Electronic Journal of Statistics
Vol. 13 (2019) 2830–2864
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1590

Multidimensional linear functional estimation in sparse Gaussian models and robust estimation of the mean

Olivier Collier[1,2] **and Arnak S. Dalalyan**[2]

Modal'X, Université Paris-Nanterre[1] *and CREST, ENSAE*[2]

Abstract: We consider two problems of estimation in high-dimensional Gaussian models. The first problem is that of estimating a linear functional of the means of n independent p-dimensional Gaussian vectors, under the assumption that at most s of the means are nonzero. We show that, up to a logarithmic factor, the minimax rate of estimation in squared Euclidean norm is between $(s^2 \wedge n) + sp$ and $(s^2 \wedge np) + sp$. The estimator that attains the upper bound being computationally demanding, we investigate suitable versions of group thresholding estimators that are efficiently computable even when the dimension and the sample size are very large. An interesting new phenomenon revealed by this investigation is that the group thresholding leads to a substantial improvement in the rate as compared to the element-wise thresholding. Thus, the rate of the group thresholding is $s^2\sqrt{p} + sp$, while the element-wise thresholding has an error of order $s^2 p + sp$. To the best of our knowledge, this is the first known setting in which leveraging the group structure leads to a polynomial improvement in the rate.

The second problem studied in this work is the estimation of the common p-dimensional mean of the inliers among n independent Gaussian vectors. We show that there is a strong analogy between this problem and the first one. Exploiting it, we propose new strategies of robust estimation that are computationally tractable and have better rates of convergence than the other computationally tractable robust (with respect to the presence of the outliers in the data) estimators studied in the literature. However, this tractability comes with a loss of the minimax-rate-optimality in some regimes.

MSC 2010 subject classifications: Primary 62J05; secondary 62G05.
Keywords and phrases: Column-sparsity, Minimax estimation, Group-sparsity, Linear transformation, High-dimensional inference, Robust estimation.

Received December 2018.

1. Introduction

Linear functionals are of central interest in statistics. The problems of estimating a function at given points, predicting the value of a future observation, testing the validity of a hypothesis, finding a dimension reduction subspace are all examples of statistical inference on linear functionals. The primary goal of this paper is to investigate the problem of estimation of a particular form of linear functional defined as the sum of the observed multidimensional signals. Although

this problem is of independent interest on its own, one of our motivations for studying it is its tight relation with the problem of robust estimation.

Various aspects of the problem of estimation of a linear functional of an unknown high-dimensional or even infinite-dimensional parameter were studied in the literature, mostly focusing on the case of a functional taking real values (as opposed to the vector valued functional considered in the present work). Early results for smooth functionals were obtained by (Koshevnik and Levit, 1977). Minimax estimation of linear functionals over various classes and models were thoroughly analyzed by (Donoho and Liu, 1987; Klemela and Tsybakov, 2001; Efromovich and Low, 1994; Golubev and Levit, 2004; Cai and Low, 2004, 2005; Laurent, Ludena and Prieur, 2008; Butucea and Comte, 2009; Juditsky and Nemirovski, 2009). There is also a vast literature on studying the problem of estimating quadratic functionals (Donoho and Nussbaum, 1990; Laurent and Massart, 2000; Cai and Low, 2006; Bickel and Ritov, 1988). Since the estimators of (quadratic) functionals can be often used as test statistics, the problem of estimating functionals has close relations with the problem of testing that were successfully exploited in (Comminges and Dalalyan, 2012, 2013; Collier and Dalalyan, 2015; Lepski, Nemirovski and Spokoiny, 1999). The problem of estimation of nonsmooth functionals was also tackled in the literature, see (Cai and Low, 2011).

Some statistical problems related to functionals of high-dimensional parameters under various types of sparsity constraints were recently addressed in several papers. The case of real valued linear and quadratic functionals was studied by Collier, Comminges and Tsybakov (2017) and Collier et al. (2016), focusing on the Gaussian sequence model. Verzelen and Gassiat (2016) analyzed the problem of the signal-to-noise ratio estimation in the linear regression model under various assumptions on the design. In a companion paper of the present submission, (Collier and Dalalyan, 2018) considered the problem of a vector valued linear functional estimation when the observations are drawn from a Poisson distribution. It turns out that the result established in the present work for the group (hard and soft) thresholding estimators are valid for the Poisson model as well, but it is not the case for the results on the greedy estimator studied in Section 2.1.

We first investigate the order of magnitude of the worst-case risk of three types of estimators of a linear functional: the greedy subset selection (GSS), the group (hard and soft) thresholding (GHT and GST) and the componentwise thresholding (HT). We then establish a non-asymptotic lower bound on the minimax risk that shows its dependence on the three main parameters of the model: the sample size n, the dimension p and the (column-)sparsity s. This lower bound implies that the greedy subset selection is minimax rate optimal in the sparse regime $s = O(p \vee \sqrt{n})$, whereas the group thresholding is minimax rate optimal in the super-sparse case $s = O(\sqrt{p})$. The advantage of the group thresholding as compared to the greedy subset selection is that the former is computationally efficient, whereas the latter is not. In all these considerations, we neglect logarithmic factors. Table 1 summarizes our main contributions related to the problem of linear functional estimation.

TABLE 1
A summary of our results related to the estimation of a linear functional. The risk is normalized by σ^2, the variance of the noise, and the bounds of the second column hide logarithmic factors and multiplicative universal constants.

Estimator	Risk Bound (up to log factors)	Computationally efficient	Stated in
GSS	$sp + s^2 \wedge np$	No	Theorem 1
GHT	$sp + s^2\sqrt{p} \wedge np$	Yes	Theorem 2
GST	$sp + s^2\sqrt{p} \wedge np$	Yes	Theorem 4
HT	$sp + s^2 p \wedge np$	Yes	Theorem 3
Lower bound	$sp + s^2 \wedge n$	–	Theorem 5

In particular, one can observe that the ratio of the worst-case risk of the group thresholding procedure and that of the component-wise thresholding might be as small as $O(p^{-1/2})$. Indeed, it is not hard to prove that the performance of the component-wise thresholding estimator is of the order of p times the performance of this same estimator in the scalar case, *i.e.,* when $p = 1$. But it is proven in Theorem 1 in (Collier, Comminges and Tsybakov, 2017) that the optimal rate of estimation of the (scalar) linear functional is $\sigma^2 s^2 \log(n/s^2)$, which is in fact reached by hard-thresholding.

To the best of our knowledge, this is the first known setting in which leveraging the group structure leads to such an important improvement of the rate. In previous results, the improvement was of at most logarithmic order. Another interesting remark is that the group soft thresholding estimator we investigate here has a data-dependent threshold[1]. Finally, note that while the thresholding estimators are natural candidates for solving the problem under consideration in the sparsity setting, the greedy subset selection is a new procedure introduced in this paper to get the best known upper bound on the minimax risk.

A second problem studied in this work is the robust estimation of the mean of a Gaussian vector. As explained in forthcoming sections, this problem has close relations to that of estimation of a linear functional. In order to explain this relation, let us recall that one of the most popular mathematical framework for analyzing robust estimators is the Huber contamination model (Huber, 1964). It assumes that there is a reference distribution $\mathbf{P}_{\boldsymbol{\mu}}$, parametrized by $\boldsymbol{\mu} \in \mathcal{M}$, the precise value of which is unknown, and a contamination distribution $\mathbf{Q}$, which is completely unknown. The data points Y_i, $i = 1, \dots, n$ are independent random variables drawn from the mixture distribution $\mathbf{P}_{\epsilon,\boldsymbol{\mu},\mathbf{Q}} = (1-\epsilon)\mathbf{P}_{\boldsymbol{\mu}} + \epsilon\,\mathbf{Q}$, where $\epsilon \in [0, 1]$ is the rate of contamination. The goal is then to estimate the parameter $\boldsymbol{\mu}$, see the papers (Chen, Gao and Ren, 2015; Chen, Gao and Ren, 2016) for some recent results. This means that among the n observations, there are s inliers drawn from $\mathbf{P}_{\boldsymbol{\mu}}$ and $(n-s)$ outliers drawn from $\mathbf{Q}$, all these observations being independent and s being a binomial random variable with parameters n

[1] Although we do not have a formal proof of that, but all the computations we did make us believe that it is impossible to get such a small risk bound for the group soft thresholding estimator based on a threshold that does not depend on data.

and $(1-\epsilon)$. Thus, the specificity of the model is that all the outliers are assumed to be drawn from the same distribution, $\mathbf{Q}$.

We suggest here to consider an alternative model for the outliers. In the general setting, it corresponds to considering the number of outliers, s, as a deterministic value and to assuming that the outliers $\{Y_i : i \in O\}$ (where $O \subset [n]$ is of cardinality s) are independent and satisfy $Y_i \sim \mathbf{P}_{\boldsymbol{\mu}_i}$. Thus, we do not assume in this model that the outliers are all generated by the same random mechanism. This model and the Huber model are two different frameworks for assessing the quality of the estimators. It is quite likely that in real world applications none of these two models are true. However, both of them are of interest for comparing various outlier-robust estimators and investigating optimality properties.

To explain the connection between the robust estimation and the problem of estimation of a linear functional, let us consider the contamination model of the previous paragraph. That is, we assume that the observations $\boldsymbol{Y}_i$ are independent and drawn from $\mathbf{P}_{\boldsymbol{\mu}_i}$, with $\boldsymbol{\mu}_i = \boldsymbol{\mu}$ for every inlier $i \in O^c = \{1, \dots, n\} \setminus O$. In addition, let $\boldsymbol{\mu}$ be the mean of $\mathbf{P}_{\boldsymbol{\mu}}$ and the family $\{\mathbf{P}_{\boldsymbol{\mu}}\}$ be translation invariant (meaning that for every vector $\boldsymbol{a}$, the random variable $\boldsymbol{Y}_i - \boldsymbol{a}$ is drawn from $\mathbf{P}_{\boldsymbol{\mu}_i - \boldsymbol{a}}$). If we have an initial estimator $\widehat{\boldsymbol{\mu}}_0$ of $\boldsymbol{\mu}$, which is consistent but not necessarily rate-optimal, then we can define the centered observations $\boldsymbol{Y}'_i = \boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}_0$. Each observation $\boldsymbol{Y}'_i$ will have a distribution close to $\mathbf{P}_{\boldsymbol{\theta}_i}$, where $\{\boldsymbol{\theta}_i \triangleq \boldsymbol{\mu}_i - \boldsymbol{\mu}, i \in [n]\}$ is a sparse set of vectors, so that $\frac{1}{n}\sum_{i\in[n]} \boldsymbol{Y}_i$ is a natural estimator of $\boldsymbol{\mu} + \frac{1}{n}\sum_{i\in[n]} \boldsymbol{\theta}_i$. The strategy we propose here is to use an estimator $\widehat{\boldsymbol{L}}_n$—based on the transformed observations $\boldsymbol{Y}'_i$— of the linear functional $\boldsymbol{L}_n = \frac{1}{n}\sum_{i\in[n]} \boldsymbol{\theta}_i$ and then to update the estimator of $\boldsymbol{\mu}$ by the formula $\widehat{\boldsymbol{\mu}}_1 = \frac{1}{n}\sum_{i\in[n]} \boldsymbol{Y}_i - \widehat{\boldsymbol{L}}_n$. This procedure can be iterated using $\widehat{\boldsymbol{\mu}}_1$ as an initial estimator of $\boldsymbol{\mu}$. We elaborate on this approach in the case of the normal distribution, $\mathbf{P}_{\boldsymbol{\mu}} = \mathcal{N}_p(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_p)$, in the second part of the present work.

1.1. Organization

The rest of the paper is organized as follows. Section 2 is devoted to the problem of linear functional estimation. It contains the statements of the main results concerning the risk bounds of different relevant estimators and some lower bounds on the minimax risk. The problem of robust estimation is addressed in Section 3. We summarize our findings and describe some directions of future research in Section 4. The proofs of main theorems are postponed to Section 5, whereas the proofs of technical lemmas are gathered in Section 6. Some well-known results frequently used in the present work are recalled in Section 7.

1.2. Notation

We denote by $[k]$ the set of integers $\{1, \dots, k\}$. The k-dimensional vectors containing only ones and only zeros are denoted by $\mathbf{1}_k$ and $\mathbf{0}_k$, respectively. As usual, $\|\boldsymbol{u}\|_2$ stands for the Euclidean norm of a vector $\boldsymbol{u} \in \mathbb{R}^k$. The $k \times k$ identity matrix is denoted by $\mathbf{I}_k$. For every $p \times n$ matrix $\mathbf{M}$ and every $T \subset [n]$,

we denote by $\mathbf{M}_T$ the submatrix of $\mathbf{M}$ obtained by removing the columns with indices lying outside T. The Frobenius norm of $\mathbf{M}$, denoted by $\|\mathbf{M}\|_F$, is defined by $\|\mathbf{M}\|_F^2 = \mathbf{tr}(\mathbf{M}^\top \mathbf{M})$, and its spectral norm, denoted by $\|\mathbf{M}\|$, is defined as square-root of the largest eigenvalue of $\mathbf{M}^T\mathbf{M}$. We will use the notation $\boldsymbol{L}(\mathbf{M})$ for the linear functional $\mathbf{M}\mathbf{1}_n$ equal to the sum of the columns of $\mathbf{M}$.

2. Estimation of a linear functional

We assume that we are given a $p \times n$ matrix $\mathbf{Y}$ generated by the following model:

$$\mathbf{Y} = \boldsymbol{\Theta} + \sigma \boldsymbol{\Xi}, \quad \xi_{i,j} \overset{\text{iid}}{\sim} \mathcal{N}(0,1). \tag{1}$$

This means that the deterministic matrix $\boldsymbol{\Theta}$ is observed in Gaussian white noise of variance σ^2. Equivalently, the columns $\boldsymbol{Y}_i$ of $\mathbf{Y}$ satisfy

$$\boldsymbol{Y}_i = \boldsymbol{\theta}_i + \sigma \boldsymbol{\xi}_i, \quad \boldsymbol{\xi}_i \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p), \quad i = 1, \dots, n.$$

Our goal is to estimate the vector $\boldsymbol{L}(\boldsymbol{\Theta}) \in \mathbb{R}^p$, where $\boldsymbol{L} : \mathbb{R}^{p \times n} \to \mathbb{R}^p$ is the linear transformation defined by

$$\boldsymbol{L}(\boldsymbol{\Theta}) = \sum_{i=1}^n \boldsymbol{\theta}_i = \boldsymbol{\Theta}\mathbf{1}_n. \tag{2}$$

Let us first explain that this is a nontrivial statistical problem, at least when both p and n are large. In fact, the naive solution to the aforementioned problem consists in replacing in (2) the unknown matrix $\boldsymbol{\Theta}$ by the noisy observation $\mathbf{Y}$. This leads to the estimator $\widehat{\boldsymbol{L}} = \mathbf{Y}\mathbf{1}_n$, the risk of which can be easily shown to be

$$\mathbf{E}_{\boldsymbol{\Theta}} \|\widehat{\boldsymbol{L}} - \boldsymbol{L}(\boldsymbol{\Theta})\|_2^2 = \sigma^2 np.$$

When the matrix $\boldsymbol{\Theta}$ has at most s nonzero columns with s being much smaller than n, it is possible to design estimators that perform much better than the naive estimator $\widehat{\boldsymbol{L}}_n$. Indeed, an oracle who knows the sparsity pattern $S = \{i \in [n] : \boldsymbol{\theta}_i \neq 0\}$ may use the oracle-estimator $\widehat{\boldsymbol{L}}_S = \boldsymbol{L}(\mathbf{Y}_S)$ which has a risk equal to $\sigma^2 sp$. It is not difficult to show that there is no estimator having a smaller risk uniformly over all the matrices $\boldsymbol{\Theta}$ with a given sparsity pattern S of cardinality s. Thus, we have two benchmarks: the very slow rate $\sigma^2 np$ attained by the naive estimator and the fast rate $\sigma^2 sp$ attained by the oracle-estimator that is unavailable in practice. The general question that we study in this work is the following: what is the best possible rate in the range $[\sigma^2 sp, \sigma^2 np]$ that can be obtained by an estimator that does not rely on the knowledge of S?

In what follows, we denote by $\mathcal{M}(p, n, s)$ the set of all $p \times n$ matrices with real entries having at most s nonzero columns:

$$\mathcal{M}(p, n, s) = \Big\{ \boldsymbol{\Theta} \in \mathbb{R}^{p \times n} : \sum_{i=1}^n \mathbb{1}(\|\boldsymbol{\theta}_i\|_2 > 0) \leq s \Big\}.$$

2.1. Greedy subset selection

Let us consider a greedy estimator that tries to successively recover various pieces of the sparsity pattern S. We start by setting $I_1 = [n]$ and $\mathcal{I}_1 = \big\{J \subseteq I_1 : \|\boldsymbol{L}(\mathbf{Y}_J)\|_2^2 \geq 12\sigma^2(|J|p + \lambda|J|^2)\big\}$. If $\mathcal{I}_1$ is empty, then we set $\widehat{J}_1 = \varnothing$ and terminate. Otherwise, *i.e.,* when $\mathcal{I}_1$ is not empty, we set $\widehat{J}_1 = \arg\min\big\{|J| : J \in \mathcal{I}_1\big\}$ and $I_2 = I_1 \setminus \widehat{J}_1$. In the next step, we define $\mathcal{I}_2$, $\widehat{J}_2$ and I_3 in the same way using as starting point I_2 instead of I_1. We repeat this procedure until we get $\widehat{J}_\ell = \varnothing$ or $I_{\ell+1} = \varnothing$. Then we set

$$\widehat{S} = \widehat{J}_1 \cup \ldots \cup \widehat{J}_\ell \quad \text{and} \quad \widehat{\boldsymbol{L}}^{\mathrm{GSS}} = \boldsymbol{L}(\mathbf{Y}_{\widehat{S}}).$$

The detailed pseudo-code for this algorithm is given in Algorithm 1 below.

