smooth (and convex, which we assume throughout). (Later on, our conjecture - verified in simulations - will cover the case of non-smooth ρ.) In accordance with the usage in robust statistics, we call

$$\psi(x) = \rho'(x) \ .$$

We now work with the simplified version of the problem, namely Equation (2); we can do so without loss of generality, as explained above. The gradient characterization of $\widehat{\beta}$ (a.k.a normal equations) is

$$\sum X_i \psi(\epsilon_i - X_i'\widehat{\beta}) = 0 \ . \tag{4}$$

We call the fitted values for this problem $\hat{y}_i = X_i'\widehat{\beta}$ and call the residuals

$$R_i = \epsilon_i - X_i'\widehat{\beta} = y_i - \hat{y}_i \ ,$$

since in the simplified problem, $\beta_0 = 0$ so $y_i = \epsilon_i$.

2.2.1 Leaving out an observation

Let us call $\widehat{\beta}_{(i)}$ the usual leave one out estimator (i.e the estimator we get by not using X_i in our regression problem). It solves

$$\sum_{j \neq i} X_j \psi(\epsilon_j - X_j'\widehat{\beta}_{(i)}) = 0 \ . \tag{5}$$

Note that when $\{X_i\}_{i=1}^n$ are independent, $\widehat{\beta}_{(i)}$ is independent of X_i. For all j, $1 \leq j \leq n$, we call $\tilde{r}_{j,(i)}$

$$\tilde{r}_{j,(i)} = \epsilon_j - X_j'\widehat{\beta}_{(i)} \ .$$

When $j \neq i$, these are the residuals from this leave-one-out situation. For $j = i$, $\tilde{r}_{i,(i)}$ is the prediction error for observation i.

Intuitively, it is reasonable to assume that, when X_i's are i.i.d, for $i \neq j$, $R_j \simeq \tilde{r}_{j,(i)}$. On the other hand, it is easy to convince oneself (by looking e.g at the least-squares situation) that $\tilde{r}_{i,(i)}$ is very different from R_i in high-dimension. The expansion we will get below will indeed confirm this fact in a more general setting than least-squares.

Using this approximation for $j \neq i$, we can do a Taylor expansion in Equation (5). Taking the difference between Equations (4) and (5) we get (assuming that we can neglect the higher order terms)

$$X_i \psi(\epsilon_i - X_i'\widehat{\beta}) + \sum_{j \neq i} \psi'(\tilde{r}_{j,(i)}) X_j X_j' (\widehat{\beta}_{(i)} - \widehat{\beta}) \simeq 0 \ .$$

We call

$$S_i = \sum_{j \neq i} \psi'(\tilde{r}_{j,(i)}) X_j X_j' \ .$$

This suggests that

$$\widehat{\beta} - \widehat{\beta}_{(i)} \simeq S_i^{-1} X_i \psi(\epsilon_i - X_i' \widehat{\beta}) \;.$$

Note that S_i is independent of X_i. Hence, multiplying the previous expression by X_i', we get, using the approximation given in Equation (3) (which amounts to assuming that the largest eigenvalue of S_i^{-1} is not too large),

$$R_i - \tilde{r}_{i,(i)} \simeq -\text{trace}\left(S_i^{-1}\right) \psi(R_i) \;.$$

Our experience in random matrix theory and the form of the matrix S_i suggest that it is not impossible that trace $\left(S_i^{-1}\right)$ could have a deterministic limit. Then, by symmetry between the observations, all trace $\left(S_i^{-1}\right)$ should be approximately the same, i.e,

$$\text{trace}\left(S_i^{-1}\right) \simeq c \;,$$

in which case we would get

$$\boxed{R_i - \tilde{r}_{i,(i)} \simeq -c\psi(R_i) \;.} \tag{6}$$

Note that since X_i and $\widehat{\beta}_{(i)}$ are independent when $\{X_i\}_{i=1}^n$ are independent, much can be said about the distribution of $\tilde{r}_{i,(i)}$. However, at this point it is not clear what the value of c should be. As we will see below, it is possible, from the definition of c to find another equation that characterizes it and links it to $\|\widehat{\beta}\|$. We now look for further information concerning this latter quantity.

2.2.2 Leaving out a predictor

Let us consider what happens when we leave the p-th predictor out (this is in part motivated by the fact that the last column of X is independent of its first (p-1) columns). Because we are assuming that X_i is $\mathcal{N}(0, \text{Id}_p)$, all the predictors play a symmetric role, so we pick the p-th to simplify notations. There is nothing particular about it.

Let us call $\widehat{\gamma}$ the corresponding optimal regression vector (corresponding to the loss function ρ; of course $\widehat{\gamma} \in \mathbb{R}^{p-1}$) and use the notations

$$X_i = \begin{bmatrix} V_i \\ X_i(p) \end{bmatrix} \;, \text{ where } V_i \text{ is a } (p-1) \text{ dimensional vector, and}$$

$$\widehat{\beta} = \begin{bmatrix} \widehat{\beta}_{S_p} \\ \widehat{\beta}_p \end{bmatrix} \;.$$

Naturally, $\widehat{\gamma}$ satisfies

$$\sum_{i=1}^n V_i \psi(\epsilon_i - V_i' \widehat{\gamma}) = 0 \;.$$

We call

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

i.e the residuals based on $p-1$ predictors. Note that $\{r_{i,[p]}\}_{i=1}^n$ is independent of $\{X_i(p)\}_{i=1}^n$ under our assumptions (because V_i is independent of $X_i(p)$ and the X_i's are i.i.d).

It is intuitively clear that $R_i \simeq r_{i,[p]}$, for all i, since adding a predictor will not help us much in estimating the 0 vector (the true regression vector in the situation we investigate), and hence the residuals should not be much affected by the addition of one predictor. Taking the difference of the equations defining $\widehat{\beta}$ and $\widehat{\gamma}$, we get

$$\sum_i X_i \psi(\epsilon_i - X_i' \widehat{\beta}) - \begin{bmatrix} V_i \\ 0 \end{bmatrix} \psi(\epsilon_i - V_i' \widehat{\gamma}) = 0 \;.$$

This p-dimensional equation separates into a scalar and a vector equation, namely,

$$\sum_i X_i(p)\psi(\epsilon_i - X_i'\widehat{\beta}) = 0 \ ,$$
$$\sum_i V_i[\psi(R_i) - \psi(r_{i,[p]})] = 0_{p-1} \ .$$

Using a Taylor expansion of $\psi(R_i)$ around $\psi(r_{i,[p]})$ and noting that $R_i - r_{i,[p]} = V_i'(\widehat{\gamma} - \widehat{\beta}_{S_p}) - X_i(p)\widehat{\beta}_p$, we can transform the first equation above into (neglecting the higher-order terms)

$$\sum_i X_i(p)\left[\psi(r_{i,[p]}) + \psi'(r_{i,[p]})(V_i'(\widehat{\gamma} - \widehat{\beta}_{S_p}) - X_i(p)\widehat{\beta}_p)\right] \simeq 0 \ .$$

This in turns gives the near identity

$$\widehat{\beta}_p \simeq \frac{\sum X_i(p)[\psi(r_{i,[p]}) + \psi'(r_{i,[p]})V_i'(\widehat{\gamma} - \widehat{\beta}_{S_p})]}{\sum X_i^2(p)\psi'(r_{i,[p]})} \ .$$

Working similarly on the equations involving V_i, we get

$$\sum_i \psi'(r_{i,[p]})V_i[R_i - r_{i,[p]}] \simeq 0 \ .$$

Since $R_i - r_{i,[p]} = -\widehat{\beta}_p X_i(p) + V_i'(\widehat{\gamma} - \widehat{\beta}_{S_p})$, the previous equation reads

$$\left[\sum_i \psi'(r_{i,[p]})V_iV_i'\right](\widehat{\gamma} - \widehat{\beta}_{S_p}) - \widehat{\beta}_p \sum_i \psi'(r_{i,[p]})V_iX_i(p) \simeq 0 \ .$$

Calling

$$\mathfrak{S}_p = \sum_i \psi'(r_{i,[p]})V_iV_i' \ , \text{ and } u_p = \sum_i \psi'(r_{i,[p]})V_iX_i(p) \ ,$$

we see that

$$(\widehat{\gamma} - \widehat{\beta}_{S_p}) \simeq \widehat{\beta}_p \mathfrak{S}_p^{-1} u_p \ .$$

Therefore, after we plug this approximation in the previous equation for $\widehat{\beta}_p$, we have

$$\widehat{\beta}_p \simeq \frac{\sum X_i(p)\psi(r_{i,[p]})}{\sum X_i^2(p)\psi'(r_{i,[p]})} + \widehat{\beta}_p \frac{u_p'\mathfrak{S}_p^{-1}u_p}{\sum X_i^2(p)\psi'(r_{i,[p]})} \ ,$$

and finally

$$\boxed{\widehat{\beta}_p \simeq \frac{\sum X_i(p)\psi(r_{i,[p]})}{\sum X_i^2(p)\psi'(r_{i,[p]}) - u_p'\mathfrak{S}_p^{-1}u_p} \ .} \tag{7}$$

(As an aside, it is interesting to notice that the first ratio in the penultimate formula above is very similar to the classical case where $p/n \to 0$.)

On $u_p'\mathfrak{S}_p^{-1}u_p$

If D is a diagonal matrix with diagonal $\psi'(r_{i,[p]})$, we see that $\mathfrak{S}_p = V'DV$, where V is the $n \times (p-1)$ dimensional matrix containing the V_i's as rows. With these notations, we can write u_p in matrix form as

$$u_p = V'DX(p) \ ,$$

where $X(p)$ is the n-dimensional vector whose i-th coordinate is $X_i(p)$. So, if we call $B = D^{1/2}V$ and $A = (DV)(V'DV)^{-1}V'D$, we have $A = D^{1/2}B(B'B)^{-1}B'D^{1/2}$, and

$$u_p'\mathfrak{S}_p^{-1}u_p = X(p)'AX(p) \;,$$

Conditional on $\{V_i, \epsilon_i\}_{i=1}^n$, $u_p'\mathfrak{S}_p^{-1}u_p$ is therefore a weighted χ^2. In high-dimension, if the eigenvalues of A are not dominated by a few ones, we can write (essentially appealing to a concentration of measure argument - see Equation (3))

$$X(p)'AX(p) \simeq \operatorname{trace}(A) = \sum_{i=1}^{n} \psi'(r_{i,[p]})p_{ii} \;,$$

where p_{ii} is the i-th entry on the diagonal of P, the projection matrix $P = B(B'B)^{-1}B'$. Using this approximation, we see (based on Equation (7)) that

$$\widehat{\beta}_p \simeq \frac{\sum X_i(p)\psi(r_{i,[p]})}{\sum X_i^2(p)\psi'(r_{i,[p]})(1-p_{ii})} \;. \tag{8}$$

Work on this denominator
$\{X_i(p)\}_{i=1}^n$ is independent of $\{r_{i,[p]}\}_{i=1}^n$ and $\{p_{ii}\}_{i=1}^n$, and it is reasonable to expect (at least for well-behaved ψ') by exchangeability that a law of large number phenomenon might happen and help us understand the denominator. In that case, we might be able to replace $\{X_i^2(p)\}_{i=1}^p$ by their expected value, namely 1, to understand the behavior of the sum (this amounts to taking an expectation conditional on $\{r_{i,[p]}\}_{i=1}^n$ and $\{p_{ii}\}_{i=1}^n$). In other words, we expect that

$$\frac{1}{n}\sum X_i^2(p)\psi'(r_{i,[p]})(1-p_{ii}) \simeq \frac{1}{n}\sum \psi'(r_{i,[p]})(1-p_{ii}) \;.$$

Using the rank-1 update formula for matrix inversion (see Horn and Johnson (1990), p. 19 and our discussion above), if we call $\mathfrak{S}_p(i) = \mathfrak{S}_p - \psi'(r_{i,[p]})V_iV_i'$, we have

$$p_{ii} = \psi'(r_{i,[p]})\frac{V_i'[\mathfrak{S}_p(i)]^{-1}V_i}{1+\psi'(r_{i,[p]})V_i'[\mathfrak{S}_p(i)]^{-1}V_i}$$

and therefore

$$1-p_{ii} = \frac{1}{1+\psi'(r_{i,[p]})V_i'[\mathfrak{S}_p(i)]^{-1}V_i} \;.$$

$\mathfrak{S}_p(i)$ is not independent of V_i, because V_i plays a role in the computation of $\{r_{j,[p]}\}_{j\neq i}$. However, it is perhaps not unreasonable to believe that, at least for well-behaved ψ', the diagonal matrix with entries $\{\psi'(r_{j,[p]})\}_{j\neq i}$ could be approximated by another diagonal matrix, which would be independent of V_i (to do this rigorously, we will essentially have to do a leave-one-observation-out type of argument, leaving out V_i and working on the corresponding diagonal matrix).

