Electronic Journal of Statistics
Vol. 6 (2012) 38–90
ISSN: 1935-7524
DOI: 10.1214/12-EJS666

Minimax risks for sparse regressions: Ultra-high dimensional phenomenons

Nicolas Verzelen

INRA, UMR 729 MISTEA, F-34060 Montpellier, France
e-mail: nicolas.verzelen@supagro.inra.fr

Abstract: Consider the standard Gaussian linear regression model $\mathbf{Y} = \mathbf{X}\theta_0 + \boldsymbol{\epsilon}$, where $\mathbf{Y} \in \mathbb{R}^n$ is a response vector and $\mathbf{X} \in \mathbb{R}^{n\times p}$ is a design matrix. Numerous work have been devoted to building efficient estimators of θ_0 when p is much larger than n. In such a situation, a classical approach amounts to assume that θ_0 is approximately sparse. This paper studies the minimax risks of estimation and testing over classes of k-sparse vectors θ_0. These bounds shed light on the limitations due to high-dimensionality. The results encompass the problem of prediction (estimation of $\mathbf{X}\theta_0$), the inverse problem (estimation of θ_0) and linear testing (testing $\mathbf{X}\theta_0 = 0$). Interestingly, an elbow effect occurs when the number of variables $k\log(p/k)$ becomes large compared to n. Indeed, the minimax risks and hypothesis separation distances blow up in this ultra-high dimensional setting. We also prove that even dimension reduction techniques cannot provide satisfying results in an ultra-high dimensional setting. Moreover, we compute the minimax risks when the variance of the noise is unknown. The knowledge of this variance is shown to play a significant role in the optimal rates of estimation and testing. All these minimax bounds provide a characterization of statistical problems that are so difficult so that no procedure can provide satisfying results.

AMS 2000 subject classifications: Primary 62J05; secondary 62F35, 62C20.
Keywords and phrases: Adaptive estimation, dimension reduction, high-dimensional regression, high-dimensional geometry, minimax risk.

Received May 2011.

1. Introduction

In many important statistical applications, including remote sensing, functional MRI and gene expressions studies the number p of parameters is much larger than the number n of observations. An active line of research aims at developing computationally fast procedures that also achieve the best possible statistical performances in this "p larger than n" setting. A typical example is the study of l_1-based penalization methods for the estimation of linear regression models. However, if p is really too large compared to n, all these new procedures fail to achieve a good estimation.

Thus, there is a need to understand the intrinsic limitations of a statistical problem: what is the best rate of estimation or testing achievable by a procedure? Is it possible to design good procedures for arbitrarily large p or are there theoretical limitations when p becomes "too large"? These limitations tell us what kind of data analysis problems are too complex so that no statistical procedure is able to provide reasonable results. Furthermore, the knowledge of such limitations may drive the research towards areas where computationally efficient procedures are shown to be suboptimal.

1.1. Linear regression and statistical problems

We observe a response vector $\mathbf{Y} \in \mathbb{R}^n$ and a real design matrix $\mathbf{X}$ of size $n \times p$. Consider the linear regression model

$$\mathbf{Y} = \mathbf{X}\theta_0 + \boldsymbol{\epsilon} \ , \tag{1.1}$$

where the vector θ_0 of size p is unknown and the random vector $\boldsymbol{\epsilon}$ follows a centered normal distribution $\mathcal{N}(0_n, \sigma^2 I_n)$. Here, 0_n stands for the null vector of size n and I_n for the identity matrix of size n.

In some cases, the design $\mathbf{X}$ is considered as *fixed* either because it has been previously chosen or because we work conditionally to the design. In other cases, the rows of the design matrix $\mathbf{X}$ correspond to a n-sample of a random vector X of size p. The design $\mathbf{X}$ is then said to be *random*. A specific class of random design is made of Gaussian designs where X follows a centered normal distribution $\mathcal{N}(0_p, \Sigma)$. The analysis of fixed and Gaussian designs share many common points. In this work, we shall enhance the similarities and the differences between both settings.

There are various statistical problems arising in the linear regression model (1.1). Let us list the most classical issues:

($\mathbf{P_1}$): **Linear hypothesis testing**. In general, the aim is to test whether θ_0 belongs to a linear subspace of $\mathbb{R}^p$. Here, we focus on testing the null hypothesis $\mathbf{H_0}$: $\{\theta_0 = 0_p\}$. In Gaussian design, this is equivalent to testing whether $\mathbf{Y}$ is independent from $\mathbf{X}$.