Algorithm 1: Greedy subset selection algorithm

input : matrix $\mathbf{Y}$ and noise variance σ, threshold λ.
output: vector $\widehat{\boldsymbol{L}}^{\mathrm{GSS}}$.
initialization $I \leftarrow [n]$ and $\widehat{S} \leftarrow \varnothing$.
repeat
  Set $\mathcal{I} \leftarrow \big\{J \subseteq I : \|\boldsymbol{L}(\mathbf{Y}_J)\|_2^2 \geq 12\sigma^2(|J|p + \lambda|J|^2)\big\}$.
  if $\mathcal{I} = \varnothing$ **then**
    $\widehat{J} \leftarrow \varnothing$
  else
    Set $\widehat{J} \leftarrow \arg\min\big\{|J| : J \in \mathcal{I}\big\}$. (Ties can be broken arbitrarily.)
  end if
  Update $\widehat{S} \leftarrow \widehat{S} \cup \widehat{J}$.
  Update $I \leftarrow I \setminus \widehat{J}$.
until *I is empty or $\widehat{J}$ is empty*
return $\widehat{\boldsymbol{L}}^{\mathrm{GSS}} \leftarrow \sum_{i \in \widehat{S}} \boldsymbol{Y}_i$

Theorem 1. *Let $\delta \in (0,1)$ be a prescribed tolerance level. The greedy subset selection estimator with $\lambda = 3/2 \log(2n/\delta)$ satisfies*

$$\sup_{\boldsymbol{\Theta} \in \mathcal{M}(p,n,s)} \mathbf{P}_{\boldsymbol{\Theta}}\Big(\|\widehat{\boldsymbol{L}}^{\mathrm{GSS}} - \boldsymbol{L}(\boldsymbol{\Theta})\|_2^2 \leq 60\sigma^2 s(p + \lambda s)\Big) \geq 1 - \delta. \tag{3}$$

This result tells us that the worst-case rate of convergence of the GSS estimator over the class $\mathcal{M}(p,n,s)$ is $\sigma^2 s(p + s\log n)$. As a consequence, the minimax risk of estimating the functional $\boldsymbol{L}(\boldsymbol{\Theta})$ over the aforementioned class is at most of order $\sigma^2 s(p + s\log n)$. As we will see below, this rate is optimal up to a logarithmic factor.

However, from a practical point of view, the GSS algorithm has limited applicability because of its high computational cost. It is therefore appealing to look for other estimators that can be computed efficiently even though their estimation error does not decay at the optimal rate for every possible configuration on (p,n,s). Let us note here that using standard tools it is possible to establish an upper bound similar to (3) that holds in expectation.

2.2. Group hard thresholding estimator

A natural approach to the problem of estimating $\boldsymbol{L}(\boldsymbol{\Theta})$ consists in filtering out all the signals $\boldsymbol{Y}_i$ that have a large norm and computing the sum of the remaining signals. This is equivalent to solving the following optimization problem

$$\widehat{\boldsymbol{\Theta}}^{\mathrm{GHT}} = \arg\min_{\mathbf{T}} \Big\{ \|\mathbf{Y} - \mathbf{T}\|_F^2 + \lambda^2 \sum_{i=1}^{n} \mathbb{1}_{\boldsymbol{t}_i \neq \mathbf{0}} \Big\},$$

where $\lambda > 0$ is a tuning parameter. The estimator $\widehat{\boldsymbol{\Theta}}^{\mathrm{GHT}}$, hereafter referred to as group hard thresholding, minimizes the negative log-likelihood penalized by the number of non-zero columns in $\boldsymbol{\Theta}$. One easily checks that the foregoing optimization problem can be solved explicitly and the resulting estimator is

$$\widehat{\boldsymbol{\theta}}_i^{\mathrm{GHT}} = \boldsymbol{Y}_i \mathbb{1}_{\|\boldsymbol{Y}_i\|_2 \geq \lambda}, \qquad i \in [n].$$

Using the group hard thresholding estimator of $\boldsymbol{\Theta}$ and the method of substitution, we can estimate $\boldsymbol{L}(\boldsymbol{\Theta})$ by

$$\widehat{\boldsymbol{L}}^{\mathrm{GHT}} = \boldsymbol{L}(\widehat{\boldsymbol{\Theta}}^{\mathrm{GHT}}). \tag{4}$$

It is clear that this estimator is computationally far more attractive than the GSS estimator presented above. Indeed, the computation of the GHT estimator requires at most $O(pn)$ operations. However, as stated in the next theorem, this gain is achieved at the expense of a higher statistical error.

Theorem 2. *Let $\widehat{\boldsymbol{L}}^{\mathrm{GHT}}$ be the estimator defined in (4) with the tuning parameter*

$$\lambda^2/\sigma^2 = p + 4\big\{ \log(1 + n/s^2) \vee p^{1/2} \log^{1/2}(1 + n^2 p/s^4) \big\}.$$

There exists a universal constant $c > 0$ such that, for every $\boldsymbol{\Theta} \in \mathcal{M}(p, n, s)$, it holds

$$\mathbf{E}_{\boldsymbol{\Theta}} \big[\big\| \widehat{\boldsymbol{L}}^{\mathrm{GHT}} - \boldsymbol{L}(\boldsymbol{\Theta}) \big\|_2^2 \big] \leq c\sigma^2 \Big(s^2 p^{1/2} \log^{1/2}(1 + n^2 p/s^4) + s^2 \log(1 + n/s^2) + sp \Big).$$

Using the fact that $\log(1 + x) \leq x$, we infer from this theorem that the rate of the group hard thresholding for fixed σ is of order $s^2\sqrt{p} \wedge np + sp$, up to a logarithmic factor. Moreover, the rate obtained in this theorem can not be improved, up to logarithmic factors, as stated in the next theorem.

Proposition 1. *Let us denote by $\widehat{\boldsymbol{L}}_\lambda^{\mathrm{GHT}}$ the estimator defined in (4) with a threshold $\lambda > 0$. There are two universal constants $p_0 \in \mathbb{N}$ and $c > 0$, such that for any $p \geq p_0$ and $s \leq n/61$, the following lower bound holds*

$$\inf_{\lambda > 0} \sup_{\boldsymbol{\Theta} \in \mathcal{M}(p,n,s)} \mathbf{E}_{\boldsymbol{\Theta}} \big\| \widehat{\boldsymbol{L}}_\lambda^{\mathrm{GHT}} - \boldsymbol{L}(\boldsymbol{\Theta}) \big\|_2^2 \geq c\sigma^2 \Big((s^2 p^{1/2}) \wedge (np) + s^2 \log(1 + n/s^2) + sp \Big).$$

The proofs of these claims being deferred to Section 5, let us comment on the stated results. At first sight the presence of the sparsity s in the definition of the threshold λ in Theorem 2 might seem problematic, since this quantity is unknown in most practical situations. However, one can easily modify the claim of Theorem 2 replacing n/s^2 and $np^{1/2}/s^2$ respectively by n and $np^{1/2}$ both in the definition of λ and the subsequent risk bound. This will lead to the rate $s^2p^{1/2} + sp$, up to a logarithmic factor. In order to obtain the better rate $s^2p^{1/2} \wedge (np)$, it suffices to project the estimator $\widehat{\boldsymbol{L}}_\lambda^{\mathrm{GHT}}$ onto the ball in $\mathbb{R}^p$ centered at $\widehat{\boldsymbol{L}}_0^{\mathrm{GHT}} = \boldsymbol{L}(\mathbf{Y})$ with squared radius $O(np \log n\, \sigma)$, as shown in the following result.

Proposition 2. *Let $\widehat{\boldsymbol{L}}$ be any estimator of $\boldsymbol{L}(\boldsymbol{\Theta})$. Define the projected estimator $\widehat{\boldsymbol{L}}^{\mathrm{pr}} = \arg\min\{\|\boldsymbol{u} - \widehat{\boldsymbol{L}}\|_2 : \|\boldsymbol{u} - \boldsymbol{L}(\mathbf{Y})\|_2^2 \le 7\sigma^2 np \log n\}$. For $n \ge 3$, $p \ge 2$ and any matrix $\boldsymbol{\Theta}$, we have*

$$\mathbf{E}_{\boldsymbol{\Theta}} \big\| \widehat{\boldsymbol{L}}^{\mathrm{pr}} - \boldsymbol{L}(\boldsymbol{\Theta}) \big\|_2^2 \le 10\sigma^2 + (28\sigma^2 np \log n) \wedge \mathbf{E}_{\boldsymbol{\Theta}} \big\| \widehat{\boldsymbol{L}} - \boldsymbol{L}(\boldsymbol{\Theta}) \big\|_2^2.$$

A second remark concerns the rate optimality. If we neglect the logarithmic factors in this discussion, the rate of the GHT estimator is shown to be at most of order $\sigma^2(s^2\sqrt{p} \wedge np + sp)$. This coincides with the optimal rate (and the one of the GSS estimator) when $s = O(\sqrt{p})$ and has an extra factor $\sqrt{p}$ in the worst-case. The latter corresponds to the regime in which $s^2\sqrt{p} \wedge np + sp = \Theta(np)$ and $s^2 \wedge n + sp = \Theta(n)$. This is equivalent to $p = O(s^4/n^2 \wedge n/s)$. When there is a limit on the computational budget, that is when the attention is restricted to the estimators computable in polynomial (in s, p, n) time, we do not know whether such a deterioration of the risk can be avoided.

An inspection of the proof of Theorem 2 shows that if all the nonzero signals $\boldsymbol{\theta}_i$ are large enough, that is when $\min_{i \in S} \|\boldsymbol{\theta}_i\|_2^2 \ge cp$ for some constant $c > 0$, the extra factor $\sqrt{p}$ disappears and the GHT achieves the optimal rate. Put differently, the signals at which the GHT estimator fails to achieve the optimal rate are those having an Euclidean norm of order $p^{1/4}$. This is closely related to the minimax rate of separation in hypotheses testing. It is known that the separation rate for testing $H_0 : \boldsymbol{\theta} = \mathbf{0}$ against $H_1 : \|\boldsymbol{\theta}\|_2 \ge \rho$, when one observes $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\theta}, \sigma^2 \mathbf{I}_p)$ is of order $\sigma p^{1/4}$. This follows for example from Theorem 12 in (Collier, Comminges and Tsybakov, 2017).

Our last remark on Theorem 2 concerns the relation with element-wise hard thresholding. The idea is the following: any column-sparse matrix $\boldsymbol{\Theta}$ is also sparse in the most common sense of sparsity. That is, the number of nonzero entries of the matrix $\boldsymbol{\Theta}$ is only a small fraction of the total number of entries. Therefore, one can estimate the entries of $\boldsymbol{\Theta}$ by thresholding those of $\mathbf{Y}$ and then estimate $\boldsymbol{L}(\boldsymbol{\Theta})$ by the method of substitution. The statistical complexity of this estimator is quantified in the next theorem, the proof of which is similar to the corresponding theorem in (Collier, Comminges and Tsybakov, 2017).

Theorem 3. *Let $\widehat{\boldsymbol{L}}^{\mathrm{HT}}$ be the element-wise hard thresholding estimator defined by $\widehat{\boldsymbol{L}}_i^{\mathrm{HT}} = \sum_{j=1}^n \boldsymbol{Y}_{i,j} \mathbb{1}_{\boldsymbol{Y}_{i,j} > \lambda}$ for $i \in [p]$. If the threshold λ is chosen so that*

$\lambda^2 = 2\sigma^2 \log(1 + n/s^2)$, *then*

$$\sup_{\boldsymbol{\Theta} \in \mathcal{M}(p,n,s)} \mathbf{E}_{\boldsymbol{\Theta}} \big[\|\widehat{\boldsymbol{L}}^{\mathrm{HT}} - \boldsymbol{L}(\boldsymbol{\Theta})\|^2 \big] \le c\sigma^2 s^2 p \log(1 + n/s^2),$$

where $c > 0$ *is a universal constant.*

A striking feature of the problem of linear functional estimation uncovered by Theorem 2 and Theorem 3, is that exploiting the group structure leads to an improvement of the risk which may attain a factor $p^{-1/2}$ (for the squared Euclidean norm). To the best of our knowledge, this is the first framework in which the grouping is proved to have such a strong impact. This can be compared to the problem of estimating the matrix $\boldsymbol{\Theta}$ itself under the same sparsity assumptions. Provable guarantees in such a setting show only a logarithmic improvement due to the use of the sparsity structure (Lounici et al., 2011; Bunea, Lederer and She, 2014).

2.3. Group-soft-thresholding estimator

A natural question is whether the results obtained above for the group hard thresholding can be carried over a suitable version of the soft-thresholding estimator. Such an extension could have two potential benefits. First, the soft thresholding is defined as a solution to a convex optimization problem, whereas hard thresholding minimizes a nonconvex cost function. This difference makes the soft thresholding method more suitable to deal with various statistical problems. The simplest example is the problem of linear regression: the extension of the soft thresholding estimator to the case of non-orthogonal design is the LASSO, that can be computed even when the dimension is very large. In the same problem, the extension of the hard thresholding is the BIC-type estimator, the computation of which is known to be prohibitively complex when the dimension is large.

A second reason motivating our interest in the soft thresholding is its smooth dependence on the data. This smoothness implies that the estimator is less sensitive to changes in the data than the hard thresholding. Furthermore, it makes it possible to design a SURE-type algorithm for defining an unbiased estimator of the risk and, eventually, selecting the tuning parameter in a data-driven way.

Finally, anticipating the results of the next section, we would like to stress that the analysis of the group-soft-thresholding estimator prepares the ground for the robust estimator of the mean studied in Section 3.

In the model under consideration, the group soft thresholding estimator $\widehat{\boldsymbol{\Theta}}^{\mathrm{GST}}$ can be defined as the minimizer of the group-LASSO cost function, that is

$$\widehat{\boldsymbol{\Theta}}^{\mathrm{GST}} = \arg\min_{\mathbf{T}} \Big\{ \sum_{i=1}^{n} \|\boldsymbol{Y}_i - \boldsymbol{t}_i\|_2^2 + \sum_{i=1}^{n} \lambda_i \|\boldsymbol{t}_i\|_2 \Big\}.$$

This problem has an explicit solution given by

$$\widehat{\boldsymbol{\theta}}_i^{\mathrm{GST}} = \Big(1 - \frac{\lambda_i}{2\|\boldsymbol{Y}_i\|_2}\Big)_+ \boldsymbol{Y}_i. \tag{5}$$

It is natural then to define the plug-in estimator as $\widehat{\boldsymbol{L}}^{\mathrm{GST}} = \boldsymbol{L}(\widehat{\boldsymbol{\Theta}}^{\mathrm{GST}})$. The next theorem establishes the performance of this estimator.

Theorem 4. *The estimator* $\widehat{\boldsymbol{L}}^{\mathrm{GST}} = \boldsymbol{L}(\widehat{\boldsymbol{\Theta}}^{\mathrm{GST}})$ *defined in* (5) *with*[2]

$$\lambda_i = \frac{2\sigma\gamma\|\boldsymbol{Y}_i\|_2}{\big(\|\boldsymbol{Y}_i\|_2^2 - \sigma^2 p\big)_+^{1/2}}, \quad \gamma^2 = 4\big\{\log(1+n/s^2) \vee p^{1/2}\log^{1/2}(1+n^2p/s^4)\big\}$$

satisfies, for every $\boldsymbol{\Theta} \in \mathcal{M}(p,n,s)$,

$$\mathbf{E}_{\boldsymbol{\Theta}}\big[\big\|\widehat{\boldsymbol{L}}^{\mathrm{GST}} - \boldsymbol{L}(\boldsymbol{\Theta})\big\|^2\big] \le c\sigma^2\Big(s^2p^{1/2}\log^{1/2}(1+n^2p/s^4) + s^2\log(1+n/s^2) + sp\Big),$$

where $c > 0$ *is some universal constant.*

The comments made after the statement of Theorem 2 can be repeated here. The dependence of γ on s is not crucial; one can replace s by 1 in the expression for γ, this will not have a strong impact on the risk bound. The bound in expectation can be complemented by a bound in deviation. The rate obtained for the soft thresholding is exactly of the same order as the obtained in Theorem 2 for the group hard thresholding. A notable difference, however, is that in the case of soft thresholding the tuning parameter λ suggested by the theoretical developments is data dependent.