In that case, we would be inclined to think that, for well-behaved ψ' functions, by Equation (3) (concentration of measure, again),

$$V_i'[\mathfrak{S}_p(i)]^{-1}V_i \simeq \operatorname{trace}\left([\mathfrak{S}_p(i)]^{-1}\right) \simeq \operatorname{trace}\left(\mathfrak{S}_p^{-1}\right) \simeq c \;.$$

(The second near equality comes from applying the rank-1 update formula to $\mathfrak{S}_p^{-1}$ and realizing that if $\operatorname{trace}\left([\mathfrak{S}_p(i)]^{-1}\right)$ is of order 1, then $\operatorname{trace}\left([\mathfrak{S}_p(i)]^{-2}\right)$ should be close to 0.)

Because P is a projection matrix, which we expect (for well-behaved ψ' and X_i having a continuous distribution) to be of rank $p-1$, we have

$$\sum_{i=1}^n (1-p_{ii}) = n-(p-1) \simeq n-p \; .$$

Therefore, we also expect that (given our arguments leading to $1-p_{ii} \simeq 1/(1+\psi'(r_{i,[p]})c)$),

$$\boxed{\frac{1}{n}\sum \frac{1}{1+\psi'(r_{i,[p]})c} \simeq 1-\frac{p}{n} \; .} \tag{9}$$

In particular,

$$\frac{1}{n}\sum_{i=1}^n \psi'(r_{i,[p]})(1-p_{ii}) \simeq \frac{1}{n}\sum \frac{\psi'(r_{i,[p]})}{1+c\psi'(r_{i,[p]})} = \frac{1}{c}\frac{1}{n}\sum[1-\frac{1}{1+c\psi'(r_{i,[p]})}] \simeq \frac{1}{c}\frac{1}{n}\sum p_{ii} \simeq \frac{p}{n\,c} \; .$$

Recalling that we expect that $\frac{1}{n}\sum X_i^2(p)\psi'(r_{i,[p]})(1-p_{ii}) \simeq \frac{1}{n}\sum \psi'(r_{i,[p]})(1-p_{ii})$, we finally get, after plugging all these approximations in Equation (8), that

$$\boxed{\widehat{\beta}_p \simeq \frac{1}{p}\text{trace}\left(\mathfrak{S}_p^{-1}\right)\sum X_i(p)\psi(r_{i,[p]}) \simeq \frac{c}{p}\sum X_i(p)\psi(r_{i,[p]}) \; .} \tag{Approx}$$

Since $r_{i,[p]}$ is independent of $X_i(p)$ and $\{X_i(p)\}_{i=1}^n$ are independent of each other, we have

$$\mathbf{E}\left(\widehat{\beta}_p^2|\{V_i,\epsilon_i\}_{i=1}^n\right) = \frac{1}{p}\frac{n}{p}\left[\frac{1}{n}\sum c^2\psi^2(r_{i,[p]})\right] \simeq \frac{1}{p}\frac{n}{p}\left[\frac{1}{n}\sum c^2\psi^2(R_i)\right] \; ,$$

where the last approximate equality relies on $r_{i,[p]} \simeq R_i$ and ψ^2 being smooth ("most of the time").

In fact, our argument had nothing to do with the last coordinate and can be repeated for all coordinates of $\widehat{\beta}$. Doing so and summing over all coordinates, we have the approximation

$$\mathbf{E}\left(\|\widehat{\beta}\|^2\right) \simeq \frac{n}{p}\left[\frac{1}{n}\sum \mathbf{E}\left(c^2\psi^2(R_i)\right)\right] \; . \tag{10}$$

We also note that if we assume that $\{c\psi(R_i)\}_{i=1}^n$ is well-behaved (so a few terms do not dominate the others in the L^2 sense) and it has a (deterministic) limiting distribution with a finite second moment, we could drop the expectation in the right hand side of the previous expression.

It should also be noted that Equation (Approx) gives us a distributional approximation for $\widehat{\beta}_p$ (!).

2.2.3 Asymptotically deterministic character of $\|\widehat{\beta}\|$

It is also reasonable, given the approximate stochastic representation we have for $\widehat{\beta}_p$ in Equation (Approx), to gather that $\|\widehat{\beta}\|$ should have a deterministic limit. We give now an argument tied to our derivation, though this could also be seen as coming from more general principles tied to the Efron-Stein inequality and some of our earlier approximations.

Indeed, for well-behaved ψ, the independence of $r_{i,[p]}$ and $X_i(p)$ gives (assuming the approximation (Approx)) that

$$\frac{p}{\sqrt{n}}\widehat{\beta}_p \Longrightarrow \mathcal{N}\left(0, \frac{c^2}{n}\sum_{i=1}^n \psi^2(r_{i,[p]})\right) \; ,$$

where we are also implicitly assuming that $\frac{c^2}{n}\sum_{i=1}^n \psi^2(r_{i,[p]})$ converges in probability to a deterministic limit (as usual, $\Longrightarrow$ denotes weak convergence). If that is the case, the rotational invariance of $\widehat{\beta}$ (see Subsubsection 2.1.1) guarantees that $\|\widehat{\beta}\|$ is asymptotically deterministic. As a matter of fact, we know (according to Subsubsection 2.1.1) that, in the case where X_i are i.i.d $\mathcal{N}(0, \mathrm{Id})$,

$$\widehat{\beta}_p \overset{\mathcal{L}}{=} \|\widehat{\beta}\|\nu_p \ ,$$

where ν_p is independent of $\|\widehat{\beta}\|$ and ν_p is the last coordinate of a vector taken at random on the unit sphere. Hence, $\nu_p \overset{\mathcal{L}}{=} Z_p/\sqrt{Z_p^2 + \chi_{p-1}^2}$, where Z_p is $\mathcal{N}(0,1)$ and independent of χ_{p-1}^2. Consequently, $\sqrt{p}\nu_p$ is asymptotically $\mathcal{N}(0,1)$. Therefore, $\sqrt{p}\widehat{\beta}_p$ has in general a scale mixture of normal distribution, and the only way it can be asymptotically normal is when $\|\widehat{\beta}\|$ is asymptotically deterministic. (To do this formally, one can compare the 4-th moment of $\widehat{\beta}_p$ to its second moment, use normality of $\widehat{\beta}_p$ to draw a relation between the two and conclude that the only way this relation can be true is when var $\left(\|\widehat{\beta}\|^2\right) \simeq 0$.)

2.2.4 Summary of our approximations

Recall the notations

$$\begin{aligned} \tilde{r}_{i,(i)} &= \epsilon_i - X_i'\widehat{\beta}_{(i)} \\ R_i &= \epsilon_i - X_i'\widehat{\beta} \ . \end{aligned}$$

The work of the previous section gives us the following intuition: we should have, according to Equations (6), (9), (Approx), and the approximation $\psi'(r_{i,[p]}) \simeq \psi'(R_i)$,

$$\begin{aligned} \tilde{r}_{i,(i)} &= R_i + c\psi(R_i) \ , \\ \frac{1}{n}\sum \frac{1}{1+\psi'(R_i)c} &\simeq 1 - \frac{p}{n} \ , \text{ and} \\ \mathbf{E}\left(\|\widehat{\beta}\|^2\right) &\simeq \frac{n}{p}\left[\frac{1}{n}\sum c^2\psi^2(R_i)\right] \ , \end{aligned}$$

where c is

$$c = \text{trace}\left(\left[\sum_{j\neq i}\psi'(\tilde{r}_{j,(i)})X_jX_j'\right]^{-1}\right) \ ,$$

and $\tilde{r}_{j,(i)} = \epsilon_j - X_j'\widehat{\beta}_{(i)}$.

The key idea is the following: we note that when X_i are i.i.d $\mathcal{N}(0, \mathrm{Id}_p)$ (as they have been assumed to be to get the previous equations),

$$\tilde{r}_{i,(i)} \sim \epsilon_i - \|\widehat{\beta}_{(i)}\|Z$$

where Z is $\mathcal{N}(0,1)$. It is also reasonable to expect that $\|\widehat{\beta}_{(i)}\| \simeq \|\widehat{\beta}\|$ as n and p tend to infinity. Hence, the law of $\tilde{r}_{i,(i)}$ is going to be approximable by

$$\tilde{r}_{i,(i)} \sim \epsilon_i - \|\widehat{\beta}\|Z \text{ where } Z \sim \mathcal{N}(0,1)$$

So from all these heuristics we are going to extract a system of two equations in two (deterministic) unknowns, c and $\|\widehat{\beta}\|$ and solve it.

2.3 Setting up the system of equations for c and $\|\widehat{\beta}\|$

As we have argued before, we expect (when our approximations are valid) that $\|\widehat{\beta}\|$ will be asymptotically deterministic. So we will from now on replace quantities like $\mathbf{E}\left(\|\widehat{\beta}\|^k\right)$ by $\|\widehat{\beta}\|^k$.

2.3.1 First attempt at characterizing the solution

Let us call

$$g_c(x) = x + c\psi(x) \;, \tag{11}$$

where c is the particular quantity we referred to before with this notation (not a dummy variable). Note that c is unknown at this point, but $c > 0$ given its definition. Also, if ρ is convex, ψ is increasing so g_c is invertible. The equation $\tilde{r}_{i,(i)} = R_i + c\psi(R_i)$ (Equation (6) above) can be rewritten

$$R_i = g_c^{-1}(\tilde{r}_{i,(i)}) \;. \tag{12}$$

We now make the following important remark (which is easily verified in hindsight but considerably harder to see a priori - indeed, it was probably the single hardest step in discovering our conjectured solution to the problem): Equation (9) means in this language (assuming the $\tilde{r}_{i,(i)}$ are not too correlated so we can replace the empirical expectation by the population version) that

$$\mathbf{E}\left(\left(g_c^{-1}\right)'(\tilde{r}_{i,(i)})\right) \simeq 1 - \frac{p}{n} \;, \tag{13}$$

since $g_c'(x) = 1 + c\psi'(x)$. Finally, Equation (10) can be rewritten

$$\frac{p}{n}\mathbf{E}\left(\|\widehat{\beta}\|^2\right) \simeq \mathbf{E}\left((\tilde{r}_{i,(i)} - g_c^{-1}(\tilde{r}_{i,(i)}))^2\right) \;. \tag{14}$$

So **our new problem** is the following: find c and $\|\widehat{\beta}\|$, two constants, such that for g_c defined above in Equation (11),

$$\boxed{\begin{cases} \mathbf{E}\left(\left(g_c^{-1}\right)'(\tilde{r}_{i,(i)})\right) & \simeq 1 - \frac{p}{n} \;, \\ \frac{p}{n}\mathbf{E}\left(\|\widehat{\beta}\|^2\right) & \simeq \mathbf{E}\left((\tilde{r}_{i,(i)} - g_c^{-1}(\tilde{r}_{i,(i)}))^2\right) \;. \end{cases}} \tag{Key Functional System}$$

Of course, and this is crucial, the distribution of $\tilde{r}_{i,(i)}$ is known (up to $\|\widehat{\beta}\|$ and it is a convolution of a normal (with unknown variance) and ϵ_i).

The first of this equation will give us a functional relationship between $\|\widehat{\beta}\|$ and c. Plugging that in the second one, we should or might be able to obtain c and/or $\|\widehat{\beta}\|$.

2.3.2 A formulation for more general ρ: final version of the system

It is of course of great interest for statisticians to be able to work with non-differentiable functions ρ: this happens for instance in quantile regression Koenker (2005). However, because in the derivation of our system we had to use ψ', it is not clear, even heuristically, what one should do or guess when the loss function ρ is not differentiable.

However, it is well-known in optimization that, if ∂f is the subdifferential of a closed proper convex function f (and hence a multivalued function), $(\text{Id} + t\partial f)^{-1}$ is a single valued function. This function is closely related to the Moreau-Yosida regularization of a convex function (see Moreau

(1965), Rockafellar (1997), Hiriart-Urruty and Lemaréchal (2001)) and is often called the prox-function or proximal mapping. We give more background (based on Moreau (1965)) on it in the Appendix, subsubsection B-1.

Hence a better conjecture is derived from the following definitions. We consider for $t > 0$,

$$f_t(y; u) = \left\{ f(u) + \frac{1}{2t}(u - y)^2 \right\}$$

and call

$$\text{prox}_t(f)(y) = \text{argmin}_u \, f_t(y; u)$$

Under regularity conditions, namely when f is a closed proper convex function, there is a unique minimizer to $f_t(y; u)$ at t and y given. Note that $\text{prox}_t(f)$ can be defined even when f is not differentiable, as is the case when working with l_1 loss.

We are now in position to state our conjecture.

Heuristic Result 1. *Consider*

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

where ϵ_i are i.i.d and independent of X_i, which are i.i.d $\mathcal{N}(0, \text{Id}_p)$. ϵ_i are also assumed to be reasonably nice, for instance they have infinitely many moments. Assume further that $p \leq n$ and p/n stays bounded away from 1 (i.e $\limsup p/n < 1$).

Then $\|\widehat{\beta}\|$ is asymptotically deterministic (as n and p tend to infinity) and is characterized through the following system of equations.