($\mathbf{P_2}$): **Prediction**. We focus on predicting the expectation $\mathbb{E}[\mathbf{Y}]$ in fixed design and the conditional expectation $\mathbb{E}[\mathbf{Y}|\mathbf{X}]$ in Gaussian design.

($\mathbf{P_3}$): **Inverse problem**. The primary interest lies in estimating θ_0 itself and the corresponding loss function is $\|\widehat{\theta} - \theta_0\|_p^2$, where $\|.\|_p$ is the l_2 norm in $\mathbb{R}^p$.

($\mathbf{P_4}$): **Support estimation** aims at recovering the support of θ_0, that is the set of indices corresponding to non-zero coefficients. The easier problem of **dimension reduction** amounts to estimate a set $\widehat{M} \subset \{1, \ldots p\}$ of "reasonable" size that contains the support of θ_0 with high probability.

Much work have been devoted to these statistical questions in the so-called high-dimensional setting, where the number of covariates p is possibly much larger than n. A classical approach to perform a statistical analysis in this setting is to assume that θ_0 is *sparse*, in the sense that most of the components of θ_0 are equal to 0. For the problem of prediction ($\mathbf{P_2}$), procedures based on complexity penalization are proved to provide good risk bounds for known variance [11] and unknown variance [6] but are computationally inefficient. In contrast, convex penalization methods such as the Lasso or the Dantzig selector are fast to compute, but only provide good performances under restrictive assumptions on the design $\mathbf{X}$ (e.g. [8, 13, 50]). Exponential weighted aggregation methods [18, 40] are another example of fast and efficient methods. The l_1 penalization methods have also been analyzed for the inverse problem $(\mathbf{P_3})$ [8] and for support estimation $(\mathbf{P_4})$ [36, 49]. Dimension reduction methods are often studied in more general settings than linear regression [17, 26]. In the linear regression model, the SIS method [25] based on the correlation between the response and the covariates allows to perform dimension reduction. The problem of high-dimensional hypothesis testing $(\mathbf{P_1})$ has so far attracted less attention. Some testing procedures are discussed in [7, 3] for fixed design and in [44, 34] for Gaussian design.

1.2. Sparsity and ultra-high dimensionality

Given a positive integer k, we say that the vector θ_0 is k-sparse if θ_0 contains at most k non-zero components. We call k the sparsity parameter. In this paper, we are interested in the setting $k < n < p$. We note $\Theta[k, p]$ the set of k-sparse vectors in $\mathbb{R}^p$.

In linear regression, most of the results about classical procedures require that the triplet (k, n, p) satisfies $k[1 + \log(p/k)] < n$. When k is "small", this corresponds to assuming that p is subexponential with respect to n. The analysis of the Lasso in prediction, inverse problems [8], and support estimation [38] entail such assumptions. In dimension reduction, the SIS method [25] also requires this assumption. If the multiple testing procedure of [7] can be analyzed for $k[1+\log(p/k)]$ larger than n, it exhibits a much slower rate of testing in this case. In noiseless problems ($\sigma = 0$), compressed sensing methods [23] fail when $k[1+\log(p/k)]$ is large compared to n (see [22] for numerical illustrations). In the sequel, we say that the problem is *ultra-high dimensional*[1] when $k[1+\log(p/k)]$ is large compared to n. Observe that ultra-high dimensionality does not necessarily imply that p is exponential with respect to n. As an example, taking $p = n^3$ and $k = n/\log\log(n)$ asymptotically yields an ultra-high dimensional problem.

Why should we care about ultra-high dimensional problem? In this setting, there are so many variables that statistical questions such as the estimation of θ_0 $(\mathbf{P_3})$ or its support $(\mathbf{P_4})$ are likely to be difficult. Nevertheless, if the signal over noise ratio is large, do there exist estimators that perform relatively well? The answer is no. We prove in this paper that a phase transition phenomenon occurs in an ultra-high dimensional setting and that most of the estimation and testing problems become hopeless. This phase transition phenomenon implies that some statistical problems that are tackled in postgenomic of functional MRI cannot actually be addressed properly.