2.4. Lower bounds and minimax rate optimality

We now address the question of the optimality of our estimators. In (Collier, Comminges and Tsybakov, 2017), the case $p = 1$ was solved with lower and upper bounds matching up to a constant. In particular, Theorem 1 in (Collier, Comminges and Tsybakov, 2017) yields the following proposition.

Proposition 3. *There is a universal constant* $c > 0$ *such that, for any* $s \in [n]$,

$$\inf_{\widehat{\boldsymbol{L}}} \sup_{\boldsymbol{\Theta}\in\mathcal{M}(1,n,s)} \mathbf{E}_{\boldsymbol{\Theta}}\big(\widehat{\boldsymbol{L}} - \boldsymbol{L}(\boldsymbol{\Theta})\big)^2 \ge c\sigma^2 s^2\log(1+n/s^2).$$

Note that when $n = s$, this rate is of the order of $\sigma^2 s$. It is straightforward that this rate generalizes to $\sigma^2 sp$ in the multidimensional case. Furthermore, if we knew in advance the sparsity pattern S, then we could restrict the matrix of observations to the indices in S, and we would get the oracle rate $\sigma^2 sp$. These remarks are made formal in the following theorem.

[2]Note that $\lambda_i = +\infty$ if $\|\boldsymbol{Y}_i\|_2^2 \le \sigma^2 p$. This reflects the fact that there is no need to fit the signals of very low magnitude.

Theorem 5. *Assume that $1 \le s \le n$, then there is a universal constant $c > 0$ such that*

$$\inf_{\widehat{\boldsymbol{L}}} \sup_{\boldsymbol{\Theta} \in \mathcal{M}(p,n,s)} \mathbf{E}_{\boldsymbol{\Theta}} \big\| \widehat{\boldsymbol{L}} - \boldsymbol{L}(\boldsymbol{\Theta}) \big\|^2 \ge c \Big[\sigma^2 s^2 \log(1 + n/s^2) + \sigma^2 sp \Big].$$

Therefore, the greedy subset selector in Section 2.1 is provably rate-optimal in the case $s = O(\sqrt{n})$. A question that remains open is the rate optimality when $\sqrt{n} = O(s)$. The lower bound of Theorem 5 is then of order $\sigma^2(n + sp)$, whereas the upper bound of Theorem 1 is of order $\sigma^2(s^2 + sp)$. Taking into account the fact that the naive estimator $\boldsymbol{L}(\mathbf{Y})$ has a risk of order $\sigma^2 np$, we get that the minimax risk is upper bounded by $\sigma^2(s^2 \wedge np + sp)$.Thus, there is a gap of order p when $p + \sqrt{n} = O(s)$.

Note that none of the estimators discussed earlier in this work attain the upper bound $\sigma^2(s^2 \wedge np + sp)$; indeed, the latter is obtained as the minimum of the risk of two estimators. Interestingly, one can design a single estimator that attains this rate. Previous sections contain all the necessary ingredients for this. We will illustrate the trick in the case of the GSS estimator, but similar technique can be applied to any estimator for which an "in deviation" risk bound is established.

The idea is to combine the GSS estimator and the naive estimator $\widehat{\boldsymbol{L}} = \boldsymbol{L}(\mathbf{Y})$, with the aim of choosing the "best" one. The combination can be performed using the Lepski method (Lepskii, 1991), also known as intersection of confidence intervals (Goldenshluger and Nemirovski, 1997). The method is described in Algorithm 2. The construction is based on the following two facts:

1. The true value $\boldsymbol{L}(\boldsymbol{\Theta})$ lies with probability $1 - \delta/2$ in the ball $\mathcal{B}(\boldsymbol{L}(\mathbf{Y}); r_1)$ with $(r_1/\sigma)^2 = 2np + 3n \log(2/\delta)$.
2. The true value $\boldsymbol{L}(\boldsymbol{\Theta})$ lies with probability $1 - \delta/2$ in the ball $\mathcal{B}(\widehat{\boldsymbol{L}}^{\mathrm{GSS}}; r_2)$ with $(r_2/\sigma)^2 = 60s(p + \lambda s)$ (cf. Theorem 1).

Algorithm 2: Adaptive GSS

input : matrix $\mathbf{Y}$, noise variance σ^2 and confidence level δ.
output: vector $\widehat{\boldsymbol{L}}^{\mathrm{adGSS}}$.

Set dist $\leftarrow \sigma^{-1} \|\widehat{\boldsymbol{L}}^{\mathrm{GSS}} - \boldsymbol{L}(\mathbf{Y})\|_2$.
Set $\lambda \leftarrow 3/2 \log(4n/\delta)$.
Set $\widehat{s} \leftarrow \min \Big\{ k \in [n] : \text{dist} \le \sqrt{60k(p + \lambda k)} + \sqrt{n(2p + 3\log(2/\delta))} \Big\}$ (if the set is empty, set $\widehat{s} \leftarrow n$)
if $60\,\widehat{s}(p + \lambda \widehat{s}) \le 2np + 3n \log(2/\delta)$ **then**
  $\widehat{\boldsymbol{L}}^{\mathrm{adGSS}} \leftarrow \widehat{\boldsymbol{L}}^{\mathrm{GSS}}$
else
  $\widehat{\boldsymbol{L}}^{\mathrm{adGSS}} \leftarrow \boldsymbol{L}(\mathbf{Y})$
end if
return $\widehat{\boldsymbol{L}}^{\mathrm{adGSS}}$

These two facts imply that with probability at least $1 - \delta$ the balls $\mathcal{B}(\boldsymbol{L}(\mathbf{Y}); r_1)$ and $\mathcal{B}(\widehat{\boldsymbol{L}}^{\mathrm{GSS}}; r_2)$ have nonempty intersection. As a consequence, in this event,

we have $\|\boldsymbol{L}(\mathbf{Y}) - \widehat{\boldsymbol{L}}^{\mathrm{GSS}}\|_2 \le r_1 + r_2$ and, therefore, $\widehat{s} \le s$. Now, if $60\widehat{s}(p + \lambda\widehat{s}) \le 2np + 3n\log(2/\delta)$, then $\widehat{\boldsymbol{L}}^{\mathrm{adGSS}} = \widehat{\boldsymbol{L}}^{\mathrm{GSS}}$ and we have

$$\|\widehat{\boldsymbol{L}}^{\mathrm{adGSS}} - \boldsymbol{L}(\boldsymbol{\Theta})\|_2 = \|\widehat{\boldsymbol{L}}^{\mathrm{GSS}} - \boldsymbol{L}(\boldsymbol{\Theta})\|_2 \le r_2$$

along with

$$\begin{aligned}\|\widehat{\boldsymbol{L}}^{\mathrm{adGSS}} - \boldsymbol{L}(\boldsymbol{\Theta})\|_2 &= \|\widehat{\boldsymbol{L}}^{\mathrm{GSS}} - \boldsymbol{L}(\mathbf{Y})\|_2 + \|\boldsymbol{L}(\mathbf{Y}) - \boldsymbol{L}(\boldsymbol{\Theta})\|_2 \\ &\le \{\sigma\sqrt{60\widehat{s}(p + \lambda\widehat{s})} + r_1\} + r_1 \le 3r_1.\end{aligned}$$

Thus, $\|\widehat{\boldsymbol{L}}^{\mathrm{adGSS}} - \boldsymbol{L}(\boldsymbol{\Theta})\|_2 \le 3(r_1 \wedge r_2)$. In the second case, $60\widehat{s}(p + \lambda\widehat{s}) \ge 2np + 3n\log(2/\delta)$, we have $\|\widehat{\boldsymbol{L}}^{\mathrm{adGSS}} - \boldsymbol{L}(\boldsymbol{\Theta})\|_2 = \|\boldsymbol{L}(\mathbf{Y}) - \boldsymbol{L}(\boldsymbol{\Theta})\|_2 \le r_1 = r_1 \wedge r_2$, where the last equality follows from the fact that $\widehat{s} \le s$. Thus, we have established the following result.

Proposition 4. *Let $\delta \in (0,1)$ be a prescribed confidence level. With probability at least $1 - \delta$, the adaptive greedy subset selection estimator $\widehat{\boldsymbol{L}}^{\mathrm{adGSS}}$ defined in Algorithm 2 satisfies $\|\widehat{\boldsymbol{L}}^{\mathrm{adGSS}} - \boldsymbol{L}(\boldsymbol{\Theta})\|_2 \le 3\sigma\big\{(60sp + 90s^2\log(4n/\delta)) \wedge (2np + 3n\log(2/\delta))\big\}^{1/2}$.*

Let us summarize the content of this section. We have established a lower bound on the minimax risk, showing that the latter is at least of order $sp + s^2 \wedge n$, up to a logarithmic factor. We have also obtained upper bounds, which imply that the minimax risk is at most of order $sp + s^2 \wedge (np)$. Furthermore, this rate can be attained by a single estimator (adaptive greedy subset selection).

3. The problem of robust estimation

The problem of linear functional estimation considered in the previous section has multiple connections with the problem of robust estimation of a Gaussian mean. In the latter problem, the observations $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$ in $\mathbb{R}^p$ are assumed to satisfy

$$\boldsymbol{Y}_i = \boldsymbol{\mu} + \boldsymbol{\theta}_i + \sigma\boldsymbol{\xi}_i, \quad \boldsymbol{\xi}_i \overset{\mathrm{iid}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_p), \tag{6}$$

where $\mathbf{I}_p$ is the identity matrix of dimension $p \times p$. We are interested in estimating the vector $\boldsymbol{\mu}$, under the assumption that most vectors $\boldsymbol{\theta}_i$ are equal to zero. All the observations $\boldsymbol{Y}_i$ such that $i \in S = \{\ell : \|\boldsymbol{\theta}_\ell\|_2 = 0\}$ are considered as inliers, while all the others are outliers. In this problem, the vectors $\boldsymbol{\theta}_i$ are unknown, but their estimation is not our primary aim. They are rather considered as nuisance parameters. In some cases, it might be helpful to use the matrix notation of (6):

$$\mathbf{Y} = \boldsymbol{\mu}\mathbf{1}_n^\top + \boldsymbol{\Theta} + \sigma\boldsymbol{\Xi}. \tag{7}$$

The obvious connection with the problem considered in the previous section is that if we know that $\boldsymbol{\mu} = \mathbf{0}_p$ in (6), then we recover model (1). This can be expressed in a more formal way as shown in the next proposition.

Proposition 5. *The problem of estimating the linear functional* $\boldsymbol{L}_n(\boldsymbol{\Theta}) = (1/n)\sum_{i\in[n]}\boldsymbol{\theta}_i$ *in model* (7) *is not easier, in the minimax sense, than that of estimating* $\boldsymbol{\mu}$*. More precisely, we have*

$$\frac{\sigma^2 p}{n} \le \inf_{\widehat{\boldsymbol{\mu}}} \sup_{\boldsymbol{\mu},\boldsymbol{\Theta}} \mathbf{E}[\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2^2] \le 2 \inf_{\widehat{\boldsymbol{L}}_n} \sup_{\boldsymbol{\Theta}} \mathbf{E}[\|\widehat{\boldsymbol{L}}_n - \boldsymbol{L}_n(\boldsymbol{\Theta})\|_2^2] + \frac{2\sigma^2 p}{n},$$

where the sup in the left-hand side and in the right-hand side are taken, respectively, over all $\boldsymbol{\Theta} \in \mathcal{M}(p,n,s)$ *and over all* $(\boldsymbol{\mu},\boldsymbol{\Theta}) \in \mathbb{R}^p \times \mathcal{M}(p,n,s)$*.*

Proof. The first inequality is a consequence of the fact that when all the entries of $\boldsymbol{\Theta}$ are zero, the optimal estimator of $\boldsymbol{\mu}$ in the minimax sense is the sample mean of $\boldsymbol{Y}_i$'s. To prove the second inequality, let $\widehat{\boldsymbol{L}}_n$ be an estimator of $\boldsymbol{L}_n(\boldsymbol{\Theta})$. We can associate with $\widehat{\boldsymbol{L}}_n$ the following estimator of $\boldsymbol{\mu}$: $\widehat{\boldsymbol{\mu}}(\widehat{\boldsymbol{L}}_n) = \boldsymbol{L}_n(\mathbf{Y}) - \widehat{\boldsymbol{L}}_n$. These estimators satisfy

$$\begin{aligned}
\mathbf{E}[\|\widehat{\boldsymbol{\mu}}(\widehat{\boldsymbol{L}}_n) - \boldsymbol{\mu}\|_2^2] &= \mathbf{E}[\|\boldsymbol{L}_n(\mathbf{Y}) - \widehat{\boldsymbol{L}}_n - \boldsymbol{\mu}\|_2^2] \\
&= \mathbf{E}[\|\boldsymbol{L}_n(\boldsymbol{\Theta}) + \sigma\boldsymbol{L}_n(\boldsymbol{\Xi}) - \widehat{\boldsymbol{L}}_n\|_2^2] \\
&\le 2\mathbf{E}[\|\boldsymbol{L}_n(\boldsymbol{\Theta}) - \widehat{\boldsymbol{L}}_n\|_2^2] + 2\sigma^2\mathbf{E}[\|\boldsymbol{L}_n(\boldsymbol{\Xi})\|_2^2].
\end{aligned}$$

Since $\boldsymbol{L}_n(\boldsymbol{\Xi})$ is drawn from the Gaussian distribution $\mathcal{N}_p(\mathbf{0},(1/n)\mathbf{I}_p)$, we have $\mathbf{E}[\|\boldsymbol{L}_n(\boldsymbol{\Xi})\|_2^2] = p\sigma^2/n$ and the claim of the proposition follows. □

Another important point that we would like to mention here is the relation between model (6) and the Huber contamination model (Huber, 1964) frequently studied in the statistical literature (we refer the reader to (Chen, Gao and Ren, 2015; Chen, Gao and Ren, 2016) for recent overviews). Recall that in Huber's contamination model, the observations $\boldsymbol{X}_1,\dots,\boldsymbol{X}_n$ are n iid p-dimensional vectors drawn from the mixture distribution $(1-\frac{s}{n})\mathcal{N}_p(\boldsymbol{\mu},\mathbf{I}_p) + \frac{s}{n}\mathbf{Q}$. The particularity of this model is that it assumes all the outliers to be generated by the same distribution $\mathbf{Q}$; the latter, however, can be an arbitrary distribution on $\mathbb{R}^p$. In contrast with this, our model (6) allows for a wider heterogeneity of the outliers. On the downside, our model assumes that the outliers are blurred by a Gaussian noise that has the same covariance structure as the noise that corrupts the inliers. The relation between these two models is formalized in the next result.

Proposition 6. *Let* $\widehat{\boldsymbol{\mu}} : \mathbb{R}^{p\times n} \to \mathbb{R}^p$ *be an estimator of* $\boldsymbol{\mu}$ *that can be applied both to the data matrix* $\mathbf{X} = [\boldsymbol{X}_1,\dots,\boldsymbol{X}_n]$ *from Huber's model and to* $\mathbf{Y}$ *from our model* (6)*. Then, we have*

$$\underbrace{\sup_{\mathbf{Q}} \mathbf{E}_{\boldsymbol{\mu},\mathbf{Q}}[\|\widehat{\boldsymbol{\mu}}(\mathbf{X}) - \boldsymbol{\mu}\|_2^2]}_{\text{risk in the Huber model}} \le \mathbf{E}_{\widehat{s}\sim\mathcal{B}(n,s/n)}\Bigg[\underbrace{\sup_{\boldsymbol{\Theta}\in\mathcal{M}(n,p,\widehat{s})} \mathbf{E}_{\boldsymbol{\mu},\boldsymbol{\Theta}}\big[\|\widehat{\boldsymbol{\mu}}(\mathbf{Y}) - \boldsymbol{\mu}\|_2^2\big]}_{\text{risk in our model (6)}}\Bigg].$$

The supremum of the left-hand side is over all probability distributions $\mathbf{Q}$ *on* $\mathbb{R}^p$ *such that*[3] $\mathbf{Q} = \mathbf{Q}_0 * \mathcal{N}_p(\mathbf{0},\sigma^2\mathbf{I}_p)$*, where* $\mathbf{Q}_0$ *is an arbitrary distribution on* $\mathbb{R}^p$*. The notation* $\mathcal{B}(n,s/n)$ *stands for the binomial distribution.*

[3] We denote by $*$ the convolution of the distributions.