If c is another deterministic parameter and if $\hat{z}_\epsilon \sim \mathcal{N}(0, \|\widehat{\beta}\|^2) + \epsilon$, where ϵ has the same distribution as ϵ_i and is independent of the normal component of $\hat{z}_\epsilon$,

$$\boxed{\left\{ \begin{array}{ll} \mathbf{E}\left((prox_c(\rho))' (\hat{z}_\epsilon) \right) & \simeq 1 - \frac{p}{n} \ , \\ \frac{p}{n} \mathbf{E}\left(\|\widehat{\beta}\|^2 \right) & \simeq \mathbf{E}\left((\hat{z}_\epsilon - prox_c(\rho)(\hat{z}_\epsilon))^2 \right) \ . \end{array} \right.} \qquad \text{(KeyProxSystem)}$$

We further conjecture that this result remains true when X_i satisfies mild concentration requirements for quadratic forms as outlined above.

We will discuss a lack of robustness conjecture below. We recall that c is conjectured to be the limit of $c = \text{trace}\left(\sum_{j \neq i} (\psi'(\tilde{r}_{j,(i)}) X_j X_j')^{-1} \right)$,.

Finally, it should also be noted that the system can be reformulated by using only ρ^*, the (Fenchel-Legendre) conjugate of ρ, for we have the identity $x - \text{prox}_c(\rho)(x) = \text{prox}_1((c\rho)^*)(x)$ (see Appendix B). Hence the first equation in our system reads $p/n = \mathbf{E}\left([\text{prox}_1((c\rho)^*)]'(\hat{z}_\epsilon) \right)$.

Joint distributional behavior of $\widehat{\boldsymbol{\beta}}$: We also recall that by invariance, when X_i are i.i.d $\mathcal{N}(0, \text{Id}_p)$, $\widehat{\beta}/\|\widehat{\beta}\|$ is uniformly distributed on the unit sphere in dimension p and is independent of $\|\widehat{\beta}\|$. Hence, our conjecture, if and when verified would completely characterize the joint(!) distributional behavior of $\widehat{\beta}$.

Dealing with $\boldsymbol{\beta_0 \neq 0}$ and $\mathbf{cov}\,\boldsymbol{(X_i) = \Sigma}$: A few simple observations are in order: for the original problem, where $y_i = X_i'\beta_0 + \epsilon_i$, X_i are i.i.d $\mathcal{N}(0, \Sigma)$, and $\widehat{\beta} = \text{argmin}_\beta \sum_{i=1}^n \rho(y_i - X_i'\beta)$, the previous conjecture completely characterizes $\Sigma^{1/2}(\widehat{\beta} - \beta_0)$. We also note that our invariance arguments (Subsubsection 2.1.1) give a complete characterization of the law of $\Sigma^{1/2}(\widehat{\beta} - \beta_0)$, at least in the Gaussian case. Our experience with random matrix theory suggests that a number of those properties will be conserved when we replace the Gaussianity assumption on X_i by an assumption on concentration of quadratic forms in X_i, i.e a guarantee that for A independent of X_i, $X_i'AX_i \simeq \text{trace}\,(A)$ when X_i has covariance Id_p.

Remarks on the residuals and the fitted values: It should also be noted that as a by-product of the heuristic analysis, we obtain a conjecture for the distributions of the residuals and the fitted values. These distributions turn out to be in general quite complicated. For the residuals, which we called earlier R_i, we had $R_i \simeq \text{prox}_c(\rho)(\tilde{r}_{i,(i)})$ and hence the marginal distribution of the residuals should converge to $\text{prox}_c(\rho)(\hat{z}_\epsilon)$ where $\hat{z}_\epsilon = \epsilon + \mathcal{N}(0, \|\widehat{\beta}\|^2)$.
For the fitted values, we can use the representation $X_i'\widehat{\beta} = X_i'(\widehat{\beta} - \widehat{\beta}_{(i)}) + X_i'\widehat{\beta}_{(i)}$ and our approximations to see that

$$X_i'\widehat{\beta} \simeq \epsilon_i - \text{prox}_c(\rho)(\tilde{r}_{i,(i)}) ,$$

where $\tilde{r}_{i,(i)} \simeq \epsilon_i + \|\widehat{\beta}\|Z$ and Z is $\mathcal{N}(0,1)$. In general, the distribution of the fitted values is going to be very different from a normal distribution - a fact already pointed out in the least squares case in Huber (1973), pp. 802-804. (In the least-squares situation, one can look at $X_i'\widehat{\beta}_{LS}$ and use the rank-1 update formula to arrive at this conclusion rigorously and find the corresponding distribution.)
This might be a bit surprising at first, given that we have argued that the finite dimensional distributions of $\widehat{\beta}$ are going to be normal (indeed more is true when the predictors are Gaussian). However, this can be understood as a manifestation of the fact that $\widehat{\beta}$ (and in particular the angle $\widehat{\beta}/\|\widehat{\beta}\|$) is correlated (in a non-negligible fashion) with each individual X_i's and we cannot neglect this dependence when projecting $\widehat{\beta}$ along X_i. In other words, in high-dimension, there is no asymptotic independence between the two. This is at a high level similar to the dependence issues we hinted at in Subsubsection 2.1.3.

2.3.3 Other possible approaches

We tried a variety of approaches to understand the behavior of $\widehat{\beta}$, at this point all heuristics, and were unsuccessful with the other attempts we made.

We tried a belief-propagation approach (Mézard and Montanari (2009)), which has been successfully applied to understand the behavior of Lasso (Donoho et al. (2009b) and the related Donoho et al. (2009a)), but in our applications gave only the right prediction for Gaussian errors and l_2 loss. Indeed it seemed to predict in the general case that neither ρ nor the distribution of ϵ_i mattered. However, we are not experts in this method and it is possible that someone who is very used to it might be able to successfully use it (or one of its variants) in our setting. (Of course, understanding of the applicability of heuristics varies greatly with fields and experience.)

We also tried to look at relevant computations in the spin-glass literature, specifically in Talagrand (2003). The questions tackled there can be seen as studying the properties of random probability distributions some of which having density of the form

$$f(x) = \frac{1}{Z}\exp[-t\sum_{i=1}^{n}\rho(X_i'x)] ,$$

over certain domains for x (for instance $\{-1,1\}^p$) and for various ρ's. However, we were unsuccessful (after a limited amount of time) in extracting heuristics that could guide us to a result for our problem (which could be seen (at least heuristically) as a limiting case of a difficult spin glass problem) from the very precise, detailed, and rigorous investigations done in this book. (Of course, this is more a measure of our current limited understanding of that subject and the techniques used there than a reflection on the book.) We note that leave-one-out is also used in spin glass problems (under the name "cavity method") but our approach to the problem seems completely different from what is done in Talagrand (2003).

In summary, we see several benefits to the approach we have presented here: it does not rely on heuristics such as the replica method or belief propagation techniques that still appear to us as somewhat unclear and whose validity are difficult to assess a priori (even by some specialists it seems, who could not answer these validity questions when we asked them at various conferences). By contrast, what we have presented here has the benefit of simplicity, is fairly grounded in rigorous mathematics (though of course there are gaps (in rigor) since it is only a heuristic at this point), and clearly points to the impact of our assumptions on the results - specifically the role of the concentration of measure phenomenon in driving the results.

2.3.4 The case $p > n$

The case where $p > n$ is of course interesting and presents itself in modern data analysis. When $p > n$, practitioners often like to add penalties to the problem. Our method of analysis also yields results in this penalized situation but they will be presented in another paper. The main issue for our current problem is that when $p > n$, there are infinitely many solutions because we can find infinitely many $\widehat{\beta}$'s such that, for all i, $X_i'\beta_0 + \epsilon_i = X_i'\widehat{\beta}$. Therefore, one needs to make a choice between them. A common choice is to pick in this collection of $\widehat{\beta}$'s the one with minimum l_2 norm, which generally makes sense when one is interested in controlling expected prediction error. It is well-known that the solution to this problem is given by

$$\widehat{\beta} = X^{\dagger}(X\beta_0 + \epsilon) ,$$

where $X^{\dagger}$ is the Moore-Penrose inverse of X (see Penrose (1956) and Ben-Israel and Greville (2003), p.109). The computation of $\|\widehat{\beta}\|$ and $\|\widehat{\beta} - \beta_0\|$ can be done using standard tools of Wishart theory (this has effectively nothing to do with the main points of the current paper so we do not delve on this question) and we have done it in a separate paper (not yet submitted).

3 Applications

We present here some applications of our conjecture and essentially verify it in simulations. We will use our heuristic result to predict the behavior of

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

for a variety of convex functions ρ. (Some examples are in the main text. More can be found in Appendix A, where we tackle Huber loss functions, quantile regression and a few other examples.)

This is statistically very interesting in our opinion, because it should help us shed light on the question of the impact of the distribution of the errors on the behavior of the solution and also on the role of the loss function ρ. These are highly non-trivial a priori.

3.1 A preliminary remark on the Gaussian error case.

Before we give concrete examples, we make the remark that the case of Gaussian errors is particularly simple, because $\hat{z}_\epsilon$ is then Gaussian.

Furthermore, the system (KeyProxSystem) simplifies slightly when $\hat{z}_\epsilon \sim \mathcal{N}(0, s^2)$. Indeed, recalling the classic Gaussian integration by part formula (a.k.a Stein's formula, Stein (1981), Lemma 1 and Equation 2.3), we know that when a random variable W is $\mathcal{N}(0,\sigma^2)$, then under mild technical conditions on the function f,

$$\mathbf{E}\left(Wf(W)\right)=\sigma^2\mathbf{E}\left(f'(W)\right)\ .$$

Since the first equation of our system fixes the value of $\mathbf{E}\left([\text{prox}_c(\rho)]'(\hat{z}_\epsilon)\right)$ at $1-\frac{p}{n}$, the second equation in the system reads, after expanding the square

$$\frac{p}{n}\mathbf{E}\left(\|\widehat{\beta}\|^2\right)=s^2-2s^2(1-\frac{p}{n})+\mathbf{E}\left((\text{prox}_c(\rho)(sZ))^2\right)\ ,\text{ where } Z\sim\mathcal{N}(0,1)\ .$$

3.2 Case $\rho(x)=x^2/2$; least-squares regression

The case where $\rho(x)=x^2/2$ amounts to solving a least-squares problem. Fortunately, in this case the solution is known explicitly. If X is an $n\times p$ matrix whose rows are X_i, and Y is our vector of responses,

$$\widehat{\beta}_{l_2}=(X'X)^{-1}X'Y\ .$$

In particular, when $y_i=\epsilon_i$ and $\text{cov}\left(\{\epsilon_i\}_{i=1}^n\right)=\sigma_\epsilon^2\text{Id}_n$, we see that $\mathbf{E}\left(\|\widehat{\beta}_{l_2}\|^2\right)=\sigma_\epsilon^2\mathbf{E}\left(\text{trace}\left((X'X)^{-1}\right)\right)$. This latter quantity is equal to $\sigma_\epsilon^2 p/(n-p)$ from classical Wishart theory (Muirhead (1982) or Anderson (2003)); the result holds at least when $n-p>2$.

Let us now compare this rigorous result with the predictions given by our conjecture.

In this case, $\psi(x)=x$, so

$$g_c(x)=(1+c)x\ .$$

Hence

$$g_c^{-1}(y)=\frac{1}{1+c}y\ .$$

Also, $y-g_c^{-1}(y)=c/(1+c)y$. Equation (13) becomes

$$\mathbf{E}\left(\frac{1}{1+c}\right)=1-\frac{p}{n}=\frac{1}{1+c}\ .$$

Equation (14) becomes

$$\frac{p}{n}\mathbf{E}\left(\|\widehat{\beta}\|^2\right)=\left(\frac{c}{1+c}\right)^2\mathbf{E}\left(\hat{z}_\epsilon^2\right)=\left(\frac{c}{1+c}\right)^2\left(\mathbf{E}\left(\|\widehat{\beta}\|^2\right)+\sigma_\epsilon^2\right).$$

So we get

$$c_{l_2}=\frac{\frac{p}{n}}{1-\frac{p}{n}}$$

and

$$\mathbf{E}\left(\|\widehat{\beta}_{l_2}\|^2\right)\simeq\frac{\frac{p}{n}}{1-\frac{p}{n}}\sigma_\epsilon^2\ .$$

Hence our conjecture agrees with classical results derived from Wishart theory. The analysis and implementation of our conjecture also clearly shows why the distribution of ϵ_i should not matter beyond the value of the variance, σ_ϵ^2.

3.3 Case $\rho(x) = |x|$; median regression

Here the Taylor expansion we relied on to derive the conjecture are very suspect, but let us see what the system (KeyProxSystem) predicts. In this case $\psi(x) = \text{sign}(x)$ (except at 0) and it is well-known and easily verified that, if ρ_{l_1} is such that $\rho_{l_1}(x) = |x|$,

$$\text{prox}_t(\rho_{l_1})(y) = (y - t)1_{y \geq t} \text{ when } y \geq 0 \; .$$

By symmetry, $\text{prox}_t(\rho_{l_1})(-y) = -\text{prox}_t(\rho)(y)$. In other words, the prox-function is the soft-thresholding function.