[1] In some papers, the expression ultra-high dimensional has been used to characterize problems such that $\log(p) = O(n^\beta)$ with $\beta < 1$. We argue in this paper that that as soon as $k\log(p)/n$ goes to 0, the case $\log(p) = O(n^\beta)$ is not intrinsically more difficult than conditions such as $p = O(n^\delta)$ with $\delta > 0$.

Example 1.1 (Motivating example)**.** In some gene network inference problems (e.g. [16]), the number p of genes can be as large as 5000 while the number n of microarray experiments is only of order 50. Let us consider a gene A. We note $\mathcal{G}_A$ the set of genes that interact with the gene A and k stands for the cardinality of $\mathcal{G}_A$. How large can be k so that it is still "reasonable" to estimate $\mathcal{G}_A$ from the microarray experiments? In statistical terms, inferring the set of genes interacting with A amounts to estimate the support of a vector θ_0 in a linear regression model (see e.g. [38]). Our answer is that if k is larger than 4, then the problem of network estimation becomes extremely difficult. We will come back to this example and explain this answer in Section 7.

1.3. Minimax risks

A classical way to assess the performance of an estimator $\widehat{\theta}$ is to consider its maximal risk over a class $\Theta \subset \mathbb{R}^p$. This is the minimax point of view. For the time being, we only define the notions of minimaxity for estimation problems ($\mathbf{P_2}$ and $\mathbf{P_3}$). Their counterpart in the case of testing ($\mathbf{P_1}$) and dimension reduction ($\mathbf{P_4}$) will be introduced in subsequent sections. Given a loss function $l(.,.)$ and estimator $\widehat{\theta}$, the maximal risk of $\widehat{\theta}$ over $\Theta[k,p]$ for a design $\mathbf{X}$ (or a covariance Σ in the Gaussian design case) and a variance σ^2 is defined by $\sup_{\theta_0\in\Theta[k,p]} \mathbb{E}_{\theta_0,\sigma}[l(\widehat{\theta},\theta_0)]$. Taking the infimum of the maximal risk over all possible estimators $\widehat{\theta}$, we obtain the *minimax risk*

$$\inf_{\widehat{\theta}} \sup_{\theta_0\in\Theta[k,p]} \mathbb{E}_{\theta_0,\sigma}[l(\widehat{\theta},\theta_0)] \ .$$

We say that an estimator $\widehat{\theta}$ is minimax if its maximal risk over $\Theta[k,p]$ is close to the minimax risk.

In practice, we do not know the number k of non-zero components of θ_0 and we seldom know the variance σ^2 of the error. If an estimator $\widehat{\theta}$ does not require the knowledge of k and nearly achieves the minimax risk over $\Theta[k,p]$ for a range of k, we say that $\widehat{\theta}$ is adaptive to the sparsity. Similarly, an estimator $\widehat{\theta}$ is adaptive to the variance σ^2, if it does not require the knowledge of σ^2 and nearly achieves the minimax risk for all $\sigma^2 > 0$. When possible, the main challenge is to build adaptive procedures. In some statistical problems considered here, adaptation is in fact impossible and there is an unavoidable loss when the variance or the sparsity parameter is unknown. In such situations, it is interesting to quantify this unavoidable loss.

1.4. Our contribution and related work

In the specific case of the Gaussian sequence model, where $n = p$ and $\mathbf{X} = I_n$, the minimax risks over k-sparse vectors have been studied for a long time. Donoho and Johnstone [21, 35] have provided the asymptotic minimax risks of prediction ($\mathbf{P_2}$). Baraud [5] has studied the optimal rate of testing from a non-asymptotic point of view while Ingster [31, 32, 33] has provided the asymptotic optimal rate of testing with exact constants.

Recently, some high-dimensional problems have been studied from a minimax point of view. Wainwright [45, 46] provides minimax lower bounds for the problem of support estimation ($\mathbf{P_4}$). Raskutti et al. [39] and Rigollet and Tsybakov [40] have provided minimax upper bounds and lower bounds for ($\mathbf{P_2}$) and ($\mathbf{P_3}$) over l_q balls for general fixed designs $\mathbf{X}$ when the variance σ^2 is known (see also Ye and Zhang [47] and Abramovich and Grinshtein [1]). Arias-Castro et al. [3] and Ingster et al. [34] have computed the asymptotic minimax detection boundaries for the testing problem ($\mathbf{P_1}$) for some specific designs. However, their study only encompasses reasonable dimensional problems (p grows polynomially with n). Some minimax lower bounds have also been stated for testing ($\mathbf{P_1}$) and prediction ($\mathbf{P_2}$) problems with Gaussian design [42, 44]. All the aforementioned results do not cover the ultra-high dimensional case and do not tackle the problem of adaptation to both k and σ.