The proof of this proposition is a simple exercise and is left to the reader. Although some statistical properties of robust estimators in a framework of the same spirit as (6) were already explored in the literature (Dalalyan and Keriven, 2012; Dalalyan and Chen, 2012; Balmand and Dalalyan, 2015; Nguyen and Tran, 2013; Klopp, Lounici and Tsybakov, 2017; Cherapanamjeri, Gupta and Jain, 2016), the entire picture in terms of matching upper and lower bounds was not available until very recently. It has been established in (Chen, Gao and Ren, 2015) that the minimax rate of estimating $\boldsymbol{\mu}$ in Huber's contamination model is

$$r_{\mathrm{mmx}}^{\mathrm{all}}(n,p,s) = \sigma^2\Big(\frac{p}{n} \vee \frac{s^2}{n^2}\Big).$$

As one might expect, this rate is proved to be minimax-optimal in our contamination model as well (Carpentier et al., 2018, row 2, Table 1), at least up to a factor poly-logarithmic in s. Furthermore, Chen, Gao and Ren (2015) showed that this rate is achieved by Tukey's median, *i.e.*, the minimizer of Tukey's depth. An important observation is that the evaluation of Tukey's median is a hard computational problem: there exists no algorithm to date capable of approximating Tukey's median in a number of operations that scales polynomially in p, n and the approximation precision. Simple computationally tractable robust estimators of $\boldsymbol{\mu}$, such as the element-wise median or the geometric median, have a rate of order (Chen, Gao and Ren, 2015; Lai, Rao and Vempala, 2016)

$$\sigma^2\Big(\frac{p}{n} \vee \frac{s^2 p}{n^2}\Big)$$

containing an extra factor p in front of $(s/n)^2$. We shall show in this section that a suitable adaptation of the group soft thresholding estimator presented in the previous section leads to a rate that can be arbitrarily close to

$$\sigma^2\Big(\frac{p}{n} \vee \frac{s^2}{n^2} \vee \frac{s^4 p}{n^4}\Big).$$

This shows that if we restrict our attention to the estimators that have a computational complexity that is at most polynomial, the minimax rate satisfies, for every $\nu \in (0, 1/\log p)$,

$$\sigma^2\Big(\frac{p}{n} \vee \frac{s^2}{n^2}\Big) \lesssim r_{\mathrm{mmx}}^{\mathrm{poly}}(n,p,s) \lesssim \sigma^2\Big(\frac{p}{n} \vee \frac{s^2}{n^2} \vee \Big\{\frac{s^4 p}{n^4}\Big\}^{1-\nu}\Big),$$

where $\lesssim$ means inequality up to logarithmic factors.

3.1. Maximum of profile likelihood with group LASSO penalty

A computationally tractable estimator that allows to efficiently deal with structured sparsity and has provably good statistical complexity is the group LASSO (Yuan and Lin, 2006; Lin and Zhang, 2006; Chesneau and Hebiri, 2008; Meier,

van de Geer and Bühlmann, 2009; Lounici et al., 2011). We define the group-LASSO estimator by

$$(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Theta}}) \in \arg\min_{\boldsymbol{m},\mathbf{T}} \Big\{ \sum_{i=1}^{n} \|\boldsymbol{Y}_i - \boldsymbol{m} - \boldsymbol{t}_i\|_2^2 + \sum_{i=1}^{n} \lambda_i \|\boldsymbol{t}_i\|_2 \Big\}. \tag{8}$$

where the λ_i are some positive numbers to be defined later. The estimator $\widehat{\boldsymbol{\mu}}$ can be seen as the maximum of a profile penalized likelihood, where the penalty is proportional to the $\ell_{2,1}$ norm (also known as the group LASSO penalty) of the nuisance parameter $\boldsymbol{\Theta}$. The above optimization problem is convex and can be solved numerically even when the dimension and the sample size are large. It is also well known that $\widehat{\boldsymbol{\mu}}$ from (8) is exactly the Huber M-estimator (Donoho and Montanari, 2016, Section 6). In addition, these estimators can also be written as

$$\begin{aligned} \widehat{\boldsymbol{\mu}} &= \frac{1}{n} \sum_{i=1}^{n} \big(\boldsymbol{Y}_i - \widehat{\boldsymbol{\theta}}_i\big) = \boldsymbol{L}_n(\mathbf{Y}) - \boldsymbol{L}_n(\widehat{\boldsymbol{\Theta}}), \\ \widehat{\boldsymbol{\Theta}} &\in \arg\min_{\mathbf{T}} \Big\{ \sum_{j=1}^{p} \|\boldsymbol{\Pi}(\boldsymbol{Y}^j - \boldsymbol{t}^j)\|_2^2 + \sum_{i=1}^{n} \lambda_i \|\boldsymbol{t}_i\|_2 \Big\}, \end{aligned} \tag{9}$$

where $\boldsymbol{\Pi}$ denotes the orthogonal projection in $\mathbb{R}^n$ onto the orthogonal complement of the constant vector $\mathbf{1}_n$. Unfortunately, we were unable to establish a risk bound for this estimator that improves on the element-wise median. The best result that we get is the following.

Theorem 6. *Consider the estimators of $\boldsymbol{\Theta}$ and $\boldsymbol{\mu}$ defined in (8) with $\lambda^2 = 32\sigma^2 p + 256\sigma^2 \log(n/\delta)$. Then, with probability at least $1-\delta$ and provided that $s \le n/32$, we have*

$$\|\boldsymbol{\Theta} - \widehat{\boldsymbol{\Theta}}\|_F^2 \le 9s\lambda^2, \quad \|\boldsymbol{L}_n(\widehat{\boldsymbol{\Theta}}) - \boldsymbol{L}_n(\boldsymbol{\Theta})\|_2^2 \le \frac{288 s^2 \lambda^2}{n^2}$$
$$\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2^2 \le \frac{288 s^2 \lambda^2}{n^2} + \frac{4\sigma^2 p}{n} + \frac{8\sigma^2 \log(2/\delta)}{n}.$$

This result, proved in Section 5.2, shows that the rate of the profiled penalized likelihood estimator of $\boldsymbol{\mu}$, with a group LASSO penalty, converges at the rate $\sigma^2\big(\frac{s^2 p}{n^2} \vee \frac{p}{n}\big)$, which coincides with the one obtained[4] by (Chen, Gao and Ren, 2015). In the rest of this section, we will propose an estimator which improves on this rate. To this end, we start with obtaining a simplified expression for the group LASSO estimator $\widehat{\boldsymbol{\Theta}}$.

First, using the fact that $\boldsymbol{\Pi}\boldsymbol{t}^j = \boldsymbol{t}^j - (\mathbf{I}_n - \boldsymbol{\Pi})\boldsymbol{t}^j$, we get $\|\boldsymbol{\Pi}(\boldsymbol{Y}^j - \boldsymbol{t}^j)\|_2^2 = \|\boldsymbol{\Pi}\boldsymbol{Y}^j - \boldsymbol{t}^j\|_2^2 - (1/n)\big(\mathbf{1}_n^\top \boldsymbol{t}^j\big)^2$, so that

$$\widehat{\boldsymbol{\Theta}} \in \arg\min \Big\{ \sum_{i=1}^{n} \|(\mathbf{Y}\boldsymbol{\Pi})_i - \boldsymbol{t}_i\|_2^2 - \frac{1}{n}\Big\| \sum_{i=1}^{n} \boldsymbol{t}_i \Big\|_2^2 + \sum_{i=1}^{n} \lambda_i \|\boldsymbol{t}_i\|_2 \Big\}.$$

[4] To be precise, (Chen, Gao and Ren, 2015) establish only a lower bound for the element-wise median, but a matching upper bound can be proved as well.

Recall that $\boldsymbol{L}_n(\boldsymbol{\Theta}) = (1/n)\boldsymbol{L}(\boldsymbol{\Theta})$. The first-order necessary conditions imply that, for every i such that $\widehat{\boldsymbol{\theta}}_i \neq \mathbf{0}_p$,

$$-2\big((\mathbf{Y}\boldsymbol{\Pi})_i - \widehat{\boldsymbol{\theta}}_i\big) - 2\boldsymbol{L}_n(\widehat{\boldsymbol{\Theta}}) + \frac{\lambda_i \widehat{\boldsymbol{\theta}}_i}{\|\widehat{\boldsymbol{\theta}}_i\|_2} = \mathbf{0}_p.$$

Furthermore, $\widehat{\boldsymbol{\theta}}_i = \mathbf{0}_p$ if and only if $\big\|2(\mathbf{Y}\boldsymbol{\Pi})_i + 2\boldsymbol{L}_n(\widehat{\boldsymbol{\Theta}})\big\|_2 \le \lambda_i$. We infer that

$$\widehat{\boldsymbol{\theta}}_i = \frac{(\mathbf{Y}\boldsymbol{\Pi})_i + \boldsymbol{L}_n(\widehat{\boldsymbol{\Theta}})}{\big\|(\mathbf{Y}\boldsymbol{\Pi})_i + \boldsymbol{L}_n(\widehat{\boldsymbol{\Theta}})\big\|_2} \left(\big\|(\mathbf{Y}\boldsymbol{\Pi})_i + \boldsymbol{L}_n(\widehat{\boldsymbol{\Theta}})\big\|_2 - \frac{\lambda_i}{2}\right)_+$$

for every i. Finally, denoting $\boldsymbol{Z}_i = (\mathbf{Y}\boldsymbol{\Pi})_i + \boldsymbol{L}_n(\widehat{\boldsymbol{\Theta}})$, we get

$$\widehat{\boldsymbol{\theta}}_i = \boldsymbol{Z}_i\Big(1 - \frac{\lambda_i}{2\|\boldsymbol{Z}_i\|_2}\Big)_+.$$

This formula shows the clear analogy between the group LASSO estimator $\widehat{\boldsymbol{\Theta}}$ and the soft thresholding estimator studied in the previous section. This analogy suggests to choose the tuning parameters in a data driven way; namely, it is tempting to set

$$\lambda_i = \frac{2\gamma\sigma\|\boldsymbol{Z}_i\|_2}{(\|\boldsymbol{Z}_i\|_2^2 - \sigma^2 p)_+^{1/2}} \qquad \Longrightarrow \qquad \widehat{\boldsymbol{\theta}}_i = \boldsymbol{Z}_i\left(1 - \frac{\gamma\sigma}{(\|\boldsymbol{Z}_i\|_2^2 - \sigma^2 p)_+^{1/2}}\right)_+. \tag{10}$$

Unfortunately, such a choice is impossible to realize since this λ_i depends on the solution $\widehat{\boldsymbol{\Theta}}$ of the optimization problem, which in turn is defined through λ_i. To circumvent this problem, we suggest to use an iterative algorithm that starts from an initial estimator $\widehat{\boldsymbol{L}}_n$ of $\boldsymbol{L}_n(\boldsymbol{\Theta})$, defines the vectors $\boldsymbol{Z}_i = (\mathbf{Y}\boldsymbol{\Pi})_i + \widehat{\boldsymbol{L}}_n$ and then updates $\widehat{\boldsymbol{L}}_n$ by the formula $\widehat{\boldsymbol{L}}_n = \boldsymbol{L}_n(\widehat{\boldsymbol{\Theta}})$, where the columns of the matrix $\widehat{\boldsymbol{\Theta}}$ are defined by the second equality in (10). This algorithm, called iterative soft thresholding, is described in Algorithm 3.

Prior to stating the theorem that describes the statistical complexity of this estimator, we present a result that explains why such an iterative algorithm succeeds in improving the convergence rate.

Proposition 7. *Let us set $\boldsymbol{Z}_i = (\mathbf{Y}\boldsymbol{\Pi})_i + \widetilde{\boldsymbol{L}}_n$, where $\widetilde{\boldsymbol{L}}_n$ is a preliminary estimator of $\boldsymbol{L}_n(\boldsymbol{\Theta})$. Let $\delta \in (0,1)$ be a tolerance level. Consider the estimator of $\boldsymbol{\Theta}$ defined by concatenating the vectors*

$$\widehat{\boldsymbol{\theta}}_i = \boldsymbol{Z}_i\left(1 - \frac{\sigma\gamma}{(\|\boldsymbol{Z}_i\|_2^2 - \frac{n-1}{n}\sigma^2 p)_+^{1/2}}\right)_+, \tag{11}$$

where $\gamma^2 > 4\log(4n/\delta) + 4\{p\log(4n/\delta)\}^{1/2}$ is a tuning parameter. Define the event

$$\Omega_1 = \left\{\|\widetilde{\boldsymbol{L}}_n - \boldsymbol{L}_n(\boldsymbol{\Theta})\|_2 < \frac{\sigma\gamma^2}{4\sqrt{p+\gamma^2}}\right\}.$$

Algorithm 3: Iterative Soft Thresholding

input : matrix $\mathbf{Y}$, noise variance σ, number of outliers s.
parameters: number of iterations N, confidence level δ.
output : vectors $\widehat{\boldsymbol{L}}_n^{\mathrm{IST}}$ and $\widehat{\boldsymbol{\mu}}^{\mathrm{IST}}$.

initialization
  $\widehat{\boldsymbol{\Theta}} \leftarrow$ solution of (9) with $\lambda^2 = 32\sigma^2(p + 8\log(n/\delta))$
  $\epsilon \leftarrow \sqrt{288}\,\dfrac{s\lambda}{n\sigma}$
for $k = 1, \dots, N$ **do**
  Set $\widehat{\boldsymbol{L}}_n \leftarrow \boldsymbol{L}_n(\widehat{\boldsymbol{\Theta}})$
  Set $\gamma^2 \leftarrow 8\epsilon^2 + 4\sqrt{4\epsilon^4 + p\epsilon^2}$
  for $i = 1, \dots, n$ **do**
    Set $\boldsymbol{Z}_i \leftarrow (\mathbf{Y}\boldsymbol{\Pi})_i + \widehat{\boldsymbol{L}}_n$
    Set $\boldsymbol{\theta}_i \leftarrow \boldsymbol{Z}_i\Big(1 - \dfrac{\sigma\gamma}{(\|\boldsymbol{Z}_i\|_2^2 - \frac{n-1}{n}\sigma^2 p)_+^{1/2}}\Big)_+$.
  end for
  update
    $\widehat{\boldsymbol{\Theta}} \leftarrow [\widehat{\boldsymbol{\theta}}_1, \dots, \widehat{\boldsymbol{\theta}}_n]$.
    $\epsilon \leftarrow {}^4\!/_n\,(\,s\gamma + s + \sqrt{sp} + \{2s\log(4/\delta)\}^{1/2})$
end for
return $\widehat{\boldsymbol{L}}_n^{\mathrm{IST}} \leftarrow \boldsymbol{L}_n(\widehat{\boldsymbol{\Theta}})$ *and* $\widehat{\boldsymbol{\mu}}^{\mathrm{IST}} = \boldsymbol{L}_n(\mathbf{Y}) - \widehat{\boldsymbol{L}}_n^{\mathrm{IST}}$.

There is an event $\bar{\Omega}$ (the same for all estimators $\widetilde{\boldsymbol{L}}_n$) of probability at least $1-\delta$, such that on $\Omega_1 \cap \bar{\Omega}$, we have

$$\|\boldsymbol{L}(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})\|_2 \le 4\sigma\Big(\,s\gamma + s + \sqrt{sp} + \{2s\log(4/\delta)\}^{1/2}\Big).$$

It follows from this theorem that at each iteration of the algorithm we improve the precision of estimation of $\boldsymbol{L}_n(\boldsymbol{\Theta})$. Indeed, if ϵ_k is an upper bound on the error $\|\widehat{\boldsymbol{L}}_n{}^{(k)} - \boldsymbol{L}_n(\boldsymbol{\Theta})\|_2/\sigma$ at the kth iteration, then we get from the last theorem that

$$\epsilon_{k+1} \le \frac{8s}{n}\big(2\epsilon_k^2 + (4\epsilon_k^4 + p\epsilon_k^2)^{1/2}\big)^{1/2} + a, \tag{12}$$

with $a = (4/n)(s + \sqrt{sp} + \{2s\log(4/\delta)\}^{1/2})$.

Lemma 1. *If $\epsilon_0^2 \le p$, $n \ge 33s$ and $a \le 0.5\sqrt{p}$, then*

$$\epsilon_k \le \Big\{p^{1/2}\Big(\frac{33^2 s^2}{n^2}\Big)^{1-(1/2)^k}\Big\} \vee 2a. \tag{13}$$

Combining all these results, we arrive at the following risk bound for the iterative soft thresholding estimator.

Theorem 7. *Let $\delta \in (0,1)$, $N \in \mathbb{N}$ and let $\widehat{\boldsymbol{L}}_n^{\mathrm{IST}}(N)$ be the iterative soft thresholding estimator obtained after N iterations. Assume that $p \ge \log(8/\delta)$ and $N \ge \log\log p$. There are some universal strictly positive constants c_1, c_2, c_3 such that if the condition*

$$s \le c_1 n$$

is satisfied then, with probability at least $1-2\delta$*, the following inequalities hold true:*

$$\|\widehat{\boldsymbol{L}}_n^{\mathrm{IST}}(N)-\boldsymbol{L}_n(\boldsymbol{\Theta})\|_2 \le c_2\sigma\left\{p^{1/2}\Big(\frac{s^2}{n^2}\Big)^{1-2^{-N}}+\frac{s+\sqrt{sp}}{n}\right\},$$
$$\|\widehat{\boldsymbol{\mu}}^{\mathrm{IST}}-\boldsymbol{\mu}\|_2 \le c_3\sigma\left\{p^{1/2}\Big(\frac{s^2}{n^2}\Big)^{1-2^{-N}}+\frac{s}{n}+\Big(\frac{p}{n}\Big)^{1/2}\right\}.$$

This implies, in particular, that if $p \le C(n/s)^{2-4\nu}$ for some $\nu \in (0,1/2)$ close to zero, then performing $N=\log_2(1/\nu)$ iterations of the IST algorithm we will recover the mean $\boldsymbol{\mu}$ of the inliers at an optimal rate $(s/n)^2 \vee (p/n)$.