Therefore, we have

$$y \geq 0 \, : \, \left[\text{prox}_t(\rho_{l_1})\right]'(y) = 1_{y \geq t} \; ,$$

and the first equation of the system (KeyProxSystem) becomes

$$\mathbf{E}\left(\left[\text{prox}_c(\rho_{l_1})\right]'(\hat{z}_\epsilon)\right) = \mathbf{E}\left(1_{|\hat{z}_\epsilon| \geq c}\right) = P(|\hat{z}_\epsilon| \geq c) \simeq 1 - \frac{p}{n} \; . \tag{15}$$

The reader might be potentially concerned about our taking the derivative of the soft-thresholding function. However, since $\hat{z}_\epsilon$ has a density which is the convolution of a Gaussian density and another density, the fact that $\text{prox}_t(\rho_{l_1})$ is not differentiable at t and $-t$ does not cause any trouble in computing the expectation.

3.3.1 The case of Gaussian errors

The Gaussian error case, $\hat{z}_\epsilon \sim \mathcal{N}(0, \|\widehat{\beta}\|^2 + \sigma_\epsilon^2)$. Calling

$$s^2 = \|\widehat{\beta}\|^2 + \sigma_\epsilon^2 \; ,$$

our first relation between $\|\widehat{\beta}\|$ (or s) and c, i.e Equation (15) becomes

$$P(|Z| \geq \frac{c}{s}) = 1 - \frac{p}{n} \; , \text{ where } Z \sim \mathcal{N}(0,1) \; .$$

After some algebra, we get that

$$\frac{c}{s} = \Phi^{-1}\left(\frac{1}{2}\left[1 + \frac{p}{n}\right]\right) \; ,$$

where Φ^{-1} is the quantile function for $\mathcal{N}(0,1)$.

Now, when $y \geq 0$,

$$y - \text{prox}_c(\rho_{l_1})(y) = y1_{y \leq c} + c1_{y \geq c} \; ,$$

and therefore

$$[y - \text{prox}_c(\rho_{l_1})(y)]^2 = y^2 1_{y \leq c} + c^2 1_{y \geq c} \; .$$

We hence have

$$\mathbf{E}\left((\hat{z}_\epsilon - \text{prox}_c(\rho_{l_1})(\hat{z}_\epsilon))^2\right) = s^2 \mathbf{E}\left(Z^2 1_{|Z| \leq \frac{c}{s}}\right) + c^2 P(|Z| \geq \frac{c}{s}) \; .$$

Recall that c/s is such that $P(|Z| \geq c/s) = 1 - \frac{p}{n}$. Also,

$$\mathbf{E}\left(Z^2 1_{|Z| \leq c/s}\right) = \frac{1}{\sqrt{2\pi}} \int_{-c/s}^{c/s} x^2 \exp(-x^2/2) dx = P\left(|Z| \leq \frac{c}{s}\right) - \frac{2}{\sqrt{2\pi}} \frac{c}{s} \exp(-c^2/(2s^2)) \; ,$$

by integration by parts, and we saw above that $c/s = \Phi^{-1}\left(\frac{1}{2}\left(1+\frac{p}{n}\right)\right)$ so we can substitute c/s by an explicit function of p/n.

Calling, for $t \in [0,1]$,

$$h\,(t) = t - \sqrt{\frac{2}{\pi}}\Phi^{-1}([1+t]/2)\exp(-[\Phi^{-1}([1+t]/2)]^2/2)\ ,$$

we have as interpretation of Equation (14),

$$\begin{aligned}\frac{p}{n}\mathbf{E}\left(\|\widehat{\beta}\|^2\right) &= s^2 h\left(\frac{p}{n}\right) + c^2\left(1-\frac{p}{n}\right)\ ,\\ &= s^2\left[h\left(\frac{p}{n}\right) + \left(1-\frac{p}{n}\right)\left(\Phi^{-1}\left[\frac{1}{2}\left(1+\frac{p}{n}\right)\right]\right)^2\right]\ .\end{aligned}$$

Since $\mathbf{E}\left(\|\widehat{\beta}\|^2\right) = s^2 - \sigma_\epsilon^2$, we can solve for s^2: calling, for $t \in [0,1]$,

$$\zeta(t) = 2\Phi^{-1}(t)\left(\varphi[\Phi^{-1}(t)] - \Phi^{-1}(t)(1-t)\right)\ ,$$

where φ is the standard normal density, we have after some algebra:

$$s^2 = \frac{\frac{p}{n}}{\zeta([1+\frac{p}{n}]/2)}\sigma_\epsilon^2$$

and

$$\boxed{\mathbf{E}\left(\|\widehat{\beta}_{l_1}\|^2\right) \simeq \frac{\frac{p}{n} - \zeta([1+\frac{p}{n}]/2)}{\zeta([1+\frac{p}{n}]/2)}\sigma_\epsilon^2\ .}$$

It is possible and not very hard to verify that $\zeta(t) \geq 0$ if $t \in [1/2, 1]$ and $2t - 1 - \zeta(t) \geq 0$ on the same interval, which insures that the prediction is positive.

We present some simulations below (Figures 1 and 2) to compare the results of our predictions and that of numerical experiments.

3.3.2 Further remarks on l_1-regression with symmetric error distribution

Our prox-function computations show that we need to solve for c such that

$$P(|\hat{z}_\epsilon| \geq c) = 1 - \frac{p}{n}\ ,$$

and then we have

$$\frac{p}{n}\mathbf{E}\left(\|\widehat{\beta}\|^2\right) = \mathbf{E}\left(\hat{z}_\epsilon^2 1_{|\hat{z}_\epsilon| \leq c}\right) + c^2(1-\frac{p}{n})\ .$$

Since $\mathbf{E}\left(\hat{z}_\epsilon^2 1_{|\hat{z}_\epsilon|\leq c}\right) = \mathbf{E}\left(\hat{z}_\epsilon^2\right) - \mathbf{E}\left(\hat{z}_\epsilon^2 1_{|\hat{z}_\epsilon|\geq c}\right)$, we see that

$$(1-\frac{p}{n})\mathbf{E}\left(\|\widehat{\beta}\|^2\right) = \mathbf{E}\left(\hat{z}_\epsilon^2 1_{|\hat{z}_\epsilon|\geq c}\right) - \sigma_\epsilon^2 - c^2(1-\frac{p}{n})\ .$$

Let us call $r = \|\widehat{\beta}\|$, which we now treat as an unknown constant. Let us call f_r the density of $\hat{z}_\epsilon$, which we assume exists. When ϵ_i are symmetric, so is $\hat{z}_\epsilon$. Hence,

$$\mathbf{E}\left(\hat{z}_\epsilon^2 1_{|\hat{z}_\epsilon|\geq c}\right) = 2\mathbf{E}\left(\hat{z}_\epsilon^2 1_{\hat{z}_\epsilon \geq c}\right) = 2\int_c^\infty y^2 f_r(y)dy\ .$$

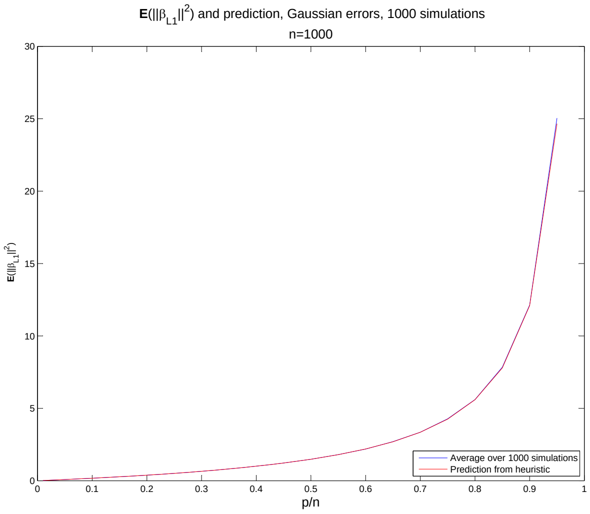

Figure 1: Prediction vs realized value of $\mathbf{E}\left(\|\widehat{\beta}\|^2\right)$, Gaussian errors, l_1 loss, 1000 simulations

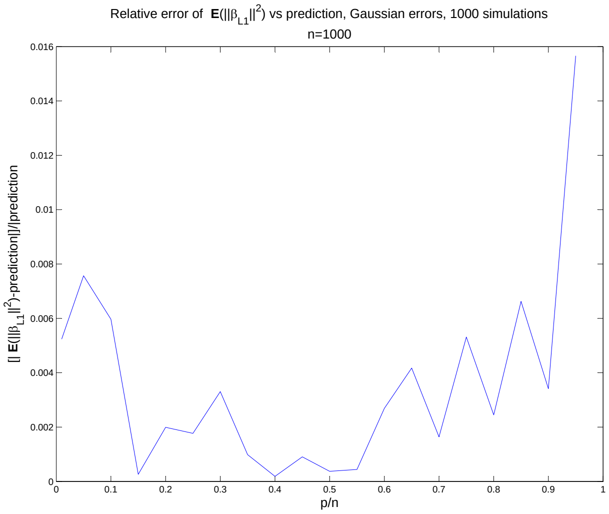

Figure 2: Relative errors: $\frac{|\mathbf{E}\left(\|\widehat{\beta}\|^2\right)-\text{prediction}|}{\text{prediction}}$, Gaussian errors, l_1 loss, 1000 simulations

We call $\bar{F}_r(t) = P(\hat{z}_\epsilon \geq t)$ and have $P(|\hat{z}_\epsilon| \geq c) = 2P(\hat{z}_\epsilon \geq c) = 2\bar{F}_r(c) = 1 - \frac{p}{n}$, by using symmetry of $\hat{z}_\epsilon$. Integrating by parts the previous display, we see that

$$\int_c^\infty y^2 f_r(y)dy = c^2\bar{F}_r(c) + 2\int_c^\infty x\bar{F}_r(x)dx = c^2\frac{1-\frac{p}{n}}{2} + 2\int_c^\infty x\bar{F}_r(x)dx \ .$$

Therefore,

$$\mathbf{E}\left(\hat{z}_\epsilon^2 1_{|\hat{z}_\epsilon|\geq c}\right) = c^2(1-\frac{p}{n}) + 4\int_c^\infty x\bar{F}_r(x)dx \ .$$

Of course, $c = \bar{F}_r^{-1}((1-\frac{p}{n})/2)$, so putting everything together we finally get that

$$\boxed{(1-\frac{p}{n})r^2 = 4\int_{\bar{F}_r^{-1}((1-\frac{p}{n})/2)}^\infty x\bar{F}_r(x)dx - \sigma_\epsilon^2} \tag{16}$$

We have to solve this equation in r for general error distributions. Analytically, it is a priori quite difficult (even establishing existence and uniqueness of the solution may not be simple).

A case of particular interest for l_1 regression is that of double-exponential errors, since we know that in that case it is more efficient in low-dimension to use l_1 rather than l_2 regression. (The l_1 objective function also corresponds to the log-likelihood of the data conditional on $\{X_i\}$ when $\{\epsilon_i\}_{i=1}^n$ are i.i.d double exponential.)

Case of double exponential errors

As we have seen the only problem is to compute the cdf of $rZ + \epsilon_i$. When ϵ_i are double exponential, if we call $F_r(t) = P(rZ + \epsilon_i \leq t)$, we have after a simple computation

$$F_r(t) = \Phi\left[\frac{t}{r}\right] + \frac{\exp(r^2/2)}{2}\left(\exp(t)\Phi\left[-\frac{t+r^2}{r}\right] - \exp(-t)\Phi\left[\frac{t-r^2}{r}\right]\right) \ .$$

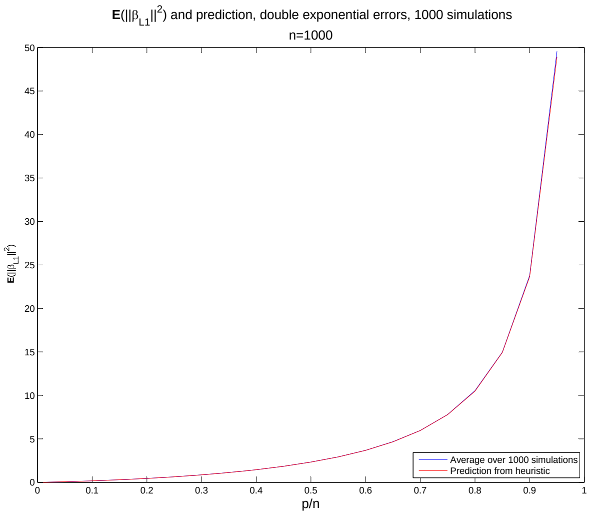

Figure 3: Prediction vs realized value of $\mathbf{E}\left(\|\widehat{\beta}\|^2\right)$, double exponential errors, l_1 loss, 1000 simulations.