This paper provides minimax lower bounds and upper bounds for the problems ($\mathbf{P_1}$), ($\mathbf{P_2}$), ($\mathbf{P_3}$) when the regression vector θ_0 is k-sparse for fixed and random designs, known and unknown variance, known and unknown sparsities. The lower and upper bounds match up to possible differences in the logarithmic terms. The main discoveries are the following:

1. **Phase transition in an ultra-high dimensional setting**. Contrary to previous work, our results cover both the high-dimensional and ultra-high dimensional setting. We establish that for each of the problems ($\mathbf{P_1}$), ($\mathbf{P_2}$) and ($\mathbf{P_3}$), an elbow effect occurs when $k\log(p/k)$ becomes large compared to n. Let us emphasize the difference between the high-dimensional and the ultra-high dimensional regimes for two problems: prediction ($\mathbf{P_2}$) and support estimation ($\mathbf{P_4}$).

   Prediction with random design. In the (non-ultra) high-dimensional setting, the minimax risk of prediction for a random design regression is of order $\sigma^2 k\log(p/k)/n$ (see Section 3). Thus, the effect of the sparsity k is linear and the effect of the number of variables p is logarithmic. In an ultra-high dimensional setting, that is when $k\log(p/k)/n$ is large, we establish that an elbow effect occurs in the minimax risk. In this setting, the minimax risk becomes of order $\sigma^2\exp[Ck\{1+\log(p/k)\}/n]$, where C is a positive constant: it grows exponentially fast with k and polynomially with p (see the red curve in Figure 1). If it was expected that the minimax risk cannot be small for such problems, we prove here that the minimax risk is in fact exponentially larger than the usual $k\log(p/k)/n$ term.

   Support estimation. In a non-ultra high dimensional setting it is known [46] that under some assumptions on the design $\mathbf{X}$ (e.g. each component of $\mathbf{X}$ is drawn from iid. standard normal distribution) the support of a k-sparse vector θ_0 is recoverable with high probability if

$$\forall i \in \text{supp}(\theta_0)\,, \quad |(\theta_0)_i| \geq C\sqrt{\log(p)/n}\sigma\ , \tag{1.2}$$

where C is a numerical constant. In an ultra-high dimensional setting, even if

$$\forall i \in \mathrm{supp}(\theta_0)\,, \quad |(\theta_0)_i| = \exp[Ck\{1+\log(p/k)\}/n]/\sqrt{k}\sigma\ , \tag{1.3}$$

it is not possible to estimate the support of θ_0 with high probability. Observe that the condition (1.3) is much stronger than (1.2). In fact, it is not even possible to reduce drastically the dimension of the problem without forgetting relevant variables with positive probability. More precisely, for any dimension reduction procedure that selects a subset of variables $\widehat{M} \subset \{1, \dots p\}$ of size p^δ with some $0 < \delta < 1$ (described in Proposition 6.7), we have $\mathrm{supp}(\theta_0) \nsubseteq \widehat{M}$ with probability away from zero (see Proposition 6.7). Thus, it is almost hopeless to have a reliable estimation of the support of θ_0 even if $\|\theta_0\|_p^2/\sigma^2$ is large. This impossibility of dimension reduction for ultra-high dimensional problems is numerically illustrated in Section 7.

2. **Adaptation to the sparsity k and to the variance σ^2**. Most theoretical results for the problems $(\mathbf{P_1})$ and $(\mathbf{P_2})$ require that the variance σ^2 is known. Here, we establish these minimax bounds for both known and unknown variance and known and unknown sparsity. The knowledge of the variance is proved to play a fundamental role for the testing problem $(\mathbf{P_1})$ when $k[1+\log(p/k)]$ is large compared to $\sqrt{n}$. The knowledge of σ^2 is also proved to be crucial for $(\mathbf{P_2})$ in an ultra-high dimensional setting. Thus, specific work is needed to develop fast and efficient procedures that do not require the knowledge of the variance. Furthermore, variance estimation is extremely difficult in an ultra-high dimensional setting.
3. **Effect of the design**. Lastly, the minimax bounds of $(\mathbf{P_1})$, $(\mathbf{P_2})$ and $(\mathbf{P_3})$ are established for fixed and Gaussian designs. Except for the problem of prediction $(\mathbf{P_2})$, the minimax risks are shown to be of the same nature for both forms of the design. Furthermore, we investigate the dependency of the minimax risks on the design $\mathbf{X}$ (resp. Σ) in Sections 4-6.