To complete this section, let us briefly note that one can use the Lepski method as described in Section 2.4 for getting an estimator of $\boldsymbol{\mu}$ that does not require the knowledge of s. This will only increase the error by a factor at worst equal to 3.

Remark 1. From an intuitive point of view, the algorithm described in Algorithm 3 can be seen as an iterative approximation of the estimator

$$\widehat{\boldsymbol{\mu}}^* \in \arg\min_{\boldsymbol{\mu}\in\mathbb{R}^p}\sum_{i=1}^n \rho_{\mathrm{H}}\left(\frac{(\|\boldsymbol{Y}_i-\boldsymbol{\mu}\|_2^2-\sigma^2 p)_+^{1/2}}{\sigma\gamma}\right), \tag{14}$$

for an appropriately chosen tuning parameter $\gamma>0$, where ρ_{H} is the Huber function. Unfortunately, the cost function in the above minimization problem is not convex with respect to the parameter $\boldsymbol{\mu}$. This implies that general purpose guarantees available for approximating solutions of convex programs are not applicable to (14). To the best of our knowledge, there is no efficient algorithm that provably approximates $\widehat{\boldsymbol{\mu}}^*$.

4. Conclusion and perspectives

In this work, we have studied two problems: the problem of estimating a multidimensional linear functional and the one of estimating the mean of p-variate random vectors when the data is corrupted by outliers. In the first problem, we have obtained upper and lower bounds on the minimax risk that match in most situations. More importantly, in both problems, we have studied computationally tractable estimators and have obtained the best known rates of convergence. A surprising outcome of our work is that exploiting the group structure of the sparsity is far more important in the problem of linear functional estimation rather than in the problem of the whole signal. We have also designed a new robust estimator of the mean that iteratively performs group soft thresholding on a suitable transformation of the data.

There are several questions related to the present work that remain open. First, it would be interesting to close the gap in the minimax rate of estimation of a linear functional when $p+\sqrt{n}=O(s)$. Second, in both problems studied in this work, a challenging question for future research is to establish lower

bounds on the minimax risk over computationally tractable estimators. For the problem of robust estimation, one may use a suitable version of the median of means (Lerasle and Oliveira, 2011; Minsker, 2015; Devroye et al., 2016; Lecué and Lerasle, 2017). However, in the multidimensional Gaussian model considered in the present work, to our knowledge there are no upper bounds on the risk of these methods that are qualitatively better than those for the element-wise median, in the sense of their dependence on the proportion of outliers.

There are also some recent papers studying the problem of robust mean estimation in the case of adversarial contamination (Lai, Rao and Vempala, 2016; Diakonikolas et al., 2016; Balakrishnan et al., 2017). Note that the methods used in those papers are quite different from ours. In particular, they are all more computationally demanding in that their iterations perform a singular values decomposition of a $p \times p$ matrix. Furthermore, the rates of convergence for those methods have always some extra poly-logarithmic factors as compared to the minimax rate.

5. Proofs of the main results

This section contains the proofs of the main theorems stated in the previous sections.

5.1. Proofs of the theorems of Section 2

Proof of Theorem 1. Using the triangle inequality several times, we get

$$\begin{aligned}\|\widehat{\boldsymbol{L}}^{\mathrm{GSS}} - \boldsymbol{L}(\boldsymbol{\Theta})\|_2 &\le \|\boldsymbol{L}(\mathbf{Y}_{\widehat{S}}) - \boldsymbol{L}(\mathbf{Y}_S)\|_2 + \|\boldsymbol{L}(\mathbf{Y}_S) - \boldsymbol{L}(\boldsymbol{\Theta}_S)\|_2 \\ &\le \|\boldsymbol{L}(\mathbf{Y}_{\widehat{S}\setminus S})\|_2 + \|\boldsymbol{L}(\mathbf{Y}_{S\setminus\widehat{S}})\|_2 + \sigma\|\boldsymbol{L}(\boldsymbol{\xi}_S)\|_2 \\ &\le \sigma\|\boldsymbol{L}(\boldsymbol{\xi}_{\widehat{S}\setminus S})\|_2 + \|\boldsymbol{L}(\mathbf{Y}_{S\setminus\widehat{S}})\|_2 + \sigma\|\boldsymbol{L}(\boldsymbol{\xi}_S)\|_2. \qquad (15)\end{aligned}$$

To upper bound the three terms of the right hand side, we introduce the event

$$\Omega_\lambda = \Big\{\|\boldsymbol{L}(\boldsymbol{\xi}_J)\|_2^2 \le 2|J|(p + \lambda|J|) \text{ for all } J \subseteq [n]\Big\}.$$

We will show that the following three claims are true for the tuning parameter λ chosen as in the statement of the theorem.

Claim 1: On the event Ω_λ, at least half of the elements of each $\widehat{J}_\ell$ belong to the true sparsity pattern S. Thus $|\widehat{S}| \le 2s$.

Claim 2: $\|\boldsymbol{L}(\mathbf{Y}_{S\setminus\widehat{S}})\|_2^2 \le 12\sigma^2 s(p + \lambda s)$.

Claim 3: The probability of Ω_λ is close to 1.

Let us first show that these claims imply the claim of the theorem. Indeed, the second term of the right hand side of (15) is bounded by $\sigma\sqrt{12s(p + \lambda s)}$ in view

of Claim 2. The third term is bounded by $\sigma\sqrt{2s(p+\lambda s)}$ on the event Ω_λ. Concerning the first term, we know that on Ω_λ it is bounded by $\sigma\sqrt{2|\widehat{S}|(p+\lambda|\widehat{S}|)}$. In view of Claim 1, $|\widehat{S}| \le 2s$. All these inequalities imply that

$$\|\widehat{\boldsymbol{L}}^{\mathrm{GSS}} - \boldsymbol{L}(\boldsymbol{\Theta})\|_2^2 \le 60\sigma^2 s(p+\lambda s)$$

on the event Ω_λ. This is exactly the claim of the theorem.

Let us prove now Claims 1-3. To prove the first claim, let us assume that there is a set J among $\widehat{J}_1, \dots, \widehat{J}_\ell$ and a subset $J_0 \subset J$ of cardinality[5] $|J|/2$ such that $J_0 \subset S^c$. This readily implies that $\|\boldsymbol{L}(\mathbf{Y}_J)\|_2^2 \ge 12\sigma^2|J|(p+\lambda|J|)$ and $\|\boldsymbol{L}(\mathbf{Y}_{J\setminus J_0})\|_2^2 < 6\sigma^2|J|(p+\lambda|J|/2)$. Using the additivity of $\boldsymbol{L}$ and the triangle inequality, we get

$$\begin{aligned}
\|\boldsymbol{L}(\mathbf{Y}_J)\|_2^2 &\le \big(\|\boldsymbol{L}(\mathbf{Y}_{J\setminus J_0})\|_2 + \sigma\|\boldsymbol{L}(\boldsymbol{\xi}_{J_0})\|_2\big)^2 \\
&\le \frac{3}{2}\|\boldsymbol{L}(\mathbf{Y}_{J\setminus J_0})\|_2^2 + 3\sigma^2\|\boldsymbol{L}(\boldsymbol{\xi}_{J_0})\|_2^2 \\
&< 9\sigma^2|J|(p+\lambda|J|/2) + 3\sigma^2|J|(p+\lambda|J|/2) \\
&< 12\sigma^2|J|(p+\lambda|J|).
\end{aligned}$$

This is in contradiction with the fact that J is one of the sets $\widehat{J}_1, \dots, \widehat{J}_L$. So, Claim 1 is proved.

The proof of Claim 2 is simpler. By construction, the set $S \setminus \widehat{S}$ is a subset of I_L, where L is the number of steps performed by the algorithm. Since the algorithm terminated after the Lth step, this means that $\mathcal{I}$ was empty, which implies that $\|\boldsymbol{L}(\mathbf{Y}_{S\setminus\widehat{S}})\|_2^2 \le 12\sigma^2|S\setminus\widehat{S}|(p+\lambda|S\setminus\widehat{S}|) \le 12\sigma^2 s(p+\lambda s)$.

It remains to prove Claim 3. This can be done using the union bound and tail bounds for χ_p^2-distributed random variables. Indeed, we have

$$\begin{aligned}
\mathbf{P}_{\boldsymbol{\Theta}}\big(\Omega_\lambda^c\big) &\le \sum_{k=1}^n \mathbf{P}_{\boldsymbol{\Theta}}\big(\exists J \subset [n] \text{ s.t. } |J|=k \text{ and } \|\boldsymbol{L}(\boldsymbol{\xi}_J)\|_2^2 > 2|J|(p+\lambda|J|)\big) \\
&\le \sum_{k=1}^n \binom{n}{k} \max_{J:|J|=k} \mathbf{P}_{\boldsymbol{\Theta}}\big(\|\boldsymbol{L}(\boldsymbol{\xi}_J)\|_2^2 > 2k(p+\lambda k)\big) \\
&\le \sum_{k=1}^n \binom{n}{k} \mathbf{P}\big(\eta > 2p + 2\lambda k\big)
\end{aligned}$$

where $\eta \sim \chi_p^2$. Using the well known bound on the tails of the χ_p^2 distribution, we get

$$\mathbf{P}\big(\Omega_\lambda^c\big) \le \sum_{k=1}^n \binom{n}{k} e^{-2\lambda k/3} = (1+e^{-2\lambda/3})^n - 1.$$

[5] To avoid uninteresting and irrelevant technicalities, we assume here that $|J|$ is even.

Therefore, for $\lambda = 3/2 \log(2n/\delta)$, we obtain that $\mathbf{P}_{\boldsymbol{\Theta}}\big(\Omega_\lambda^c\big) \le \delta$. This completes the proof of Claim 3 and of the theorem. □

Proof of Theorem 2. Recall that $\boldsymbol{\Xi} = (\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_n)$. First, denoting by S_λ the set of indices such that $\|\boldsymbol{Y}_i\|_2 \ge \lambda$, we decompose

$$\begin{aligned}\widehat{\boldsymbol{L}}^{\mathrm{GHT}} - \boldsymbol{L}(\boldsymbol{\Theta}) &= \boldsymbol{L}(\boldsymbol{Y}_S - \boldsymbol{\Theta}) - \boldsymbol{L}(\boldsymbol{Y}_{S\setminus S_\lambda}) + \boldsymbol{L}(\boldsymbol{Y}_{S_\lambda\setminus S}) \\ &= \sigma \boldsymbol{L}(\boldsymbol{\Xi}_S) - \boldsymbol{L}(\boldsymbol{Y}_{S\setminus S_\lambda}) + \sigma \boldsymbol{L}(\boldsymbol{\Xi}_{S_\lambda\setminus S}) \\ &= \sigma \boldsymbol{L}(\boldsymbol{\Xi}_{S\cap S_\lambda}) - \boldsymbol{L}(\boldsymbol{\Theta}_{S\setminus S_\lambda}) + \sigma \boldsymbol{L}(\boldsymbol{\Xi}_{S_\lambda\setminus S}). \end{aligned} \tag{16}$$

so that

$$\big\|\widehat{\boldsymbol{L}}^{\mathrm{GHT}} - \boldsymbol{L}(\boldsymbol{\Theta})\big\|_2 \le \sigma \|\boldsymbol{L}(\boldsymbol{\Xi}_{S\cap S_\lambda})\|_2 + \|\boldsymbol{L}(\boldsymbol{\Theta}_{S\setminus S_\lambda})\|_2 + \sigma \|\boldsymbol{L}(\boldsymbol{\Xi}_{S_\lambda\setminus S})\|_2. \tag{17}$$

The first term corresponds to the stochastic error of estimating the signal vectors that are correctly identified as nonzero. We can write

$$\|\boldsymbol{L}(\boldsymbol{\Xi}_{S\cap S_\lambda})\|_2 = \|\boldsymbol{\Xi}_S \mathbf{1}_{S\cap S_\lambda}\|_2 \le \sqrt{s}\|\boldsymbol{\Xi}_S\|.$$

The second-order moment of the spectral norm of the random matrix $\boldsymbol{\Xi}_S$ can be evaluated using well-known upper bounds on the spectral norm of matrices with independent Gaussian, recalled in Lemma 9 below, so that

$$\mathbf{E}_{\boldsymbol{\Theta}}\big[\|\boldsymbol{L}(\boldsymbol{\Xi}_{S\cap S_\lambda})\|_2^2\big] \le 3s^2 + 3sp + 12s. \tag{18}$$

Set $\eta_i = \boldsymbol{\theta}_i^\top \boldsymbol{\xi}_i / \|\boldsymbol{\theta}_i\|_2$. We can control the second term in (17) using the following inequality

$$\begin{aligned}\|\boldsymbol{L}(\boldsymbol{\Theta}_{S\setminus S_\lambda})\|_2 &\le \sum_{i\in S} \|\boldsymbol{\theta}_i\|_2 \mathbb{1}_{\|\boldsymbol{\theta}_i\|_2^2 < \lambda^2 - 2\sigma\boldsymbol{\theta}_i^\top\boldsymbol{\xi}_i - \sigma^2\|\boldsymbol{\xi}_i\|_2^2} \\ &\le s(\lambda^2 - \sigma^2 p)^{1/2} + 2\sigma \sum_{i\in S} |\eta_i| + \sigma \sum_{i\in S} \big|\|\boldsymbol{\xi}_i\|_2^2 - p\big|^{1/2},\end{aligned}$$

since the condition $\|\boldsymbol{\theta}_i\|^2 < \lambda^2 - 2\sigma\boldsymbol{\theta}_i^\top\boldsymbol{\xi}_i - \sigma^2\|\boldsymbol{\xi}_i\|_2^2$ implies that

$$\|\boldsymbol{\theta}_i\|^2 < |\lambda^2 - \sigma^2 p| + 2\sigma\|\boldsymbol{\theta}_i\|_2|\eta_i| + \sigma^2|\|\boldsymbol{\xi}_i\|_2^2 - p|,$$

and, if $a, b > 0$, then by a simple resolution

$$x^2 < ax + b \quad \Rightarrow \quad x < a + \sqrt{b}.$$

This readily yields

$$\mathbf{E}_{\boldsymbol{\Theta}}\big[\|\boldsymbol{L}(\boldsymbol{\Theta}_{S\setminus S_\lambda})\|_2^2\big]^{1/2} \le s(\lambda^2 - \sigma^2 p)^{1/2} + 2\sigma s + \sigma s(2p)^{1/4}. \tag{19}$$

The third term in (17) corresponds to the Type II error in the problem of support estimation. Denoting $t = (\lambda^2 - \sigma^2 p)/\sigma^2$, and using tail bounds for the chi-squared random variables (see Lemma 6 below), we get

$$\begin{aligned}\mathbf{E}_{\boldsymbol{\Theta}}[\|\boldsymbol{\Xi}_{S_\lambda\setminus S}\|_2^2] &= \mathbf{E}_{\boldsymbol{\Theta}}\left[\Big\|\sum_{i\in S^c}\boldsymbol{\xi}_i\mathbb{1}(\|\boldsymbol{\xi}_i\|_2\geq\lambda/\sigma)\Big\|_2^2\right]\\ &= \sum_{i\in S^c}\mathbf{E}_{\boldsymbol{\Theta}}[\|\boldsymbol{\xi}_i\|_2^2\mathbb{1}(\|\boldsymbol{\xi}_i\|_2^2\geq p+t)]\\ &\leq 2n\big(pe^{-t^2/32p}\,\mathbb{1}_{t<4p}+te^{-t/4}\mathbb{1}_{t\geq 4p}\big).\end{aligned}$$

Using the fact that $t=4\log(1+n/s^2)\vee\big\{16p\log(1+n^2p/s^4)\big\}^{1/2}$ we arrive at

$$\mathbf{E}_{\boldsymbol{\Theta}}[\|\boldsymbol{\Xi}_{S_\lambda\setminus S}\|_2^2]\leq\Big(8s^2\log(1+n/s^2)\Big)\vee\Big(2s^2\sqrt{p}\wedge 2np\Big).$$

The result follows from the previous upper bounds and the choice of λ. □

Proof of Proposition 1. Recall that S_λ denotes the set of indices such that $\|\boldsymbol{Y}_i\|_2\geq\lambda$. We define $\boldsymbol{\Theta}$ as the matrix with entries $\epsilon=\sigma p^{-1/4}$ in the first s columns and 0 elsewhere. Using the inequality $\forall a,b,\ (a-b)^2\geq a^2/2-b^2$ and (16), we get

$$\begin{aligned}&\mathbf{E}_{\boldsymbol{\Theta}}\big[\big\|\widehat{\boldsymbol{L}}^{\mathrm{GHT}}-\boldsymbol{L}(\boldsymbol{\Theta})\big\|_2^2\big]\\ &\quad\geq\frac{1}{2}\mathbf{E}_{\boldsymbol{\Theta}}\big[\big\|\sigma\boldsymbol{L}(\boldsymbol{\Xi}_{S_\lambda\setminus S})-\boldsymbol{L}(\boldsymbol{\Theta}_{S\setminus S_\lambda})\big\|_2^2\big]-\sigma^2\mathbf{E}_{\boldsymbol{\Theta}}\big[\big\|\boldsymbol{L}(\boldsymbol{\Xi}_{S\cap S_\lambda})\big\|_2^2\big].\end{aligned}$$