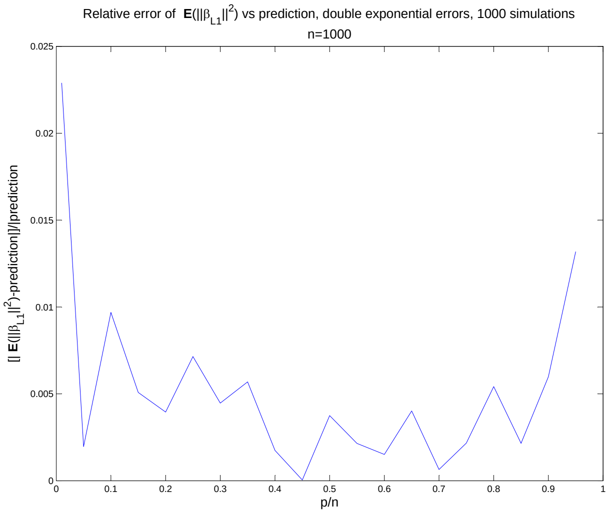

Figure 4: Relative errors: $\frac{|\mathbf{E}\left(\|\widehat{\beta}\|^2\right)-\text{prediction}|}{\text{prediction}}$, double exponential errors, l_1 loss, 1000 simulations.

Furthermore, when ϵ_i are double exponential, $\sigma_\epsilon^2 = 2$, and hence the equation we have to solve is

$$(1-\frac{p}{n})r^2 - 4\int_{\bar{F}_r^{-1}((1-\frac{p}{n})/2)}^{\infty} x\bar{F}_r(x)dx + 2 = 0 \ .$$

We found numerical solutions of this equation by doing a dichotomous search for its zeroes. Our numerical results follow. It should be noted that one needs to be careful with the numerical aspects of these questions: our experience was that naive and somewhat unrefined (or careless) implementations yielded poor results. Also, the numerics in the case p/n small become delicate and were a source of extra difficulty. It is not completely surprising since we are also potentially in that case in a different asymptotic regime. We refer the reader to Figures 3 and 4 for illustration.

A general remark about numerics

We note that a slightly better numerical implementation (especially for small p/n) might make use of the fact that $r^2 + \sigma_\epsilon^2 = 4\int_0^\infty x\bar{F}_r(x)dx$ (this is just a second moment computation) so that Equation (16) reads

$$\frac{p}{n}r^2 = 4\int_0^{\bar{F}_r^{-1}((1-\frac{p}{n})/2)} x\bar{F}_r(x)dx \ ,$$

and the integration bounds are now finite. However, our numerical illustrations come from a straight implementation of Equation (16).

3.3.3 Comparison between l_1-regression and ordinary least squares

A focus of our work is understanding how various estimators behave in high-dimension when the distribution of the errors changes and which loss function we should use for regression, if we happen to have some information about the errors. We now make these comparisons numerically and verify that our predictions remain accurate for the corresponding relative error measures.

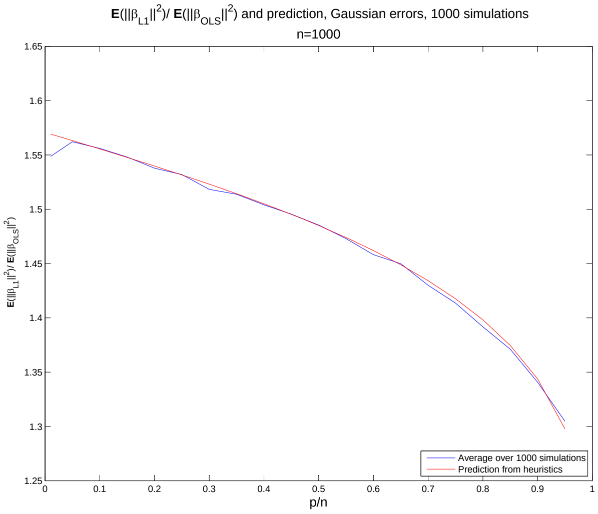

Figure 5: Prediction vs realized value of $\mathbf{E}\left(\|\widehat{\beta}_{l_1}\|^2\right)/\mathbf{E}\left(\|\widehat{\beta}_{OLS}\|^2\right)$, Gaussian errors. According to this measure, no matter the value of $\frac{p}{n}$, using ordinary least-squares yields better results than l_1-regression when the errors are Gaussian.

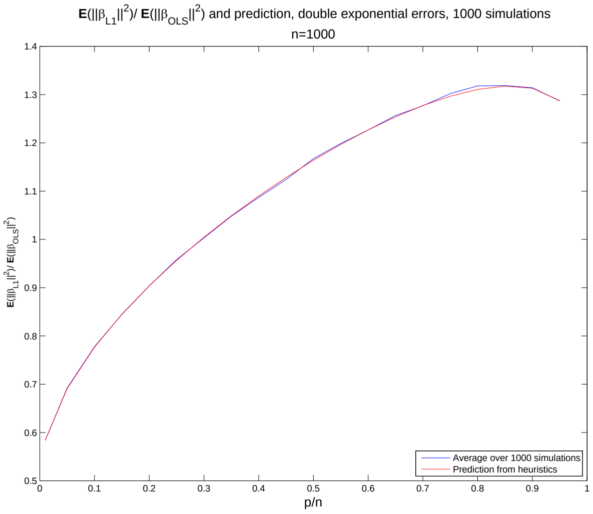

Figure 6: Prediction vs realized value of $\mathbf{E}\left(\|\widehat{\beta}_{l_1}\|^2\right)/\mathbf{E}\left(\|\widehat{\beta}_{OLS}\|^2\right)$, double exponential errors. Surprisingly, according to this measure, it becomes preferable to use ordinary least-squares rather than l_1-regression when the errors are double-exponential and $\frac{p}{n}$ is sufficiently large.

Figure 6 (p. 22) is particularly interesting as it shows that, even for double exponential errors, the performance of l_1 regression (supposedly adapted to the case of double exponential errors investigated there) becomes worse than that of ordinary least squares regression when p/n is sufficiently large (roughly $p/n > 0.3$). This is a surprising fact because it is well-known that when p/n is close to zero and the errors are double exponential, l_1-regression leads to estimate of $\widehat{\beta}$ that are more efficient than l_2 regression. Hence, classical low-dimensional intuition is upended in high-dimension.

On the other hand, Figure 5 shows that least-squares regression performs better than median regression when the errors are Gaussian, over the whole range of p/n.

3.4 Further examples

We present further examples of solutions of our system in Appendix A.

4 Robustness questions and extensions of the conjecture

We investigate two types of robustness questions for our proposed system of equations: we consider the impact of our distributional assumptions (distributional robustness) and that of the dimensionality assumptions. In the first case, we check that a meaningful perturbation of our assumptions yields a change in the system of equations one needs to solve. In the second case, we check that our system of equations allows us to recover classical results when p/n is small.

4.1 Impact of geometry

Because the conjecture relies heavily on the use of the concentration of measure phenomenon (for Gaussian random vectors) at a few key points, it is important to have an idea of the sensitivity of the conjecture to the implicit geometric assumptions that are made.

To make those more explicit, let us note that if X_i are i.i.d $\mathcal{N}(0, \mathrm{Id}_p)$ and A is a $p \times p$ deterministic symmetric matrix with $|||A|||_2 < C$ for some $C > 0$ and independent of p, we have

$$\sup_{i=1,\ldots,n} |\frac{1}{p} X_i' A X_i - \frac{\mathrm{trace}\,(A)}{p}| \to 0 \text{ in probability, when } p/n \to \kappa \in (0, \infty) \; .$$

This is an easy consequence of the concentration of measure phenomenon for 1-Lipschitz function of Gaussian random vectors. This can be shown through elementary moment computation here. For a more general view, we refer the reader to Ledoux (2001) or El Karoui (2009a) for specific interest in these questions. Taking $A = \mathrm{Id}_p$, we see that the previous remark implies that

$$\sup_{i=1,\ldots,n} |\frac{\|X_i\|^2}{p} - 1| \to 0 \text{ in probability} \; .$$

It is also easy to see that another choice of A implies that

$$\sup_{i \neq j} |\frac{X_i' X_j}{\|X_i\| \|X_j\|}| \to 0 \text{ in probability} \; .$$

In other words, our Gaussianity assumptions (and really concentration assumptions) amount to assuming - among many other things - that the data vectors live near a sphere (of radius $\sqrt{p}$) and are nearly orthogonal to one another (in fact, the same argument we made for a sphere can be made for the surface of many ellipsoids - for the sphere we picked $A = \mathrm{Id}_p$ but a different A would yield a different ellipsoid; the conditions on A for the concentration arguments to go through are very mild). As explained in the works referenced above, this is also true for Gaussian random vectors having covariance $\Sigma \neq \mathrm{Id}_p$, as long as, for instance, $|||\Sigma|||_2$ does not grow too fast with p, and more generally for a host of distributions having concentration properties for convex 1-Lipschitz functions.

4.1.1 Work with elliptical models

To depart from Gaussian-like geometry, we move from a Gaussian design matrix to an elliptical one. So we now assume that

$$X_i = \lambda_i \mathcal{X}_i$$

where λ_i is a random variable independent of $\mathcal{X}_i$ and $\mathcal{X}_i$ is $\mathcal{N}(0, \mathrm{Id}_p)$. We refer to Anderson (2003) for information on elliptical models and El Karoui (2009a) and El Karoui and Koesters (2011) for a longer discussion about the relevance of these models in high-dimensional statistics (see also Diaconis and Freedman (1984) for similar considerations in a different context). The elliptical model described here has a (genuinely) different geometry from the Gaussian model studied before because $\|X_i\|/\sqrt{p}$ is not close to a constant; indeed, it is close to $|\lambda_i|$, which is a random variable. In that sense, the model allows us to change the geometry of the dataset.

We also make the simple remark that if X is an $n \times p$ matrix whose rows are X_i, we have

$$X = D_\lambda \mathcal{X} \; ,$$

where D_λ is a diagonal matrix whose (i,i) entry is λ_i, and $\mathcal{X}$ is an $n\times p$ whose rows are $\mathcal{X}_i$. (We note that the argument that follows is essentially done conditional on D_λ and just assumes that D_λ is independent of $\mathcal{X}$. Therefore, beyond just the standard elliptical case where λ_i are i.i.d, the analysis applies to the case where D_λ is deterministic and also to the situation where its entries are not independent. What is crucial about D_λ is that some of the statistics that depend on the λ_i's below become asymptotically deterministic. For this to happen, it will generally suffice that the empirical distribution of the diagonal of D_λ converges weakly to a limit and that we do not have too many "outliers" in the λ_i's. We refer the interested reader to El Karoui (2010) were similar issues arise and are treated in (rigorous) detail.)

Let us recall the central results of our leave-one-out analyses. We ascertained and asserted that:

$$\begin{aligned}\widehat{\beta}-\widehat{\beta}_{(i)} &\simeq S_i^{-1}X_i\psi(R_i)\;,\\ \widehat{\beta}_p &\simeq \frac{\sum_{i=1}^n X_i(p)\psi(r_{i,[p]})}{\sum_{i=1}^n X_i^2(p)\psi'(r_{i,[p]})-u_p'\mathfrak{S}_p^{-1}u_p}\;,\\ u_p'\mathfrak{S}_p^{-1}u_p &= X(p)'AX(p),\ A=D^{1/2}P_VD^{1/2},\ P_V=D^{1/2}V(V'DV)^{-1}V'D^{1/2}\;,\end{aligned}$$

where D is a diagonal matrix with $D(i,i)=\psi'(r_{i,[p]})$. Multiplying by X_i' the first equation, and using the fact that it is elliptical, we get

$$R_i-\tilde{r}_{i,(i)}\simeq -\lambda_i^2c\psi(R_i)$$

where

$$c=\text{trace}\left((\sum_{i=1}^n\psi'(R_i)X_iX_i')^{-1}\right)\;.$$

Work on the denominator of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{p}}$

Using the rank-1 update for matrix inversion, we see that

$$P_V(i,i)=1-\frac{1}{1+\psi'(r_{i,[p]})V_i'[\mathfrak{S}_p(i)]^{-1}V_i}\simeq 1-\frac{1}{1+\lambda_i^2\psi'(r_{i,[p]})\text{trace}\left([\mathfrak{S}_p(i)]^{-1}\right)}\;.$$

Therefore, using the fact that P_V is a projection matrix, and the approximations trace $\left([\mathfrak{S}_p(i)]^{-1}\right)\simeq c$, and $r_{i,[p]}\simeq R_i$, the fact that trace $(P_V)=p-1$ now reads

$$\frac{1}{n}\sum_{i=1}^n\frac{1}{1+\lambda_i^2c\psi'(r_{i,[p]})}\simeq 1-\frac{p}{n}\;.$$

Using concentration properties of $\mathcal{X}_i$, conditional on $\{\lambda_i\}_{i=1}^n$, we have

$$\frac{1}{n}X(p)'AX(p)=\frac{1}{n}\mathcal{X}(p)'D_\lambda AD_\lambda\mathcal{X}(p)\simeq\frac{1}{n}\text{trace}\left(D_\lambda AD_\lambda\right)\simeq\frac{1}{n}\sum_{i=1}^n\lambda_i^2\psi'(r_{i,[p]})P_V(i,i)\;.$$

As noted above, we also have,

$$1-P_V(i,i)=\frac{1}{1+\psi'(r_{i,[p]})V_i'\mathfrak{S}(p,i)^{-1}V_i}\simeq\frac{1}{1+\lambda_i^2c\psi'(r_{i,[p]})}\;.$$