The minimax bounds stated in this paper are non asymptotic. While some upper bounds are consequences of recent results in the literature, most of the effort is spent here to derive the lower bound. These bounds rely on Fano's and Le Cam's methods [48] and on geometric considerations. In each case, near optimal procedures are exhibited.

1.5. Organization of the paper

In Section 3, we summarize the minimax bounds for specific designs called "worst-case" and "best-case" designs in order to emphasize the effects of dimensionality. The general results are stated in Section 4 for the tests and Section 5 for the problem of prediction. The problems of inverse estimation, support estimation, and dimension reduction are studied in Section 6. In Section 7, we address the following practical question: For exactly what range of (k, p, n)

should we consider a statistical problem as ultra-high dimensional? A small simulation study illustrates this answer. Section 8 contains the final discussion and side results about variance estimation. Section 9 is devoted to the proof of the mains minimax lower bounds. Specific statistical procedures allow to establish the minimax upper bounds. Most of these procedures are used as theoretical tools but should not be applied in a high dimensional setting because they are computationally inefficient. In order to clarify the statements of the results in Sections 4–6, we postpone the definition of these procedures to Section 10. The remaining proofs are described in a technical appendix [43].

2. Notations and preliminaries

We respectively note $\|.\|_n$ and $\|.\|_p$ the l_2 norms in $\mathbb{R}^n$ and $\mathbb{R}^p$, while $\langle . \rangle_n$ refers to the inner product in $\mathbb{R}^n$. For any $\theta_0 \in \mathbb{R}^p$ and $\sigma > 0$, $\mathbb{P}_{\theta_0,\sigma}$ and $\mathbb{E}_{\theta_0,\sigma}$ refer to the joint distribution of $(\mathbf{Y}, \mathbf{X})$. When there is no risk of confusion, we simply write $\mathbb{P}$ and $\mathbb{E}$. All references with a capital letter such as Section A or Eq. (B.2) refer to the technical Appendix [43].

In the sequel, we note $\text{supp}(\theta_0)$ the support of θ_0. For any $1 \leq k \leq p$, $\mathcal{M}(k,p)$ stands for the collections of all subsets of $\{1,\ldots,p\}$ with cardinality k. Given $i \in \{1,\ldots,p\}$, we note $\mathbf{X}_i$ the vector of size n corresponding to i-th column of $\mathbf{X}$. For $m \subset \{1,\ldots,p\}$, $\mathbf{X}_m$ stands for the $n \times |m|$ submatrix of $\mathbf{X}$ that contains the columns $\mathbf{X}_i$, $i \in m$. In what follows, we note $\mathbf{X}^T$ the transposed matrix of $\mathbf{X}$.

Gaussian design and conditional distribution. When the design is said to be "Gaussian", the n rows of $\mathbf{X}$ are n independent samples of a random row vector X such that $X^T \sim \mathcal{N}(0_p, \Sigma)$. Thus, $(\mathbf{Y}, \mathbf{X})$ if a n-sample of the random vector $(Y, X^T) \in \mathbb{R}^{p+1}$, where Y is defined by

$$Y = X\theta_0 + \epsilon \ , \tag{2.1}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. The linear regression model with *Gaussian* design is relevant to understand the conditional distribution of a Gaussian variable Y conditionally to a Gaussian vector since $\mathbb{E}[Y|X] = X\theta_0$ and $\text{Var}(Y|X) = \sigma^2$. This is why we shall often refer to σ^2 as the conditional variance of Y when considering Gaussian design. This model is also closely connected to the estimation of Gaussian graphical models [38, 44].

As explained later, the minimax risk over $\Theta[k,p]$ strongly depends on the design $\mathbf{X}$. This is why we introduce some relevant quantities on $\mathbf{X}$.