Moreover, $\boldsymbol{L}(\boldsymbol{\Xi}_{S_\lambda\setminus S})$ being centered and independent of $\boldsymbol{L}(\boldsymbol{\Theta}_{S\setminus S_\lambda})$, we can develop

$$\mathbf{E}_{\boldsymbol{\Theta}}\big[\big\|\sigma\boldsymbol{L}(\boldsymbol{\Xi}_{S_\lambda\setminus S})-\boldsymbol{L}(\boldsymbol{\Theta}_{S\setminus S_\lambda})\big\|_2^2\big]=\sigma^2\mathbf{E}_{\boldsymbol{\Theta}}\big[\big\|\boldsymbol{L}(\boldsymbol{\Xi}_{S_\lambda\setminus S})\big\|_2^2\big]+\mathbf{E}_{\boldsymbol{\Theta}}\big[\big\|\boldsymbol{L}(\boldsymbol{\Theta}_{S\setminus S_\lambda})\big\|_2^2\big].$$

First assume that $\lambda^2\geq\sigma^2p+\sigma^2\sqrt{p}$ and focus on the second term in the right-hand side of the last display. Using Jensen's inequality, we have

$$\begin{aligned}\mathbf{E}_{\boldsymbol{\Theta}}\big[\big\|\boldsymbol{L}(\boldsymbol{\Theta}_{S\setminus S_\lambda})\big\|_2^2\big]&\geq\big\|\mathbf{E}_{\boldsymbol{\Theta}}\big[\boldsymbol{L}(\boldsymbol{\Theta}_{S\setminus S_\lambda})\big]\big\|_2^2\\ &=p\epsilon^2s^2\max_{i\in S}\mathbf{P}_{\boldsymbol{\Theta}}\big(\|\boldsymbol{Y}_i\|_2<\lambda\big)^2\\ &=s^2\sqrt{p}\max_{i\in S}\mathbf{P}_{\boldsymbol{\Theta}}\big(\|\boldsymbol{Y}_i\|_2<\lambda\big)^2.\end{aligned}$$

On the other hand, since $\operatorname{sgn}(\boldsymbol{\xi}_i^\top\boldsymbol{\theta}_i)$ is a Rademacher random variable independent of $\|\boldsymbol{\xi}_i\|_2^2$, for every $i\in S$, we have

$$\mathbf{P}_{\boldsymbol{\Theta}}\big(\|\boldsymbol{Y}_i\|_2^2<\lambda^2\big)\geq 0.5\mathbf{P}\big(\sigma^2\|\boldsymbol{\xi}_1\|_2^2-\sigma^2p<\sigma^2\sqrt{p}-p\epsilon^2\big)=0.5\mathbf{P}\big(\|\boldsymbol{\xi}_1\|_2^2-p<0\big).$$

This last probability converges to $1/2$, so that it is larger than $1/4$ for all p large enough.

In the other case, $\lambda^2<\sigma^2p+\sigma^2\sqrt{p}$, we consider the first term:

$$\begin{aligned}\mathbf{E}_{\boldsymbol{\Theta}}\big[\big\|\boldsymbol{L}(\boldsymbol{\Xi}_{S_\lambda\setminus S})\big\|_2^2\big]&=(n-s)p-(n-s)\mathbf{E}[\|\boldsymbol{\xi}_1\|_2^2\mathbb{1}_{\|\boldsymbol{\xi}_1\|_2\leq\lambda/\sigma}]\\ &\geq(n-s)p-(n-s)(p+\sqrt{p})\mathbf{P}\big(\|\boldsymbol{\xi}_1\|_2^2-p\leq\sqrt{p}\big).\end{aligned}$$

The probability in the right-hand side converges to $\Phi(2^{-1/2}) \le 0.8$, so that for $p \ge 64$ and $s \le n/2$, we have

$$\mathbf{E}_{\boldsymbol{\Theta}}\left[\left\|\boldsymbol{L}(\boldsymbol{\Xi}_{S_\lambda \setminus S})\right\|_2^2\right] \ge (n-s)(0.2p - 0.8p^{1/2}) \ge 0.05np.$$

Finally, according to (18), we have for p large enough

$$\mathbf{E}_{\boldsymbol{\Theta}}\left[\|\boldsymbol{L}(\boldsymbol{\Xi}_{S\cap S_\lambda})\|_2^2\right] \le 3s^2 + 3sp + 12s \le \frac{s^2\sqrt{p}}{65} + 3sp$$

so that

$$\mathbf{E}_{\boldsymbol{\Theta}}\left[\left\|\widehat{\boldsymbol{L}}^{\mathrm{GHT}} - \boldsymbol{L}(\boldsymbol{\Theta})\right\|_2^2\right] \ge \frac{s^2\sqrt{p}}{65} \wedge (0.05np) - 3sp.$$

We have to distinguish between two cases. If $s^2\sqrt{p} \le 196sp$, then the result holds in view of Theorem 5. In the opposite case, the result holds as long as $s \le n/61$. □

Proof of Proposition 2. Without loss of generality, we assume $\sigma = 1$. To ease notation, we set $\boldsymbol{L} = \boldsymbol{L}(\boldsymbol{\Theta})$ and $\mathbb{B} = \{\boldsymbol{u} \in \mathbb{R}^p : \|\boldsymbol{u} - \boldsymbol{L}(\mathbf{Y})\|_2^2 \le 7np\log n\}$. On the one hand, we remark that if $\boldsymbol{L} \in \mathbb{B}$, since $\widehat{\boldsymbol{L}}^{\mathrm{pr}}$ belongs to the same ball, we have $\|\widehat{\boldsymbol{L}}^{\mathrm{pr}} - \boldsymbol{L}\|_2^2 \le 28np\log n \wedge \|\widehat{\boldsymbol{L}} - \boldsymbol{L}\|_2^2$. Therefore,

$$\mathbf{E}_{\boldsymbol{\Theta}}\left(\left\|\widehat{\boldsymbol{L}}^{\mathrm{pr}} - \boldsymbol{L}\right\|_2^2 \mathbb{1}_{\boldsymbol{L}\in\mathbb{B}}\right) \le (28np\log n) \wedge \mathbf{E}_{\boldsymbol{\Theta}}\left\|\widehat{\boldsymbol{L}} - \boldsymbol{L}\right\|_2^2. \tag{20}$$

On the other hand, since $\widehat{\boldsymbol{L}}^{\mathrm{pr}} \in \mathbb{B}$, we have

$$\left\|\widehat{\boldsymbol{L}}^{\mathrm{pr}} - \boldsymbol{L}\right\|_2 \le \left\|\widehat{\boldsymbol{L}}^{\mathrm{pr}} - \boldsymbol{L}(\mathbf{Y})\right\|_2 + \left\|\boldsymbol{L}(\mathbf{Y}) - \boldsymbol{L}\right\|_2 \le (7np\log n)^{1/2} + \|\boldsymbol{L}(\boldsymbol{\Xi})\|_2. \tag{21}$$

Since $\boldsymbol{L} \notin \mathbb{B}$ is equivalent to $\|\boldsymbol{\Xi}\|_2^2 > 7np\log n$ and $\|\boldsymbol{\Xi}\|_2^2 \sim n\chi_p^2$, we have

$$\mathbf{E}_{\boldsymbol{\Theta}}[\|\boldsymbol{L}(\boldsymbol{\Xi})\|_2^2 \mathbb{1}_{\boldsymbol{L}\notin\mathbb{B}}] = n\mathbf{E}_{\eta\sim\chi_p^2}[\eta \mathbb{1}_{\eta > 7p\log n}] \le 12np\log n e^{-1.5p\log n}, \tag{22}$$

where we have used a standard bound on the tails of χ^2 distribution recalled in Lemma 6. For $p \ge 2$ and $n \ge 3$, it is easy to see that $12np\log n e^{-1.5p\log n} = 12(n^{1-p}\log n)(pn^{-p/2}) \le 12\frac{\log 3}{3}\frac{2}{3} \le 3$. Combining this inequality with (21) and (22), and using the Minkowski inequality, we get

$$\begin{aligned}\left\{\mathbf{E}_{\boldsymbol{\Theta}}\left(\left\|\widehat{\boldsymbol{L}}^{\mathrm{pr}} - \boldsymbol{L}\right\|_2^2 \mathbb{1}_{\boldsymbol{L}\notin\mathbb{B}}\right)\right\}^{1/2} &\le (7np\log n)^{1/2}\mathbf{P}(\eta > 7p\log n)^{1/2} + \sqrt{3} \\ &\le (7np\log n)^{1/2} n^{-3p/4} + \sqrt{3} \le \sqrt{10},\end{aligned} \tag{23}$$

where we have used Lemma 5. (20) and (23) together imply the claim of the proposition. □

Proof of Theorem 4. To ease notation, for every random vector X we write $\|X\|_{L_2}$ for $(\mathbf{E}[\|X\|_2^2])^{1/2}$. We first notice that

$$\widehat{\boldsymbol{L}}^{\mathrm{GST}} - \boldsymbol{L}(\boldsymbol{\Theta}) = \sum_{i\in S} \boldsymbol{\theta}_i \left\{ \left(1 - \frac{\sigma\gamma}{(\|\boldsymbol{Y}_i\|_2^2 - \sigma^2 p)_+^{1/2}}\right)_+ - 1 \right\} \qquad (:= T_1)$$

$$+ \sigma \sum_{i\in S} \boldsymbol{\xi}_i \left(1 - \frac{\sigma\gamma}{\big(\|\boldsymbol{Y}_i\|_2^2 - \sigma^2 p\big)_+^{1/2}}\right)_+ \qquad (:= T_2)$$

$$+ \sigma \sum_{i\notin S} \boldsymbol{\xi}_i \left(1 - \frac{\sigma\gamma}{\sigma\big(\|\boldsymbol{\xi}_i\|_2^2 - p\big)_+^{1/2}}\right)_+ \qquad (:= T_3),$$

so that we only need to bound the expected squared norms of the three terms in the right-hand side. These three terms have the following meanings: the first one is the bias of estimation or the approximation error, the second term is the stochastic error on the support S, whereas the third term is the stochastic error on $S^{\complement}$.

Evaluation of the approximation error For the first term, we use the Minkowski inequality as follows

$$\|T_1\|_{L_2} \le \Big\| \sum_{i\in S} \boldsymbol{\theta}_i \mathbb{1}_{\|\boldsymbol{Y}_i\|_2^2 \le \sigma^2(p+\gamma^2)} \Big\|_{L_2} + \sigma \Big\| \sum_{i\in S} \boldsymbol{\theta}_i \frac{\gamma\, \mathbb{1}_{\|\boldsymbol{Y}_i\|_2^2 > \sigma^2(p+\gamma^2)}}{\big(\|\boldsymbol{Y}_i\|_2^2 - \sigma^2 p\big)^{1/2}} \Big\|_{L_2}. \quad (24)$$

The first part can be treated exactly as in (19) of the proof of Theorem 2, *i.e.*,

$$\Big\| \sum_{i\in S} \boldsymbol{\theta}_i \mathbb{1}_{\|\boldsymbol{Y}_i\|_2^2 \le \sigma^2(p+\gamma^2)} \Big\|_{L_2} \le \sigma s\big(\gamma + 2 + (2p)^{1/4}\big). \quad (25)$$

For assessing the second term in the right hand side of (24), we set

$$T_{1,i} = \frac{\|\boldsymbol{\theta}_i\|_2 \gamma\, \mathbb{1}_{\|\boldsymbol{Y}_i\|_2^2 > \sigma^2(p+\gamma^2)}}{(\|\boldsymbol{\theta}_i\|_2^2 + 2\sigma\boldsymbol{\theta}_i^\top \boldsymbol{\xi}_i + \sigma^2\big(\|\boldsymbol{\xi}_i\|_2^2 - p\big))^{1/2}}.$$

We consider two cases. The first case corresponds to $\big|2\sigma\boldsymbol{\theta}_i^\top \boldsymbol{\xi}_i + \sigma^2\big(\|\boldsymbol{\xi}_i\|_2^2 - p\big)\big| < \|\boldsymbol{\theta}_i\|_2^2/2$. In this case one easily checks that $T_{1,i} \le \sqrt{2}\,\gamma$. In the second case, $\big|2\sigma\boldsymbol{\theta}_i^\top \boldsymbol{\xi}_i + \sigma^2\big(\|\boldsymbol{\xi}_i\|_2^2 - p\big)\big| \ge \|\boldsymbol{\theta}_i\|_2^2/2$, we have

$$\|\boldsymbol{\theta}_i\|_2^2/4 \le 2\sigma|\boldsymbol{\theta}_i^\top \boldsymbol{\xi}_i| \quad \text{or} \quad \|\boldsymbol{\theta}_i\|_2^2/4 \le \sigma^2\big|\|\boldsymbol{\xi}_i\|_2^2 - p\big|.$$

This readily implies that

$$T_{1,i} = \frac{\|\boldsymbol{\theta}_i\| \gamma\, \mathbb{1}_{\|\boldsymbol{Y}_i\|_2^2 > \sigma^2(p+\gamma^2)}}{\big(\|\boldsymbol{Y}_i\|_2^2 - \sigma^2 p\big)^{1/2}} \le \|\boldsymbol{\theta}_i\| \le 8\sigma \frac{|\boldsymbol{\theta}_i^\top \boldsymbol{\xi}_i|}{\|\boldsymbol{\theta}_i\|_2} + 2\sigma\sqrt{|\|\boldsymbol{\xi}_i\|_2^2 - p|}.$$

Therefore, using the fact that $\mathbf{E}[|\boldsymbol{\theta}_i^\top \boldsymbol{\xi}_i|^2] = \|\boldsymbol{\theta}_i\|_2^2$ and $\mathbf{E}[|\|\boldsymbol{\xi}_i\|_2^2 - p|^2] = 2p$, we get

$$\|T_{1,i}\|_{L_2} \le 8\sigma + 2\sigma(2p)^{1/4}.$$

Combining this inequality with (24) and (25), we arrive at

$$\|T_1\|_{L_2} \le \sigma s\gamma + 2\sigma s + \sigma s(2p)^{1/4} + 8\sigma s + 2\sigma s(2p)^{1/4}$$
$$= \sigma s\gamma + 10\sigma s + 3\sigma s(2p)^{1/4}.$$

Evaluation of the stochastic error on S The second term can be treated exactly as in the proof of Theorem 2 using Wishart matrices, *i.e.*,

$$\|T_2\|_{L_2}^2 \le \sigma^2 s\mathbf{E}[\|\mathbf{\Xi}_S\|^2] \le \sigma^2(3s^2 + 3sp + 12s).$$

Evaluation of the stochastic error on $S^{\complement}$ For the third term, we write

$$\|T_3\|_{L_2}^2 = \sigma^2\left\|\sum_{i\notin S}\boldsymbol{\xi}_i\left(1 - \frac{\gamma}{\big(\|\boldsymbol{\xi}_i\|_2^2 - p\big)_+^{1/2}}\right)_+\right\|_{L_2}^2$$
$$= \sigma^2\sum_{i\notin S}\left\|\boldsymbol{\xi}_i\left(1 - \frac{\gamma}{\big(\|\boldsymbol{\xi}_i\|_2^2 - p\big)_+^{1/2}}\right)_+\right\|_{L_2}^2 \le \sigma^2\sum_{i\notin S}\mathbf{E}\big[\|\boldsymbol{\xi}_i\|_2^2\mathbb{1}_{\|\boldsymbol{\xi}_i\|_2^2 > p+\gamma^2}\big].$$

We conclude by Lemma 6 that the last term satisfies

$$\|T_3\|_{L_2}^2 \le 2n\sigma^2\big(pe^{-\gamma^4/32p}\,\mathbb{1}_{\gamma^2<4p} + \gamma^2 e^{-\gamma^2/4}\mathbb{1}_{\gamma^2\ge 4p}\big).$$

Using the fact that $\gamma^2 = 4\log(1+n/s^2) \vee \big\{16p\log(1+n^2p/s^4)\big\}^{1/2}$ we arrive at

$$\|T_3\|_{L_2}^2 \le \sigma^2\Big(8s^2\log(1+n/s^2)\Big) \vee \Big(2s^2\sqrt{p} \wedge 2np\Big).$$

This completes the proof of the theorem. □

5.2. Proofs of the theorems of Section 3

This section gathers the proofs all of the results concerning the problem of robust estimation of a Gaussian mean.