We also use the approximation

$$\frac{1}{n}\sum_{i=1}^n X_i^2(p)\psi'(r_{i,[p]})=\frac{1}{n}\sum_{i=1}^n\mathcal{X}_i^2(p)\lambda_i^2\psi'(r_{i,[p]})\simeq\frac{1}{n}\sum_{i=1}^n\lambda_i^2\psi'(r_{i,[p]})\;,$$

Hence, the denominator of $\widehat{\beta}_p$, divided by n is such that

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2(p)\psi'(r_{i,[p]}) - \frac{1}{n}u_p'\mathfrak{S}_p^{-1}u_p \simeq \frac{1}{n}\sum_{i=1}^{n}\lambda_i^2\psi'(r_{i,[p]})(1-P_V(i,i))$$

Replacing $1-P_V(i,i)$ by its value, we get that

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2(p)\psi'(r_{i,[p]}) - \frac{1}{n}u_p'\mathfrak{S}_p^{-1}u_p \simeq \frac{1}{n}\sum_{i=1}^{n}\frac{\lambda_i^2\psi'(r_{i,[p]})}{1+c\lambda_i^2\psi'(r_{i,[p]})} = \frac{1}{c}\left(1-\frac{1}{n}\sum_{i=1}^{n}\frac{1}{1+c\lambda_i^2\psi'(r_{i,[p]})}\right) \simeq \frac{1}{c}\frac{p}{n} .$$

So finally,

$$\widehat{\beta}_p \simeq \frac{\sum X_i(p)\psi(r_{i,[p]})/n}{\frac{p}{n}/c} \simeq c\frac{1}{p}\sum_{i=1}^{n}\lambda_i\psi(r_{i,[p]})\mathcal{X}_i(p) .$$

After elementary algebraic manipulations, we see (using again $\psi(r_{i,[p]}) \simeq \psi(R_i)$) that

$$\mathbf{E}\left(\|\widehat{\beta}\|^2\right) \simeq \frac{n}{p}\mathbf{E}\left(c^2\lambda_i^2\psi^2(R_i)\right) ,$$

where we used the notation $\mathbf{E}\left(c^2\lambda_i^2\psi^2(R_i)\right)$ as a shortcut for $\frac{1}{n}\sum_{i=1}^{n} c^2\lambda_i^2\psi^2(R_i)$, which we assume has a deterministic limit in our asymptotics.

Hence, with our notations from before, we see that the system to solve in the elliptical case is the following: with $\tilde{r}_{i,[p]} = \epsilon_i - \lambda_i\|\widehat{\beta}\|\mathcal{N}(0,1)$, and $g_{c\lambda_i^2}(x) = x + c\lambda_i^2\psi(x)$, we have

$$R_i = \text{prox}_{c\lambda_i^2}(\rho)(\tilde{r}_{i,(i)}) \text{ , and } c\lambda_i\psi(R_i) = \frac{\tilde{r}_{i,(i)} - \text{prox}_{c\lambda_i^2}(\rho)(\tilde{r}_{i,(i)})}{\lambda_i} .$$

System formulation for elliptical data
The first equation of the system (KeyProxSystem) becomes

$$\mathbf{E}\left([\text{prox}_{c\lambda_i^2}(\rho)]'(\tilde{r}_{i,(i)})\right) = 1 - \frac{p}{n} ,$$

where the expectation also has to be taken with respect to the distribution of λ_i.

The second equation in this system (characterizing $\|\widehat{\beta}\|^2$) is now

$$\frac{p}{n}\mathbf{E}\left(\|\widehat{\beta}\|^2\right) \simeq \mathbf{E}\left(\frac{[\tilde{r}_{i,(i)} - \text{prox}_{c\lambda_i^2}(\rho)(\tilde{r}_{i,(i)})]^2}{\lambda_i^2}\right) .$$

The conclusion is that the system of equations we have to solve is sensitive to the geometry of the data, since it depends on the distribution of λ_i. We note that in this setting we still have $\|\widehat{\beta}\|^2 \simeq \mathbf{E}\left(\|\widehat{\beta}\|^2\right)$.

We also note that when $\lambda_i^2 = 1$, we recover the system we postulated in the Gaussian predictor case.

4.1.2 Extensions of the conjecture: heteroskedasticity and weighted robust regression

Heteroskedasticity A look at our arguments reveals that we do not make strong use at any point of the assumption that the ϵ_i's are i.i.d. Our effective assumptions are more concerned with the fact that a few points do not drive the behavior of the solution. When the ϵ_i's are not i.i.d but have reasonably similar distributions (something that will be made clearer when we give a mathematical proof of all this) the system becomes, when $\beta_0 = 0$ and $\text{cov}\,(X_i) = \text{Id}_p$,

$$\begin{cases} \mathbf{E}\left([\text{prox}_c(\rho)]'(\tilde{r}_{i,(i)})\right) & = 1 - \frac{p}{n}\ , \\ \frac{p}{n}\mathbf{E}\left(\|\widehat{\beta}\|^2\right) & \simeq \mathbf{E}\left([\tilde{r}_{i,(i)} - \text{prox}_c(\rho)(\tilde{r}_{i,(i)})]^2\right)\ , \end{cases}$$

where $\tilde{r}_{i,(i)} = \|\widehat{\beta}\| Z_i + \epsilon_i$, Z_i's are i.i.d $\mathcal{N}(0,1)$ and the expectation has to take into account the fact that there is effectively a prior on the distributions of ϵ_i's.

Weighted robust regression One may also ask what happens when we solve

$$\widehat{\beta} = \text{argmin}_\beta \sum_{i=1}^n w_i \rho(\epsilon_i - X_i'\beta)\ ,\ w_i \geq 0\ ,$$

and the w_i's are not all equal. This is very relevant for a proper treatment of heteroskedasticity issues, where one might be inclined to resort to this sort of modifications of the original question. In this brief discussion, we only consider the case where $\rho(tx) = t^{1/\alpha}\rho(x)$, for $t > 0$. This is clearly relevant to the examples we have looked at. In this case, we can rewrite the objective function as

$$\sum_{i=1}^n w_i \rho(\epsilon_i - X_i'\beta) = \sum_{i=1}^n \rho(w_i^\alpha \epsilon_i - w_i^\alpha X_i'\beta)\ .$$

Hence, if w_i were picked in such a way that $w_i^\alpha \epsilon_i$ are i.i.d and X_i were $\mathcal{N}(0, \text{Id}_p)$, the problem would effectively be - as the second formulation makes clear - a problem with homoskedastic errors and elliptical-like predictors.

In general, it is clear that the weighted robust regression problems (with ρ such that $\rho(tx) = t^{1/\alpha}\rho(x), t > 0$) is effectively going to be a standard robust regression problem with heteroskedastic errors and elliptical (or elliptical-like) predictors. The system we conjectured in the elliptical case should be able to handle some amount of heteroskedasticity by following the same guidelines we just outlined, i.e the expectation should be interpreted as an expectation over the Gaussian component of $\tilde{r}_{i,(i)}$, a prior on λ_i's and a prior on ϵ_i's.

Beyond the weighted case Naturally, the next (mathematical) step would be to consider the problem

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

In this case, it is natural to think that we have a prior on the loss functions ρ's. Looking at our derivation, in the case of i.i.d errors and $\mathcal{N}(0, \text{Id}_p)$ predictors, a natural conjecture is that $\|\widehat{\beta}\|$ will be asymptotically deterministic (under some conditions on this prior on functions) and will satisfy

$$\begin{cases} \mathbf{E}\left([\text{prox}_c(\rho_i)]'(\tilde{r}_{i,(i)})\right) & = 1 - \frac{p}{n}\ , \\ \frac{p}{n}\mathbf{E}\left(\|\widehat{\beta}\|^2\right) & \simeq \mathbf{E}\left([\tilde{r}_{i,(i)} - \text{prox}_c(\rho_i)(\tilde{r}_{i,(i)})]^2\right)\ , \end{cases}$$

where now the expectation is also over the prior on ρ_i's. (This prior would have to be "reasonable" in the sense that a few ρ_i's do not drive the whole regression problem).

4.2 Another example of sensitivity to our assumptions

It is important to investigate the sensitivity of our results in several directions. We assume in this subsection that the vectors $\{X_i\}_{i=1}^n$ are i.i.d and take value $\{e_k\}_{k=1}^p$ with equal probability, where $\{e_k\}_{k=1}^p$ are the canonical basis vectors (i.e $e_k(j) = \delta_{j,k}$). We also assume for the sake of simplicity that ψ changes sign on $\mathbb{R}$ and that it is continuous. In this case, the (vector) equation defining $\widehat{\beta}$ becomes

$$\sum_{i=1}^n X_i \psi(\epsilon_i - X_i'\widehat{\beta}) = \sum_{k=1}^p \sum_{i:X_i=e_k} e_k \psi(\epsilon_i - \widehat{\beta}_k) = 0 \;.$$

Hence the vector equation separates into p scalar equations and $\widehat{\beta}$ is defined, coordinate by coordinate as

$$\forall k,\, 1 \leq k \leq p : \sum_{i:X_i=e_k} \psi(\epsilon_i - \widehat{\beta}_k) = 0 \;.$$

The number of terms in each of these equations is in expectation n/p. So it is clear that the marginal distribution of $\widehat{\beta}_k$ is not normal in general: indeed we can find it by conditioning on $N_k = \text{Card}\{i : X_i = e_k\}$ (a binomial$(n, 1/p)$ random variable, hence an essentially $Poisson(n/p)$ random variable in our asymptotics where p/n remains bounded away from 0) and solving, for $\{\epsilon_j\}_{j=1}^{N_k}$, independent of N_k:

$$\sum_{1\leq j\leq N_k} \psi(\epsilon_j - \widehat{\beta}_k) = 0 \;.$$

This will in general not be normal; even in the case of least-squares regression, $\widehat{\beta}_k$ is the mean the ϵ_j's, which is not in general normal (recall that N_k is bounded in probability here), when ϵ_j's are not normal.

This example could naturally be easily extended to the case where ρ has a subdifferential. We leave this to the interested reader.

4.3 Connection with classical statistical theory ($\kappa \simeq 0$)

Recall the system (Key Functional System). If $g_c = x + c\psi(x)$,

$$\begin{cases} \mathbf{E}\left(\left(g_c^{-1}\right)'(\tilde{r}_{i,(i)})\right) & \simeq 1 - \frac{p}{n} \;, \\ \frac{p}{n}\mathbf{E}\left(\|\widehat{\beta}\|^2\right) & \simeq \mathbf{E}\left((\tilde{r}_{i,(i)} - g_c^{-1}(\tilde{r}_{i,(i)}))^2\right) \;. \end{cases}$$

When p/n is very small we are back in the classical case. We therefore expect that $\mathbf{E}\left(\|\widehat{\beta}\|^2\right) \simeq 0$ (we have an unbiased estimator for 0 that is consistent) and the first equation reads in its expanded form

$$\frac{1}{n}\sum_{i=1}^n \frac{1}{1 + c\psi'(R_i)} = 1 - \frac{p}{n} \;,$$

where $R_i = \epsilon_i - X_i'\widehat{\beta}$. Hence, we also expect that $c \simeq 0$.

It is then natural to guess that $g_c^{-1}(y) = y - \eta f(y)$, where η is small and f unknown. Plugging this into $g_c(x) = x + c\psi(x)$, we see that

$$x = g_c(g_c^{-1}(x)) = x - \eta f(x) + c\psi(x - \eta f(x)) \simeq x - \eta f(x) + c\psi(x) \;,$$

for smooth ψ. So it is natural to conjecture that, since c is expected to be small,

$$\eta f(x) = c\psi(x)$$

and

$$g_c^{-1}(y) \simeq y - c\psi(y) \;.$$

We now take these relations as given and proceed to solve the system. Since $(g_c^{-1})'(y) \simeq 1 - c\psi'(y)$, we see that

$$c\mathbf{E}\left(\psi'(\tilde{r}_{i,(i)})\right) = \frac{p}{n} \; .$$

Now $\tilde{r}_{i,(i)} = \mathcal{N}(0, \|\widehat{\beta}\|^2) + \epsilon_i \simeq \epsilon_i$, since $\|\widehat{\beta}\|^2$ is about zero. So we are led to believe that

$$c = \frac{p}{n} \frac{1}{\mathbf{E}\left(\psi'(\epsilon)\right)} \; ,$$

which is indeed small in general. Now $y - g_c^{-1}(y) = c\psi(y)$. So the second equation gives

$$\frac{p}{n}\mathbf{E}\left(\|\widehat{\beta}\|^2\right) = c^2\mathbf{E}\left((\psi(\tilde{r}_{i,(i)}))^2\right) \simeq \left(\frac{p}{n}\right)^2 \frac{\mathbf{E}\left(\psi^2(\tilde{r}_{i,(i)})\right)}{[\mathbf{E}\left(\psi'(\epsilon)\right)]^2} \simeq \left(\frac{p}{n}\right)^2 \frac{\mathbf{E}\left(\psi^2(\epsilon)\right)}{[\mathbf{E}\left(\psi'(\epsilon)\right)]^2} \; .$$

Rewriting it, we get

$$\mathbf{E}\left(\|\widehat{\beta}\|^2\right) = \frac{p}{n} \frac{\mathbf{E}\left(\psi^2(\epsilon)\right)}{[\mathbf{E}\left(\psi'(\epsilon)\right)]^2} \; .$$

If one does classical perturbation analysis on this problem for very small p/n, that is the solution one gets.