Definition 2.1. Consider some integer $k > 0$ and some design $\mathbf{X}$.

$$\Phi_{k,+}(\mathbf{X}) := \sup_{\theta \in \Theta[k,p] \setminus \{0_p\}} \frac{\|\mathbf{X}\theta\|_n^2}{\|\theta\|_p^2} \quad \text{and} \quad \Phi_{k,-}(\mathbf{X}) := \inf_{\theta \in \Theta[k,p] \setminus \{0_p\}} \frac{\|\mathbf{X}\theta\|_n^2}{\|\theta\|_p^2} . \tag{2.2}$$

In fact, $\Phi_{k,+}(\mathbf{X})$ and $\Phi_{k,-}(\mathbf{X})$ respectively correspond to the largest and the smallest restricted eigenvalue of order k of $\mathbf{X}^T\mathbf{X}$.

Given a symmetric real square matrix A, $\varphi_{\max}(A)$ stands for the largest eigenvalue of A. Finally, $C, C_1, \ldots$ denote positive universal constants that may vary from line to line. The notation $C(.)$ specifies the dependency on some quantities.

In the propositions, the constants involved in the assumptions are not always expressly specified. For instance, sentences of the form "Assume that $n \geq C$. Then, ..." mean that "There exists an universal $C > 0$ such that if $n \geq C$, then ...".

3. Main results

The exact bounds are stated in Section 4–6. In order to explain these results, we now summarize the main minimax bounds by focusing on the role of (k, n, p) rather than on the dependency on the design $\mathbf{X}$. In order to keep the notations short, we do not provide in this section the minimal assumptions of the results. Let us simply mention that all of them are valid if the sparsity k satisfies $k \leq (p^{1/3}) \wedge (n/5)$ and that $p \geq n \geq C$ where C a positive numerical constant.

3.1. Prediction

3.1.1. Definitions

First, the results are described for the problem of prediction $(\mathbf{P_2})$ since the problem of minimax estimation is more classical in this setting. Different prediction loss functions are used for fixed and Gaussian designs. When the design is considered as fixed, we study the loss $\|\mathbf{X}(\theta_1 - \theta_2)\|_n^2/(n\sigma^2)$. For Gaussian design, we consider the integrated prediction loss function:

$$\|\sqrt{\Sigma}(\theta_1 - \theta_2)\|_p^2/\sigma^2 = \mathbb{E}\left[\{X(\theta_1 - \theta_2)\}^2\right]/\sigma^2 \ . \tag{3.1}$$

Given a design $\mathbf{X}$, the minimax risk of prediction over $\Theta[k, p]$ with respect to $\mathbf{X}$ is

$$\mathcal{R}_F[k, \mathbf{X}] = \inf_{\widehat{\theta}} \sup_{\theta_0 \in \Theta[k,p]} \mathbb{E}_{\theta_0,\sigma}[\|\mathbf{X}(\widehat{\theta} - \theta_0)\|_n^2/(n\sigma^2)] \ . \tag{3.2}$$

For a Gaussian design with covariance Σ, we study the quantity

$$\mathcal{R}_R[k, \Sigma] := \inf_{\widehat{\theta}} \sup_{\theta_0 \in \Theta[k,p]} \mathbb{E}_{\theta_0,\sigma}[\|\sqrt{\Sigma}(\widehat{\theta} - \theta_0)\|_p^2/\sigma^2] \ . \tag{3.3}$$

These minimax risks of prediction do not only depend on (k, n, p) but also on the design $\mathbf{X}$ (or on the covariance Σ). The computation of the exact dependency of the minimax risks on $\mathbf{X}$ or Σ is a challenging question. To simplify the presentation in this section, we only describe the minimax prediction risks for worst-case designs defined by

$$\mathcal{R}_F[k] := \sup_{\mathbf{X}} \mathcal{R}_F[k, \mathbf{X}], \quad \mathcal{R}_R[k] := \sup_{\Sigma} \mathcal{R}_R[k, \Sigma] \ , \tag{3.4}$$