Proof of Theorem 6. In view of (9), we have

$$\big\|(\mathbf{\Theta} - \widehat{\mathbf{\Theta}})\mathbf{\Pi} + \sigma\mathbf{\Xi}\mathbf{\Pi}\big\|_F^2 + \lambda\sum_{i=1}^n\|\widehat{\boldsymbol{\theta}}_i\|_2 \le \sigma^2\|\mathbf{\Xi}\mathbf{\Pi}\|_F^2 + \lambda\sum_{i=1}^n\|\boldsymbol{\theta}_i\|_2.$$

Developing the left-hand side, this yields

$$\|(\mathbf{\Theta} - \widehat{\mathbf{\Theta}})\mathbf{\Pi}\|_F^2 \le 2\sigma\sum_{n=1}^n(\mathbf{\Xi}\mathbf{\Pi})_i^\top(\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i) + \lambda\sum_{i=1}^n\big\{\|\boldsymbol{\theta}_i\|_2 - \|\widehat{\boldsymbol{\theta}}_i\|_2\big\}.$$

Using the Cauchy-Schwarz inequality, we have on the event

$$\mathcal{A} = \Big\{\max_{i\in[n]}\big\|(\mathbf{\Xi}\mathbf{\Pi})_i\big\|_2 \le \frac{\lambda}{4\sigma}\Big\}$$

that

$$\begin{aligned}\|(\boldsymbol{\Theta}-\widehat{\boldsymbol{\Theta}})\boldsymbol{\Pi}\|_F^2 &\le \frac{\lambda}{2}\sum_{i=1}^n\Big\{\|\widehat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i\|_2+2\|\boldsymbol{\theta}_i\|_2-2\|\widehat{\boldsymbol{\theta}}_i\|_2\Big\}\\ &\le \frac{3\lambda}{2}\sum_{i\in S}\|\widehat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i\|_2-\frac{\lambda}{2}\sum_{i\notin S}\|\widehat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i\|_2,\end{aligned}$$

where we used the triangular inequality. The last inequality implies that

$$\sum_{i\notin S}\|\widehat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i\|_2\le 3\sum_{i\in S}\|\widehat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i\|_2. \tag{26}$$

Furthermore, using the Cauchy-Schwarz inequality, we get

$$\|(\boldsymbol{\Theta}-\widehat{\boldsymbol{\Theta}})\boldsymbol{\Pi}\|_F^2\le\frac{3\lambda}{2}\sqrt{s}\|\boldsymbol{\Theta}-\widehat{\boldsymbol{\Theta}}\|_F.$$

We can then apply Lemma 4, since (26) ensures that the condition is satisfied with $a=3$: provided that $s\le n/16$, we have

$$\frac{1}{2}\|\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}\|_F^2\le\frac{3\lambda}{2}\sqrt{s}\|\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}\|_F,$$

so that

$$\|\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}\|_F\le 3\lambda\sqrt{s}.$$

The first claim follows now from Lemma 7. To show the second inequality, it suffices to remark that

$$\begin{aligned}\|\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}\|_2^2 &\le \frac{2}{n^2}\|(\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta})\mathbf{1}_n\|_2^2+2\sigma^2\|\boldsymbol{L}_n(\boldsymbol{\Xi})\|_2^2\\ &\le \frac{2}{n^2}\bigg(\sum_{i\in[n]}\|\widehat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i\|_2\bigg)^2+2\sigma^2\|\boldsymbol{L}_n(\boldsymbol{\Xi})\|_2^2\\ &\le \frac{2}{n^2}\bigg(4\sum_{i\in S}\|\widehat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i\|_2\bigg)^2+2\sigma^2\|\boldsymbol{L}_n(\boldsymbol{\Xi})\|_2^2\\ &\le \frac{32s}{n^2}\sum_{i\in S}\|\widehat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i\|_2^2+2\sigma^2\|\boldsymbol{L}_n(\boldsymbol{\Xi})\|_2^2\\ &\le \frac{288s^2\lambda^2}{n^2}+2\sigma^2\|\boldsymbol{L}_n(\boldsymbol{\Xi})\|_2^2.\end{aligned}$$

To complete the proof, we use the fact that $n\|\boldsymbol{L}_n(\boldsymbol{\Xi})\|_2^2$ is a $\chi^2(p)$ random variable, which implies that with probability at least $1-\delta/2$ it is bounded from above by $2p+4\log(2/\delta)$. □

Proof of Proposition 7. To ease notation, we set $\sigma_{n,p}^2=(n-1)\sigma^2p/n$, $w_i=\big(1-\sigma\gamma/\sqrt{(\|\boldsymbol{Z}_i\|_2^2-\sigma_{n,p}^2)_+}\big)_+$, $\widetilde{\boldsymbol{\Delta}}_n=\widetilde{\boldsymbol{L}}_n-\boldsymbol{L}_n(\boldsymbol{\Theta})$ and $\bar{\boldsymbol{\xi}}_i=\boldsymbol{\xi}_i-\bar{\boldsymbol{\xi}}$. Then,

$$\boldsymbol{Z}_i=\boldsymbol{\theta}_i+\widetilde{\boldsymbol{\Delta}}_n+\sigma\bar{\boldsymbol{\xi}}_i.$$

We first show that with high probability the weights w_i vanish outside the support S.

Lemma 2. *In the event $\Omega_1 \cap \Omega_2$, where*

$$\Omega_2 = \Big\{ \max_i \big| \|\bar{\boldsymbol{\xi}}_i\|_2^2 - (1 - 1/n)p \big| \le 0.5\gamma^2 \Big\},$$

we have $w_i = 0$ for every $i \notin S$. Furthermore, under the condition $\gamma^2 > 4\log(4n/\delta) + 4\{p\log(4n/\delta)\}^{1/2}$, the probability of Ω_2 is at least $1 - \delta/2$.

Using equation (11) and the fact that $w_i = 0$ for every $i \notin S$ in the event $\Omega_1 \cap \Omega_2$, we get

$$\boldsymbol{L}(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}) = \sum_{i=1}^n \boldsymbol{Z}_i w_i - \sum_{i \in S} \boldsymbol{\theta}_i = \sum_{i \in S} \big(\boldsymbol{Z}_i w_i - \boldsymbol{\theta}_i \big).$$

Replacing $\boldsymbol{Z}_i$ by $\boldsymbol{\theta}_i + \widetilde{\boldsymbol{\Delta}}_n + \sigma\bar{\boldsymbol{\xi}}_i$ and using the triangle inequality, we obtain

$$\begin{aligned}
\|\boldsymbol{L}(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})\|_2 &\le \Big\| \sum_{i \in S} (\boldsymbol{Z}_i - \sigma\bar{\boldsymbol{\xi}}_i)(w_i - 1) \Big\|_2 + \Big\| \sum_{i \in S} \big(\boldsymbol{Z}_i - \boldsymbol{\theta}_i \big) + \sigma\bar{\boldsymbol{\xi}}_i (w_i - 1) \Big\|_2 \\
&\le \sum_{i \in S} \|\boldsymbol{Z}_i - \sigma\bar{\boldsymbol{\xi}}_i\|_2 |w_i - 1| + \Big\| \sum_{i \in S} \big(\widetilde{\boldsymbol{\Delta}}_n + \sigma\bar{\boldsymbol{\xi}}_i w_i \big) \Big\|_2 \\
&\le \sum_{i \in S} \|\boldsymbol{Z}_i - \sigma\bar{\boldsymbol{\xi}}_i\|_2 |w_i - 1| + s\big\|\widetilde{\boldsymbol{\Delta}}_n\big\|_2 + \sigma \Big\| \sum_{i \in S} \bar{\boldsymbol{\xi}}_i w_i \Big\|_2. \qquad (27)
\end{aligned}$$

We will now evaluate the first and the third terms of the right-hand side.

Lemma 3. *There is a sequence of standard Gaussian random variables $\eta_1, \dots, \eta_n$ such that in the event $\Omega_1 \cap \Omega_2$, it holds*

$$\|\boldsymbol{Z}_i - \sigma\bar{\boldsymbol{\xi}}_i\|_2 |w_i - 1| \le 2\sigma|\eta_i| + \big\{ |2\sigma\bar{\boldsymbol{\xi}}_i^\top \widetilde{\boldsymbol{\Delta}}_n| \big\}^{1/2} + 1.96\sigma\gamma. \qquad (28)$$

Let us introduce the $p \times s$ matrix $\boldsymbol{\Xi}_S = [\bar{\boldsymbol{\xi}}_i; i \in S]$ (the matrix obtained by concatenating the vectors $\bar{\boldsymbol{\xi}}_i$ with subscript i running over S). Using the Hölder inequality, one can check that

$$\begin{aligned}
\sum_{i \in S} \big\{ 2\sigma |\bar{\boldsymbol{\xi}}_i^\top \widetilde{\boldsymbol{\Delta}}_n| \big\}^{1/2} &\le \sqrt{2\sigma}\, s^{3/4} \Big(\sum_{i \in S} |\bar{\boldsymbol{\xi}}_i^\top \widetilde{\boldsymbol{\Delta}}_n|^2 \Big)^{1/4} \\
&= \sqrt{2\sigma}\, s^{3/4} \|\boldsymbol{\Xi}_S^\top \widetilde{\boldsymbol{\Delta}}_n\|_2^{1/2} \\
&\le \sqrt{2\sigma}\, s^{3/4} \|\boldsymbol{\Xi}_S\|^{1/2} \|\widetilde{\boldsymbol{\Delta}}_n\|_2^{1/2}. \qquad (29)
\end{aligned}$$

Finally, the third term in the right-hand side of (27) can be bounded as follows:

$$\Big\| \sum_{i \in S} \bar{\boldsymbol{\xi}}_i w_i \Big\|_2 = \|\boldsymbol{\Xi}_S \boldsymbol{w}\|_2 \le \sqrt{s}\, \|\boldsymbol{\Xi}_S\|. \qquad (30)$$

Combining (27), (28), (29) and (30), we arrive at

$$\begin{aligned}\|\boldsymbol{L}(\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta})\|_2 &\le 2\sigma\sum_{i\in S}\big(|\eta_i|+\{2\sigma|\widetilde{\boldsymbol{\Delta}}_n^\top\bar{\boldsymbol{\xi}}_i|\}^{1/2}\big)+1.96\,\sigma s\gamma+s\big\|\widetilde{\boldsymbol{\Delta}}_n\big\|_2\\ &\quad+\sigma\sqrt{s}\,\|\boldsymbol{\Xi}_S\|\\ &\le 2\sigma\sum_{i\in S}|\eta_i|+2.21\,\sigma s\gamma+\sqrt{2\sigma}\,s^{3/4}\|\boldsymbol{\Xi}_S\|^{1/2}\big\|\widetilde{\boldsymbol{\Delta}}_n\big\|_2^{1/2}+\sigma\sqrt{s}\|\boldsymbol{\Xi}_S\|\\ &\le 2\sigma\sum_{i\in S}|\eta_i|+2.34\,\sigma s\gamma+2\sigma\sqrt{s}\,\|\boldsymbol{\Xi}_S\|.\end{aligned}$$

According to (Vershynin, 2012, Corollary 5.35) (recalled in Lemma 8 below for the reader's convenience) the event $\Omega_3=\big\{\|\boldsymbol{\Xi}_S\|\le\sqrt{s}+\sqrt{p}+\{2\log(8/\delta)\}^{1/2}\big\}$ has probability at least $1-\delta/8$. One can also check that with probability at least $1-\delta/8$, the event $\Omega_4=\{\sum_{i\in S}|\eta_i|\le s+\sqrt{s\log(8/\delta)}\}$ is realized. Assuming that $\delta\le 1/2$ so that $\log(8/\delta)\le(4/3)\log(4/\delta)$, this implies that in $\Omega_1\cap\bar{\Omega}$ with $\bar{\Omega}=\Omega_2\cap\Omega_3\cap\Omega_4$, we have

$$\|\boldsymbol{L}(\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta})\|_2\le 4\sigma(\,s\gamma+s+\sqrt{sp}+\{2s\log(4/\delta)\}^{1/2}).$$

This completes the proof. □

6. Some technical lemmas

Lemma 4. *Let us introduce the projection matrix* $\boldsymbol{\Pi}=\mathbf{I}_n-\frac{1}{n}\mathbf{J}_n$, *where* $\mathbf{I}_n$ *and* $\mathbf{J}_n$ *are respectively the* $n\times n$ *identity matrix and the constant matrix with all the entries equal to* 1. *Let* $\mathbf{U}$ *be a* $p\times n$ *matrix with columns* $\boldsymbol{u}_i\in\mathbb{R}^p$ *satisfying, for some set* $S\subset\{1,\dots,n\}$ *and some real number* $a>0$,

$$\sum_{i\notin S}\|\boldsymbol{u}_i\|_2\le a\sum_{i\in S}\|\boldsymbol{u}_i\|_2, \tag{31}$$

then

$$\|\mathbf{U}\boldsymbol{\Pi}\|_F^2\ge\Big(1-\frac{(1+a)^2|S|}{n}\Big)\|\mathbf{U}\|_F^2.$$

Proof. We denote by $\boldsymbol{u}_i\in\mathbb{R}^n$ the column vector corresponding to the i-th row of $\mathbf{U}$. On the one hand, since $1/n\mathbf{J}_n$ is an orthogonal projection matrix, we have by the Pythagorean theorem

$$\|\boldsymbol{\Pi}\boldsymbol{u}^i\|_2^2=\|\boldsymbol{u}^i\|_2^2-\frac{1}{n^2}\|\mathbf{J}_n\boldsymbol{u}^i\|_2^2.$$

In particular, this implies that

$$\|\mathbf{U}\boldsymbol{\Pi}\|_F^2=\|\mathbf{U}\|_F^2-\frac{1}{n^2}\|\mathbf{U}\mathbf{J}_n\|_F^2. \tag{32}$$

On the other hand,

$$\|\mathbf{U}\mathbf{J}_n\|_F = \|\mathbf{U}1_n\mathbf{1}_n^\top\|_F = \sqrt{n}\,\|\mathbf{U}\mathbf{1}_n^\top\|_2 = \sqrt{n}\,\Big\|\sum_{i=1}^n \boldsymbol{u}_i\Big\|_2 \le \sqrt{n}\sum_{i=1}^n \|\boldsymbol{u}_i\|_2.$$

Using (31) and the Cauchy-Schwarz inequality, we get

$$\begin{aligned}\sum_{i=1}^n \|\boldsymbol{u}_i\|_2 &\le (1+a)\sum_{i\in S}\|\boldsymbol{u}_i\|_2 \le (1+a)|S|^{1/2}\Big(\sum_{i\in S}\|\boldsymbol{u}_i\|_2^2\Big)^{1/2}\\ &\le (1+a)|S|^{1/2}\|\mathbf{U}\|_F.\end{aligned}$$

This readily yields $\|\mathbf{U}\mathbf{J}_n\|_F \le (1+a)(n|S|)^{1/2}\|\mathbf{U}\|_F$. Combining this inequality with (32), we get the claim of the lemma. □

Proof of Lemma 1. We prove by inductive reasoning that for every k we have $\epsilon_k^2 \le p$ and (13). This is trivially true for $k=0$. Assume that these claims hold true for some given value k. Let us check them for the value $k+1$. From recursion (12), one can infer that

$$\epsilon_{k+1} \le \frac{16.5s}{n}(\epsilon_k \vee p^{1/4}\epsilon_k^{1/2}) + a \le \frac{33s}{n}(\epsilon_k \vee p^{1/4}\epsilon_k^{1/2}) \vee 2a. \tag{33}$$

The conditions of the lemma directly imply that $\epsilon_{k+1}^2 \le p$. Therefore,

$$\epsilon_{k+1} \le \frac{33sp^{1/4}\epsilon_k^{1/2}}{n} \vee 2a.$$

Applying inequality (13), we get the claim.