This suggests that the system (Key Functional System) (or its more general version (KeyProxSystem)) is in some sense universal in terms of dimension since we can also apply it to the case p/n small and get the right result.

We should also add that in low dimensions, the geometric issues we pointed out (when questioning the robustness of the system (KeyProxSystem)) above are irrelevant since we are basically able to estimate the true regression vector β_0. Hence $\widehat{\beta} - \beta_0 \simeq 0$ - therefore the distribution of $\tilde{r}$ is always close to that of the noise.

4.4 What is settled rigorously?

As far as we know, the least squares case is settled rigorously when the predictors are Gaussian, by relying on classical Wishart theory (Anderson (2003)). Dealing with elliptical distributions in the least-squares case can also be done but requires random matrix theory - in fact the results of El Karoui (2010) and El Karoui (2009b) can be used to do so. It is also possible to move away from the Gaussianity assumption on the predictors in the case of ridge regression and this is done in El Karoui and Koesters (2011) under mild concentration assumptions for the predictors. The part of the current paper on invariance is fully rigorous. We are currently working on making the rest of the paper mathematically rigorous. However, this work, which is very near completion (indeed we can rigorously justify many of our approximations), is not included in the paper because it is long and would put us largely above the page limit requirements for this journal.

5 Conclusion

We have presented a double leave-one-out approach to understanding the behavior of robust regression estimators in high-dimension. Somewhat surprisingly, our results point to the fact that the concentration of measure phenomenon is driving this behavior. This is surprising because it suggests that these problems have - after all - much in common with questions in high-dimensional random matrix theory. Indeed, this is confirmed by our derivations, but was quite unexpected, especially for problems like median regression and more generally quantile regression, which have certain selection features built into them (see Koenker (2005)), and are in this regard quite different a priori from standard random matrix problems.

We have given a detailed heuristic (grounded in theory) for justifying the a priori non trivial conjecture (see p.14) summarized in the System (KeyProxSystem). The conjecture gives good results even for non-smooth loss functions, as the agreement between our l_1 simulations and predictions indicates. This is also quite surprising at this point given how we arrived at the system. It should also be noted that the conjecture gives information not only about the behavior of $\widehat{\beta}$, the estimated regression vector, but also about the residuals, which have in general a very complicated distribution.

It should be noted that numerically the System (KeyProxSystem) is, in general, quite hard to solve, even when the robust regression problem is not. So it is possible that if this difficult system turns out to have another interpretation in applied mathematics, the link we made with robust regression might help in solving it efficiently.

Our derivation and simulations point to the fact that great care needs to be applied when thinking about the effect of using robust regression estimators in high-dimension. We have for instance seen that when the errors are double exponential, using median regression instead of least-squares regression might lead to inefficiencies when p/n is sufficiently large (see Figure 6); it is however well-known that in the classical setting (i.e p/n small) median regression outperforms ordinary least-squares for these errors. But there is no universal rule: for Gaussian errors, using ordinary least squares is more efficient than using median regression, as we expect from classical theory. This leads us to believe that classical intuition and arguments might have to be taken with a bit of suspicion in the high-dimensional setting of interest here - the two situations have little to do with one another.

Our analysis and conjecture reveals the very complicated interaction between loss function and distribution of the errors in determining the behavior of the solution - something that - as far as we know - was not understood so far and is very central to improving our understanding of statistics in high-dimension. In that respect, it should be noted that generally the whole distribution of the errors matter, not only simple statistics like the variance. Only in the case of least-squares regression does it seem that the variance is the only characteristic of the errors that is important.

Our work also yields quite striking formulas, such as the explicit one for median regression with Gaussian errors, and reveal the key role of the prox function in this problem. However, a deeper connection with convex optimization theory (upon which we stumbled) needs to be drawn.

Using the approach presented here we are able to conjecture the behavior of $\widehat{\beta}$ when we add a penalization term (on β) to the problem, for much richer geometries of the design matrix and for a variety of related problems involving heteroskedasticity and reweighting. We do not present all of them here to avoid obscuring the main arguments.

Our presentation is heuristic here, but we are currently working on a fully rigorous mathematical analysis. We hope however to have drawn the attention of the reader to the gist of the various phenomena at stake and in particular would like to point out again the key role of measure concentration in our analysis. A benefit of this analysis is that it immediately shows its limitations and in particular the sensitivity of the results to dataset (Euclidian) geometry. Finally, we note that as p/n tends to zero and we return to the classical setting, our predictions yield well-known classical results.

APPENDIX

A Further examples of solutions of our system of equations

A-1 Huber loss functions

In this situation, we have

$$\rho_\delta(x) = \begin{cases} \frac{x^2}{2} & \text{if } |x| \leq \delta \ , \\ \delta(|x| - \frac{\delta}{2}) & \text{if } |x| \geq \delta \ . \end{cases}$$

Hence, we have, for $x \geq 0$, $\psi_\delta(x) = x 1_{x \leq \delta} + \delta 1_{x \geq \delta}$. It is therefore easy to verify that

$$\text{prox}_c(\rho_\delta)(y) = \begin{cases} \frac{y}{1+c} & \text{if } y \in (-(1+c)\delta, (1+c)\delta) \ , \\ y - c\delta & \text{if } y \geq (1+c)\delta \ , \\ y + c\delta & \text{if } y \leq -(1+c)\delta \ . \end{cases}$$

After some elementary manipulations, we see that the first equation of system (KeyProxSystem) reads

$$\frac{c}{1+c} P(\hat{z}_\epsilon \in [-(1+c)\delta, (1+c)\delta]) = \frac{p}{n} \ .$$

The second equation of system (KeyProxSystem) reads

$$\frac{p}{n} \mathbf{E}\left(\|\widehat{\beta}\|^2\right) = \left(\frac{c}{1+c}\right)^2 \mathbf{E}\left(\hat{z}_\epsilon^2 1_{\hat{z}_\epsilon \in [-(1+c)\delta, (1+c)\delta]}\right) + c^2\delta^2 (1 - \frac{p}{n}\frac{1+c}{c}) \ .$$

Gaussian case: As usual, calling $s = \sqrt{\|\widehat{\beta}\|^2 + \sigma_\epsilon^2}$, and $\alpha = (1+c)\delta/s$, the previous equation can be rewritten as

$$\frac{p}{n}(s^2 - \sigma_\epsilon^2) = \left(\frac{c}{1+c}\right)^2 s^2 \left[2\Phi(\alpha) - 1 - 2\alpha\phi(\alpha)\right] + 2c^2\delta^2(1 - \Phi(\alpha)) \ ,$$

using the fact that if $Z \sim \mathcal{N}(0,1)$, $\mathbf{E}\left(Z^2 1_{\{Z \in [-\alpha,\alpha]\}}\right) = P(Z \in [-\alpha, \alpha]) - 2\alpha\phi(\alpha)$. On the other hand, the first equation in the system reads $2\Phi(\alpha) - 1 = (p/n)(1+c)/c$.

Numerically, we can try to solve our system of two equations by collapsing it into a single equation in α. As a matter of fact, we have, using the first equation of the system (KeyProxSystem),

$$c = \frac{\frac{p}{n}}{2\Phi(\alpha) - 1 - \frac{p}{n}} \ , \text{ and } s = \frac{2\Phi(\alpha) - 1}{2\Phi(\alpha) - 1 - \frac{p}{n}} \frac{\delta}{\alpha} \ .$$

We can also rewrite the second equation as

$$\frac{p}{n}(s^2 - \sigma_\epsilon^2) = c^2\delta^2 \left[\frac{1}{\alpha^2}\left[2\Phi(\alpha) - 1 - 2\alpha\phi(\alpha)\right] + 2(1 - \Phi(\alpha))\right] \ .$$

After we replace c and s by their values in terms of α, this becomes an equation in α (δ and σ_ϵ^2 are of course given) which we can attempt to solve numerically

A-2 Quantile regression

Quantile regression is a very popular technique used in a variety of fields and primarily in Econometrics. We refer the reader to Koenker (2005) for a detailed introduction and many references.

The loss function ρ_τ used in this method is of the form

$$\rho_\tau(x) = x(\tau - 1_{x \leq 0}), \text{ where } \tau \in (0,1) ,$$

which is not symmetric in general. (Of course, when $\tau = 1/2$ we are back in the case of median regression, where $\rho(x) = |x|$.)

Elementary computations show that the prox function has the form

$$\text{prox}_c(\rho_\tau(y)) = \begin{cases} y + c(1-\tau) & \text{if } y \leq -c(1-\tau) \\ 0 & \text{if } y \in [-c(1-\tau), c\tau] \\ y - c\tau & \text{if } y \geq c\tau \end{cases}$$

and its derivative can be written as (ignoring the problems at $-c(1-\tau)$ and $c\tau$)

$$(\text{prox}_c(\rho_\tau)'(y) = 1 - 1_{y \in [-c(1-\tau), c\tau]} .$$

Hence, the first equation of the system (KeyProxSystem) can be reformulated as

$$P(\hat{z}_\epsilon \in [-c(1-\tau), c\tau]) = \frac{p}{n} ,$$

the definition of $\hat{z}_\epsilon$ taking care of the (differentiability) problems at $-c(1-\tau)$ and $c\tau$.

Gaussian error case In this case, we are able to pursue the matters a bit further. As usual, we call $s > 0$ the standard deviation of $\hat{z}_\epsilon$, i.e $s = \sqrt{\|\widehat{\beta}\|^2 + \sigma_\epsilon^2}$. Calling $\upsilon_\tau = c/s(>0)$, we see that $\upsilon_\tau(>0)$ is defined by the equation

$$\frac{p}{n} = P(Z \in [-\upsilon_\tau(1-\tau), \upsilon_\tau\tau]) ,$$

where $Z \sim \mathcal{N}(0,1)$. At τ given, this equation can be easily solved numerically for υ_τ.

On the other hand, the second equation of the system (KeyProxSystem) reads

$$\frac{p}{n}\left(s^2 - \sigma_\epsilon^2\right) = c^2(1-\tau)^2 P(Z \leq -\upsilon_\tau(1-\tau)) + c^2\tau^2 P(Z \geq \upsilon_\tau\tau) + s^2 \mathbf{E}\left(Z^2 1_{Z \in [-\upsilon_\tau(1-\tau), \upsilon_\tau\tau]}\right) .$$

Recalling that when $Z \sim \mathcal{N}(0,1)$, $\mathbf{E}\left(Z^2 1_{Z \in (a,b)}\right) = \Phi(b) - \Phi(a) + a\phi(a) - b\phi(b) = \Xi(a,b)$, where ϕ is the standard normal density and Φ is the standard normal cdf, we can finally rewrite the previous equation as

$$\frac{p}{n}\left(s^2 - \sigma_\epsilon^2\right) = s^2\left[\upsilon_\tau^2(1-\tau)^2\Phi(-\upsilon_\tau(1-\tau)) + \upsilon_\tau^2\tau^2\Phi(-\upsilon_\tau\tau) + \Xi(-\upsilon_\tau(1-\tau), \upsilon_\tau\tau)\right] .$$

Of course, almost by definition, $p/n = \Phi(\upsilon_\tau\tau) - \Phi(-\upsilon_\tau(1-\tau))$. Therefore, calling $\Theta(x) = -x\phi(x) - x^2\Phi(x)$, we finally have the conjecture

$$\begin{aligned} s^2 &\simeq \frac{p/n}{\Theta(-\upsilon_\tau(1-\tau)) + \Theta(-\upsilon_\tau\tau)}\sigma_\epsilon^2 , \\ \|\widehat{\beta}\|^2 &\simeq s^2 - \sigma_\epsilon^2 . \end{aligned}$$

A little work is needed to investigate whether the conjecture makes sense - and whether it does at least give meaningful signs to the quantities of interest. It is easy to verify (using e.g integration by parts) that for $x \leq 0$, $\Theta(x) \geq 0$. Since we also know that by definition, we should have $s^2 \geq \sigma_\epsilon^2$, we need to check that the conjectured ratio for s^2/σ_ϵ^2 is greater than 1.

Calling $b = \upsilon_\tau(1-\tau)$ and $a = -\tau\upsilon_\tau$, we see that this amounts to showing that

$$\Theta(-b) + \Theta(a) \leq \Phi(b) - \Phi(a) \ .$$

Calling Δ the function such that $\Delta(x) = (x^2-1)\Phi(x) + x\phi(x)$, we see that this is equivalent to showing that

$$1 + \Delta(-b) + \Delta(a) \geq 0 \ .$$

Elementary computations show that $\Delta'(x) = 2x\Phi(x)$ which is naturally negative when $x \leq 0$. Hence, using the fact that $-b \leq 0$ and $a \leq 0$, we see that $1 + \Delta(-b) + \Delta(a) \geq 1 + 2\Delta(0)$. Since $\Delta(0) = -\Phi(0) = -1/2$, we have shown that indeed

$$1 + \Delta(-b) + \Delta(a) \geq 0 \ ,$$

and therefore,

$$\frac{p/n}{\Theta(-\upsilon_\tau(1-\tau)) + \Theta(-\upsilon_\tau\tau)} \geq 1 \ .$$

A-3 Case $\rho(x) = |x|^3/3$; l_3 loss

The case of l_3 loss is perhaps not necessarily very relevant from a robustness standpoint (l_3 gives significant weight to outliers) but it is one of the rare cases where our equations can be solved quite explicitly, so we investigate it. It is also interesting from our standpoint of trying to contribute to a general theory of M-estimation.

Here $\rho(x) = |x|^3/3$ and hence $\psi(x) = \text{sign}(x)x^2$. Therefore,

$$g_c(x) = x + c\,\text{sign}(x)\,x^2 \ .$$

For $y \geq 0$, after some algebra we get ($c > 0$)

$$\text{prox}_c(\rho)(y) = \frac{-1 + \sqrt{1+4cy}}{2c} \ .$$

As usual $\text{prox}_c(\rho)(-y) = -\text{prox}_c(\rho)(y)$, so that

$$\forall y \, , \ \text{prox}_c(\rho)(y) = \text{sign}(y)\,\frac{-1 + \sqrt{1+4c|y|}}{2c} \ ,$$

and

$$(\text{prox}_c(\rho))'(y) = \frac{1}{\sqrt{1+4c|y|}} \ .$$

Gaussian errors case: call $\nu = cs$, where $s^2 = \|\widehat{\beta}\|^2 + \sigma_\epsilon^2$. Recall that $\hat{z}_\epsilon \sim sZ$, where Z is $\mathcal{N}(0,1)$, because $\epsilon_i \sim \mathcal{N}(0,\sigma_\epsilon^2)$. Interpreting Equation (13) in this context, we get that ν satisfies:

$$\mathbf{E}\left(\frac{1}{\sqrt{1+4\nu|Z|}}\right) = 1 - \frac{p}{n} \ .$$

So at $\frac{p}{n}$ given, ν can be found numerically.

We turn to Equation (14). In the case of Gaussian errors, $\hat{z}_\epsilon$ is $\mathcal{N}(0, s^2)$. So we can use integration by parts to say that

$$\mathbf{E}\left(\hat{z}_\epsilon \text{prox}_c(\rho)(\hat{z}_\epsilon)\right) = s^2 \mathbf{E}\left((\text{prox}_c(\rho))'(\hat{z}_\epsilon)\right) = s^2\left(1 - \frac{p}{n}\right) .$$

We conclude that in the case of Gaussian errors, Equation (14) can be interpreted as

$$\frac{p}{n}(s^2 - \sigma_\epsilon^2) = s^2 - 2(1 - \frac{p}{n})s^2 + \mathbf{E}\left((\text{prox}_c(\rho)(sZ))^2\right) .$$

In the present case we have

$$(\text{prox}_c(\rho)(y))^2 = \frac{1}{4c^2}\left(2 + 4c|y| - 2\sqrt{1 + 4c|y|}\right) .$$

Taking expectation, we get, using the fact that $\hat{z}_\epsilon \sim sZ$, and the notation $cs = \nu$,

$$\begin{aligned}\mathbf{E}\left((\text{prox}_c(\rho)(\hat{z}_\epsilon))^2\right) &= \frac{1}{2c^2}\left(1 + 2\nu\sqrt{\frac{2}{\pi}} - \mathbf{E}\left(\sqrt{1 + 4\nu|Z|}\right)\right) ,\\ &= \frac{s^2}{2\nu^2}\left(1 + 2\nu\sqrt{\frac{2}{\pi}} - \mathbf{E}\left(\sqrt{1 + 4\nu|Z|}\right)\right) .\end{aligned}$$

So finally, we can reinterpret Equation (14) as saying that

$$\boxed{\frac{p}{n}(s^2 - \sigma_\epsilon^2) = s^2\left[(2\frac{p}{n} - 1) + \frac{1}{2\nu^2}\left(1 + 2\nu\sqrt{\frac{2}{\pi}} - \mathbf{E}\left(\sqrt{1 + 4\nu|Z|}\right)\right)\right] ,}$$

where ν is found by solving the equation mentioned above (therefore ν is implicitly a function of p/n). Calling

$$\gamma(\frac{p}{n}) = \left[(2\frac{p}{n} - 1) + \frac{1}{2\nu^2}\left(1 + 2\nu\sqrt{\frac{2}{\pi}} - \mathbf{E}\left(\sqrt{1 + 4\nu|Z|}\right)\right)\right] ,$$

we get

$$s^2 = \frac{\frac{p}{n}}{\frac{p}{n} - \gamma(\frac{p}{n})}\sigma_\epsilon^2 ,$$

and

$$\boxed{\mathbf{E}\left(\|\widehat{\beta}_{l_3}\|^2\right) \simeq \frac{\gamma(\frac{p}{n})}{\frac{p}{n} - \gamma(\frac{p}{n})}\sigma_\epsilon^2 .}$$

We have checked these predictions against our simulations (for a few values of p/n) and got relative accuracy of around 1% when n was of order 1000.

A-4 Case $\rho(x) = |x|^{1.5}/1.5$

We present this case as it is also one where we can be reasonably explicit about the behavior of $\|\widehat{\beta}\|$. Here we have

$$\psi(x) = \text{sign}(x)\sqrt{x} .$$

Therefore, for $y \geq 0$, we have, after some elementary algebra

$$g_c^{-1}(y) = \frac{c^2}{4}\left(2 + \frac{4}{c^2}y - 2\sqrt{1 + \frac{4}{c^2}y}\right) .$$

We can deduce $g_c^{-1}(y)$ for $y \leq 0$ by symmetry, and we have

$$g_c^{-1}(y) = \text{sign}(y) g_c^{-1}(|y|) \ .$$

We can now compute the derivative of this function and the first equation of system (KeyProxSystem) becomes

$$\mathbf{E}\left(\frac{1}{\sqrt{1+\frac{4}{c^2}|\tilde{r}_{i,(i)}|}}\right) = \frac{p}{n} \ .$$

On the other hand, $[y - g_c^{-1}(y)]^2 = \frac{c^4}{4}\left(1-\sqrt{1+\frac{4}{c^2}|y|}\right)^2$, so the second equation reads

$$\frac{p}{n}\mathbf{E}\left(\|\widehat{\beta}\|^2\right) = \frac{c^4}{4}\mathbf{E}\left(\left(1-\sqrt{1+\frac{4}{c^2}|\tilde{r}_{i,(i)}|}\right)^2\right) \ .$$

Gaussian error computations
As usual in the case of Gaussian errors we have $\tilde{r}_{i,(i)} = sZ$, $Z \sim \mathcal{N}(0,1)$, where $s = \|\widehat{\beta}\|^2 + \sigma_\epsilon^2$. If we call $\nu = 4s/c^2$, the first equation of the system now reads

$$\mathbf{E}\left(\frac{1}{\sqrt{1+\nu|Z|}}\right) = \frac{p}{n} \ .$$

This can be solved numerically to find ν (which clearly depends on p/n). Once ν is known, the second equation can be turned into an equation in s only:

$$\frac{p}{n}(s^2 - \sigma_\epsilon^2) = \frac{4s^2}{\nu^2}\mathbf{E}\left((1-\sqrt{1+\nu|Z|})^2\right) \ .$$

We can then finally conclude that

$$s^2 = \frac{\frac{p}{n}}{\frac{p}{n} - \frac{4}{\nu^2}\mathbf{E}\left((1-\sqrt{1+\nu|Z|})^2\right)}\sigma_\epsilon^2 \ .$$

B Basic facts on the prox function

B-1 Prox functions and dealing with non-differentiable ρ's

Since we have been doing formal computations to derive our heuristics, we have of course taken some liberties with non-differientable ρ among other things. In particular, if $\rho(x) = |x|$, a priori ψ (" $= \rho'$") is multivalued at 0, since it is a subdifferential. However, if $\partial\rho$ is the subdifferential of a closed proper convex function ρ (see e.g Rockafellar (1997) for definitions, if needed), it is well-known (since Moreau (1965)) that

$$(\text{Id} + t\partial\rho)^{-1}$$

is a single valued function, a remarkable fact.

Indeed, if

$$\rho_t(y) = \min_u \rho(u) + \frac{1}{2t}(u-y)^2 \text{ and}$$
$$\text{prox}_t(\rho)(y) = \text{argmin}_u \left\{\rho(u) + \frac{1}{2t}(u-y)^2\right\} \ ,$$

$\text{prox}_t(\rho)(y)$ is the unique minimizer of $\rho(u)+\frac{1}{2t}(u-y)^2$ at y given (for this we need ρ to be a closed proper convex function), and very importantly

$$\text{prox}_t(\rho)(y)=(\text{Id}+t\partial\rho)^{-1}(y)\ .$$

See Moreau (1965), Proposition 6.a, p. 283. (Of course the prox function can be defined for functions of vector arguments, but since we do not need this extension in this paper, we do not present it.) The function $\rho_t(y)$ is known in optimization as a Moreau-Yosida regularization of ρ.

Another important relation concerning the prox function is the following: when ρ is a closed, proper, convex function, we have

$$x=\text{prox}_1(\rho)(x)+\text{prox}_1(\rho^*)(x)\ ,$$

where ρ^* is the Fenchel-Legendre conjugate of ρ, namely, in the scalar case, $\rho^*(x)=\sup_y xy-\rho(y)$. We again refer the reader to the original paper of Moreau (Moreau (1965)) or Rockafellar (1997) for statements in English.

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.

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey, J. W. (1972). *Robust estimates of location: Survey and advances.* Princeton University Press, Princeton, N.J.

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.

Ben-Israel, A. and Greville, T. N. E. (2003). *Generalized inverses.* CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC, 15. Springer-Verlag, New York, second edition. Theory and applications.

Bickel, P. J. (1975). One-step Huber estimates in the linear model. *J. Amer. Statist. Assoc.* **70**, 428–434. URL `http://www.jstor.org/stable/2285834`.

Bickel, P. J. (1981). Quelques aspects de la statistique robuste. In *Ninth Saint Flour Probability Summer School—1979 (Saint Flour, 1979)*, volume 876 of *Lecture Notes in Math.*, pp. 1–72. Springer, Berlin.

Bickel, P. J. (1984). Robust regression based on infinitesimal neighbourhoods. *Ann. Statist.* **12**, 1349–1368. URL `http://dx.doi.org/10.1214/aos/1176346796`.

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., Maleki, A., and Montanari, A. (2009a). Message-passing algorithms for compressed sensing. *PNAS* **106**, 18914–18919.

Donoho, D., Maleki, A., and Montanari, A. (2009b). Message passing algorithms for compressed sensing: I. Motivation and construction. *Available on Arxiv* URL `http://arxiv.org/abs/0911.4219v1`.

Eaton, M. L. (2007). *Multivariate statistics.* Institute of Mathematical Statistics Lecture Notes—Monograph Series, 53. Institute of Mathematical Statistics. Reprint of the 1983 original.

El Karoui, N. (2009a). 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. (2009b). On the realized risk of high-dimensional Markowitz portfolios. Technical Report 784, Department of Statistics, UC Berkeley. Submitted to SIAM Journal in Financial Mathematics.

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. 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).

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. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35**, 73–101.

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`.

Koenker, R. (2005). *Quantile regression*, volume 38 of *Econometric Society Monographs.* Cambridge University Press, Cambridge. URL `http://dx.doi.org/10.1017/CBO9780511754098`.

Ledoux, M. (2001). *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs.* American Mathematical Society, Providence, RI.

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`.

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust statistics.* Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., Chichester. URL `http://dx.doi.org/10.1002/0470010940`. Theory and methods.

Mézard, M. and Montanari, A. (2009). *Information, physics, and computation.* Oxford Graduate Texts. Oxford University Press, Oxford. URL `http://dx.doi.org/10.1093/acprof:oso/9780198570837.001.0001`.

Moreau, J.-J. (1965). Proximité et dualité dans un espace hilbertien. *Bull. Soc. Math. France* **93**, 273–299.

Muirhead, R. J. (1982). *Aspects of multivariate statistical theory.* John Wiley & Sons Inc., New York. Wiley Series in Probability and Mathematical Statistics.

Penrose, R. (1956). On best approximation solutions of linear matrix equations. *Proc. Cambridge Philos. Soc.* **52**, 17–19.

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`.

Relles, D. (1968). *Robust Regression by Modified Least Squares.* Ph.D. thesis, Yale University.

Rockafellar, R. T. (1997). *Convex analysis.* Princeton Landmarks in Mathematics. Princeton University Press, Princeton, NJ. Reprint of the 1970 original, Princeton Paperbacks.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9**, 1135–1151. URL `http://www.jstor.org/stable/2240405`.

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.

Yohai, V. J. (1974). Robust estimation in the linear model. *Ann. Statist.* **2**, 562–567. Collection of articles dedicated to Jerzy Neyman on his 80th birthday.