One can note that even in the case $\epsilon_k > \sqrt{p}$, we get from (33) that

$$\epsilon_{k+1} \le \frac{1}{2}\epsilon_k + a.$$

Therefore, if the preliminary estimator is not good enough to guarantee that $\epsilon_0 \le \sqrt{p}$, after a number of steps at most logarithmic in $\epsilon_0/(\sqrt{p}-2a)$, we will get an error ϵ_k smaller than $\sqrt{p}$. □

Proof of Lemma 2. It follows from the definition of w_i that, for every $i \notin S$, $w_i > 0$ is equivalent to

$$\|\sigma\bar{\boldsymbol{\xi}}_i + \widetilde{\boldsymbol{\Delta}}_n\|_2^2 > \sigma^2(p+\gamma^2).$$

In view of the triangle inequality, this implies that

$$\sigma\|\bar{\boldsymbol{\xi}}_i\|_2 + \|\widetilde{\boldsymbol{\Delta}}_n\|_2 > \sigma\sqrt{p+\gamma^2}.$$

In the event $\Omega_1 \cap \Omega_2$, the last inequality implies

$$\sqrt{p+0.5\gamma^2} + \frac{\gamma^2}{4\sqrt{p+\gamma^2}} > \sqrt{p+\gamma^2}.$$

It is easy to see that the last inequality is never true, implying thus that $w_i = 0$. Indeed,

$$\sqrt{p+\gamma^2} - \sqrt{p+0.5\gamma^2} = \frac{0.5\gamma^2}{\sqrt{p+\gamma^2}+\sqrt{p+0.5\gamma^2}} > \frac{0.5\gamma^2}{\sqrt{p+\gamma^2}+\sqrt{p+\gamma^2}} = \frac{\gamma^2}{4\sqrt{p+\gamma^2}}.$$

The fact that the probability of Ω_2 is at least $1-\delta/2$ is a consequence of the tail bound of a chi-squared random variable and the union bound. □

Proof of Lemma 3. The definition of $\boldsymbol{Z}_i$ yields

$$\|\boldsymbol{Z}_i - \sigma\bar{\boldsymbol{\xi}}_i\|_2 |w_i - 1| = \|\boldsymbol{Z}_i - \sigma\bar{\boldsymbol{\xi}}_i\|_2 \min\left\{1, \frac{\sigma\gamma}{\big(\|\boldsymbol{Z}_i\|_2^2 - \sigma_{n,p}^2\big)_+^{1/2}}\right\}.$$

Let us introduce the random variables $\eta_i = \frac{n}{n-1}\boldsymbol{\theta}_i^\top \bar{\boldsymbol{\xi}}_i / \|\boldsymbol{\theta}_i\|_2$. It is clear that η_i is Gaussian with zero mean and unit variance. Repeated use of the triangle inequality leads to

$$\begin{aligned}
\|\boldsymbol{Z}_i\|_2^2 &\geq \|\boldsymbol{Z}_i - \sigma\bar{\boldsymbol{\xi}}_i\|_2^2 + \|\sigma\bar{\boldsymbol{\xi}}_i\|_2^2 + 2\sigma\bar{\boldsymbol{\xi}}_i^\top(\boldsymbol{\theta}_i + \widetilde{\boldsymbol{\Delta}}_n) \\
&\geq \|\boldsymbol{Z}_i - \sigma\bar{\boldsymbol{\xi}}_i\|_2^2 + \|\sigma\bar{\boldsymbol{\xi}}_i\|_2^2 - 2\sigma\|\boldsymbol{\theta}_i\|_2|\eta_i| + 2\sigma\bar{\boldsymbol{\xi}}_i^\top\widetilde{\boldsymbol{\Delta}}_n \\
&\geq \|\boldsymbol{Z}_i - \sigma\bar{\boldsymbol{\xi}}_i\|_2^2 + \|\sigma\bar{\boldsymbol{\xi}}_i\|_2^2 - 2\sigma\|\boldsymbol{Z}_i - \sigma\bar{\boldsymbol{\xi}}_i\|_2|\eta_i| - 2\sigma\|\widetilde{\boldsymbol{\Delta}}_n\|_2|\eta_i| + 2\sigma\bar{\boldsymbol{\xi}}_i^\top\widetilde{\boldsymbol{\Delta}}_n \\
&\geq (\|\boldsymbol{Z}_i - \sigma\bar{\boldsymbol{\xi}}_i\|_2 - \sigma|\eta_i|)^2 - \sigma^2\eta_i^2 + \|\sigma\bar{\boldsymbol{\xi}}_i\|_2^2 - 2\sigma\|\widetilde{\boldsymbol{\Delta}}_n\|_2|\eta_i| + 2\sigma\bar{\boldsymbol{\xi}}_i^\top\widetilde{\boldsymbol{\Delta}}_n.
\end{aligned}$$

One can check the following simple fact: if $a, b, c, d > 0$ then

$$a\min\left(1, \frac{d}{\{(a-b)^2 - c\}_+^{1/2}}\right) \leq b + d + \sqrt{c_+}.$$

Taking in this inequality $a = \|\boldsymbol{Z}_i - \sigma\bar{\boldsymbol{\xi}}_i\|_2$, $c = \sigma^2\eta_i^2 - \|\sigma\bar{\boldsymbol{\xi}}_i\|_2^2 + 2\sigma\|\widetilde{\boldsymbol{\Delta}}_n\|_2|\eta_i| - 2\sigma\bar{\boldsymbol{\xi}}_i^\top\widetilde{\boldsymbol{\Delta}}_n + \sigma_{n,p}^2$, $b = \sigma|\eta_i|$ and $d = \sigma\gamma$, we arrive at

$$\begin{aligned}
\|\boldsymbol{Z}_i - \sigma\bar{\boldsymbol{\xi}}_i\|_2 |w_i - 1| \leq{}& 2\sigma|\eta_i| + \sigma\gamma + \|\widetilde{\boldsymbol{\Delta}}_n\|_2 \\
&+ \sqrt{(\sigma_{n,p}^2 - \sigma^2\|\bar{\boldsymbol{\xi}}_i\|_2^2)_+} + \sqrt{(2\sigma\bar{\boldsymbol{\xi}}_i^\top\widetilde{\boldsymbol{\Delta}}_n - \|\widetilde{\boldsymbol{\Delta}}_n\|_2^2)_+}.
\end{aligned}$$

In the event $\Omega_1 \cap \Omega_2$, we upper bound $\|\widetilde{\boldsymbol{\Delta}}_n\|_2$ and $|\sigma_{n,p}^2 - \sigma^2\|\bar{\boldsymbol{\xi}}_i\|_2^2|$ respectively by $0.25\sigma\gamma$ and $0.5\sigma^2\gamma^2$. This leads to the claim of the lemma. □

7. Tail bounds

In this section, we recall well-known results on the tails of some random variables appearing in the analysis of the Gaussian models of the previous sections.

Lemma 5. *If η is a random variable drawn from the χ^2_d distribution, then for every $x > 0$*

$$\mathbf{P}\big(\eta \geq d + x\big) \leq e^{-\frac{1}{16d}x(x\wedge 4d)}.$$

Lemma 6. *If η is a random variable drawn from the χ^2_d distribution, then for every $x > 0$ and $d \geq 2$,*

$$\mathbf{E}\big(\eta \mathbb{1}_{\eta \geq d+x}\big) \leq 2d\, e^{-x^2/32d}\, \mathbb{1}_{x<4d} + 2xe^{-x/4}\mathbb{1}_{x\geq 4d}.$$

Proof. First assume that $x \geq 4d$. Combining the relation

$$\mathbf{E}\big(\eta\, \mathbb{1}_{\eta\geq d+x}\big) = (d+x)\mathbf{P}(\eta \geq d+x) + \int_x^{+\infty} \mathbf{P}(\eta \geq d+t)\, dt$$

with Lemma 5, we get

$$\begin{aligned}\mathbf{E}\big(\eta\, \mathbb{1}_{\eta\geq d+x}\big) &\leq (d+x)e^{-x/4} + \int_x^{+\infty} e^{-t/4}\, dt \\ &\leq (4+d+x)e^{-x/4} \leq (4d+x)e^{-x/4} \leq 2xe^{-x/4}.\end{aligned}$$

Now assume that $x < 4d$. Then, using the Cauchy-Schwarz inequality and Lemma 5,

$$\mathbf{E}\big(\eta\, \mathbb{1}_{\eta\geq d+x}\big) \leq \sqrt{d(d+2)}e^{-x^2/32d}.$$ □

Lemma 7. *Denote $\mathbf{\Pi} = I_n - \frac{1}{n}\mathbf{J}_n$ where $\mathbf{I}_n$ is the identity matrix in dimension n and $\mathbf{J}_n$ is the constant matrix with only 1 coefficients, and assume that*

$$\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_n \overset{iid}{\sim} \mathcal{N}(0, \mathbf{I}_p).$$

Then, with probability at least $1-\delta$, the matrix $\mathbf{\Xi} = [\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_n]$ satisfies

$$\max_{i=1,\dots,n} \big\|(\mathbf{\Xi\Pi})_i\big\|_2^2 \leq 2p + 16\log(n/\delta).$$

Proof. We first notice that

$$(\mathbf{\Xi\Pi})_i = \boldsymbol{\xi}_i - \frac{1}{n}\sum_{j=1}^n \boldsymbol{\xi}_j \overset{\text{iid}}{\sim} \mathcal{N}\big(0, \tfrac{n-1}{n}\mathbf{I}_p\big).$$

This implies that the random variable $\big\|(\mathbf{\Xi\Pi})_i\big\|_2^2$ is drawn from the $\frac{n-1}{n}\chi^2_p$ distribution, and the result follows from Lemma 5. □

Lemma 8 (Corollary 5.35 in (Vershynin, 2012))**.** *Assume that $\mathbf{A}$ is a $N \times n$ random matrix with independent standard Gaussian entries. Then, for any $t \geq 0$, with probability at least $1 - 2e^{-t^2/2}$, it holds that*

$$\|\mathbf{A}\| \leq \sqrt{N} + \sqrt{n} + t.$$

We deduce from this the following lemma.

Lemma 9. *If* $\mathbf{A}$ *is a* $N \times n$ *random matrix with independent standard Gaussian entries, then* $\mathbf{E}[\|\mathbf{A}\|^2] \le 3N + 3n + 12$.

Proof. It is clear that

$$\begin{aligned}\mathbf{E}[\|\mathbf{A}\|^2] &\le 3(N+n) + \int_0^{+\infty} \mathbf{P}(\|\mathbf{A}\|^2 > 3(N+n) + x)\, dx \\ &\le 3(N+n) + 12 \int_0^{+\infty} t e^{-t^2/2}\, dt.\end{aligned}$$

The result follows from the fact that the last integral is equal to one. □

Acknowledgments

O. Collier's research has been conducted as part of the project Labex MME-DII (ANR11-LBX-0023-01). The work of A. Dalalyan was partially supported by the grant Investissements d'Avenir (ANR-11IDEX-0003/Labex Ecodec/ANR-11-LABX-0047).

References

Balakrishnan, S., Du, S. S., Li, J. and Singh, A. (2017). Computationally Efficient Robust Sparse Estimation in High Dimensions. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017* 169–212.

Balmand, S. and Dalalyan, A. S. (2015). Convex programming approach to robust estimation of a multivariate Gaussian model submitted No. 1512.04734, arXiv.

Bickel, P. J. and Ritov, Y. (1988). Estimating Integrated Squared Density Derivatives: Sharp Best Order of Convergence Estimates. *Sankhy?: The Indian Journal of Statistics, Series A (1961-2002)* **50** 381-393. MR1065550

Bunea, F., Lederer, J. and She, Y. (2014). The Group Square-Root Lasso: Theoretical Properties and Fast Algorithms. *IEEE Transactions on Information Theory* **60** 1313-1325. MR3164977

Butucea, C. and Comte, F. (2009). Adaptive estimation of linear functionals in the convolution model and applications. *Bernoulli* **15** 69–98. MR2546799

Cai, T. T. and Low, M. G. (2004). Minimax estimation of linear functionals over nonconvex parameter spaces. *Ann. Statist.* **32** 552–576. MR2060169

Cai, T. T. and Low, M. G. (2005). On adaptive estimation of linear functionals. *Ann. Statist.* **33** 2311–2343. MR2211088

Cai, T. T. and Low, M. G. (2006). Optimal adaptive estimation of a quadratic functional. *Ann. Statist.* **34** 2298–2325. MR2291501

Cai, T. T. and Low, M. G. (2011). Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *Ann. Statist.* **39** 1012–1041. MR2816346

Carpentier, A., Delattre, S., Roquain, E. and Verzelen, N. (2018). Estimating minimum effect with outlier selection. *arXiv e-prints* arXiv:1809.08330.

Chen, M., Gao, C. and Ren, Z. (2015). Robust Covariance and Scatter Matrix Estimation under Huber's Contamination Model. *ArXiv e-prints, to appear in the Annals of Statistics*. MR3845006

Chen, M., Gao, C. and Ren, Z. (2016). A general decision theory for Huber's ϵ-contamination model. *Electron. J. Statist.* **10** 3752–3774. MR3579675

Cherapanamjeri, Y., Gupta, K. and Jain, P. (2016). Nearly-optimal Robust Matrix Completion. *CoRR* **abs/1606.07315**.

Chesneau, C. and Hebiri, M. (2008). Some theoretical results on the grouped variables Lasso. *Math. Methods Statist.* **17** 317–326. MR2483460

Collier, O., Comminges, L. and Tsybakov, A. B. (2017). Minimax estimation of linear and quadratic functionals on sparsity classes. *Ann. Statist.* **45** 923–958. MR3662444

Collier, O. and Dalalyan, A. S. (2015). Curve registration by nonparametric goodness-of-fit testing. *J. Statist. Plann. Inference* **162** 20-42. MR3323102

Collier, O. and Dalalyan, A. S. (2018). Estimating linear functionals of a sparse family of Poisson means. *Statistical Inference for Stochastic Processes*. MR3824971

Collier, O., Comminges, L., Tsybakov, A. B. and Verzélen, N. (2016). Optimal adaptive estimation of linear functionals under sparsity. *ArXiv e-prints, ArXiv:1611.09744*. MR3851767

Comminges, L. and Dalalyan, A. S. (2012). Tight conditions for consistency of variable selection in the context of high dimensionality. *Ann. Statist.* **40** 2667–2696. MR3097616

Comminges, L. and Dalalyan, A. S. (2013). Minimax testing of a composite null hypothesis defined via a quadratic functional in the model of regression. *Electron. J. Statist.* **7** 146–190. MR3020417

Dalalyan, A. S. and Chen, Y. (2012). Fused sparsity and robust estimation for linear models with unknown variance. In *Advances in Neural Information Processing Systems 25: NIPS* 1268-1276.

Dalalyan, A. S. and Keriven, R. (2012). Robust estimation for an inverse problem arising in multiview geometry. *J. Math. Imaging Vision* **43** 10–23. MR2910870

Devroye, L., Lerasle, M., Lugosi, G. and Oliveira, R. I. (2016). Sub-Gaussian mean estimators. *Ann. Statist.* **44** 2695–2725. MR3576558

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A. and Stewart, A. (2016). Robust Estimators in High Dimensions without the Computational Intractability. In *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, USA* 655–664. MR3631028

Donoho, D. L. and Liu, R. C. (1987). *On Minimax Estimation of Linear Functionals. Technical report (University of California, Berkeley. Department of Statistics)*. Department of Statistics, University of California.

Donoho, D. and Montanari, A. (2016). High dimensional robust M-estimation: asymptotic variance via approximate message passing. *Probability*

Theory and Related Fields **166** 935–969. MR3568043

Donoho, D. L. and Nussbaum, M. (1990). Minimax quadratic estimation of a quadratic functional. *Journal of Complexity* **6** 290 - 323. MR1081043

Efromovich, S. and Low, M. G. (1994). Adaptive estimates of linear functionals. *Probability Theory and Related Fields* **98** 261–275. MR1258989

Goldenshluger, A. and Nemirovski, A. (1997). On spatial adaptive estimation of nonparametric regression. *Math. Meth. Statistics* **6** 135–170. MR1466625

Golubev, Y. and Levit, B. (2004). An oracle approach to adaptive estimation of linear functionals in a Gaussian model. *Mathematical Methods of Statistics* **13** 392–408. MR2126747

Huber, P. J. (1964). Robust Estimation of a Location Parameter. *Ann. Math. Statist.* **35** 73–101. MR0161415

Juditsky, A. B. and Nemirovski, A. S. (2009). Nonparametric estimation by convex programming. *Ann. Statist.* **37** 2278–2300. MR2543692

Klemela, J. and Tsybakov, A. B. (2001). Sharp Adaptive Estimation of Linear Functionals. *Ann. Statist.* **29** 1567–1600. MR1891739

Klopp, O., Lounici, K. and Tsybakov, A. B. (2017). Robust matrix completion. *Probability Theory and Related Fields* **169** 523–564. MR3704775

Koshevnik, Y. A. and Levit, B. Y. (1977). On a Non-Parametric Analogue of the Information Matrix. *Theory of Probability & Its Applications* **21** 738-753. MR0428578

Lai, K. A., Rao, A. B. and Vempala, S. (2016). Agnostic Estimation of Mean and Covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)* 665-674. MR3631029

Laurent, B., Ludena, C. and Prieur, C. (2008). Adaptive estimation of linear functionals by model selection. *Electron. J. Statist.* **2** 993–1020. MR2448602

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28** 1302–1338. MR1805785

Lecué, G. and Lerasle, M. (2017). Learning from MOM's principles: Le Cam's approach. *ArXiv e-prints*.

Lepski, O., Nemirovski, A. and Spokoiny, V. (1999). On estimation of the L_r norm of a regression function. *Probability Theory and Related Fields* **113** 221–253. MR1670867

Lepskii, O. V. (1991). On a Problem of Adaptive Estimation in Gaussian White Noise. *Theory of Probability & Its Applications* **35** 454-466. MR1091202

Lerasle, M. and Oliveira, R. I. (2011). Robust empirical mean Estimators. *ArXiv e-prints*.

Lin, Y. and Zhang, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.* **34** 2272–2297. MR2291500

Lounici, K., Pontil, M., van de Geer, S. and Tsybakov, A. B. (2011). Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.* **39** 2164–2204. MR2893865

Meier, L., van de Geer, S. and Bühlmann, P. (2009). High-dimensional

additive modeling. *Ann. Statist.* **37** 3779–3821. MR2572443

Minsker, S. (2015). Geometric median and robust estimation in Banach spaces. *Bernoulli* **21** 2308–2335. MR3378468

Nguyen, N. H. and Tran, T. D. (2013). Robust Lasso With Missing and Grossly Corrupted Observations. *IEEE Transactions on Information Theory* **59** 2036-2058. MR3043781

Vershynin, R. (2012). *Introduction to the non-asymptotic analysis of random matrices* In *Compressed Sensing: Theory and Applications* 210-268. Cambridge University Press. MR2963170

Verzelen, N. and Gassiat, E. (2016). Adaptive estimation of High-Dimensional Signal-to-Noise Ratios. *ArXiv e-prints*. MR3788186

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 49–67. MR2212574