the supremum being taken over all designs $\mathbf{X}$ of size $n \times p$ (resp. all covariance matrices Σ). The quantity $\mathcal{R}_F[k]$ corresponds to the smallest risk achievable *uniformly* over $\Theta[k,p]$ *and* all designs $\mathbf{X}$. It is shown in Section 5 that the quantity $\mathcal{R}_R[k]$ is achieved (up to constants) for a covariance $\Sigma = I_p$ while the quantity $\mathcal{R}_F[k]$ is achieved with high probability for designs $\mathbf{X}$ that are realizations of the standard Gaussian design (all the components of $\mathbf{X}$ are drawn independently from a standard normal distribution). This corresponds to designs used in compressed sensing [23]. In fact, the maximal risks $\mathcal{R}_F[k]$ and $\mathcal{R}_R[k]$ for the prediction problem correspond to typical situations where the designs is well-balanced, that is as close as possible to orthogonality.

3.1.2. Results

In the sequel, we say that $\mathcal{R}_F[k]$ is *of order* $f(k,p,n,C)$, where C is positive constant when there exist two positive universal constants C_1 and C_2 such that

$$f(k,p,n,C_1) \leq \mathcal{R}_F[k] \leq f(k,p,n,C_2) \ .$$

These minimax risks are computed in Section 5 and are gathered in Table 1. They are also depicted on Figure 1.

When $k\log(p/k)$ remains small compared to n, the minimax risk of prediction is of the same order for fixed and Gaussian design. The $k\log(p/k)/n$ risk is classical and has been known for a long time in the specific case of the Gaussian sequence model [35]. Some procedures based on complexity penalization or aggregation (e.g. [11]) are proved to achieve these risks uniformly over all designs $\mathbf{X}$. Computationally efficient procedures like the Lasso or the Dantzig selector are only proved to achieve a $k\log(p)/n$ risk under assumption on the

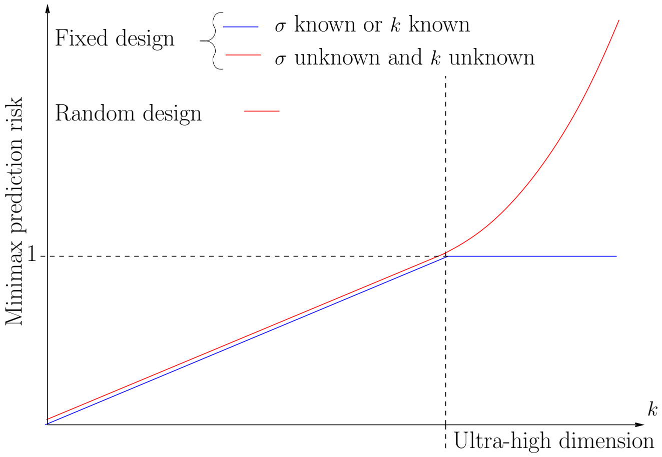

Fig 1. *Minimax prediction risk ($\mathbf{P_2}$) over $\Theta[k,p]$ as a function of k for fixed and random design and known and unknown variance. The corresponding bounds are stated in Section 5.*

Table 1
Orders of magnitude of the minimax risks of prediction $\mathcal{R}_F[k]$ and $\mathcal{R}_R[k]$ over $\Theta[k,p]$

Fixed Design: $\mathcal{R}_F[k]$	**Gaussian** Design: $\mathcal{R}_R[k]$
$C\left[\frac{k}{n}\log\left(\frac{p}{k}\right)\right]\wedge 1$	$C_1\frac{k}{n}\log\left(\frac{p}{k}\right)\exp\left[C_2\frac{k}{n}\log\left(\frac{p}{k}\right)\right]$

design $\mathbf{X}$ [8]. If the support of θ_0 is known in advance, the parametric risk is of order k/n. Thus, the price to pay for not knowing the support of θ_0 is only logarithmic in p.

In an ultra-high dimensional setting, the minimax prediction risk in fixed designs remains smaller than one. It is the minimax risk of estimation of the vector $\mathbb{E}(\mathbf{Y})$ of size n. This means that the sparsity index k does not play anymore a role in ultra-high dimension. For a Gaussian design, the minimax prediction risk becomes of order $C_1(p/k)^{C_2k/n}$: it increases exponentially fast with respect to k and polynomially fast with respect to p. Comparing this risk with the parametric rate k/n, we observe that the price to pay for not knowing the support of θ_0 is now far higher than $\log(p)$.

In Section 5, we also study the adaptation to the sparsity index k and to the variance σ^2. We prove that adaptation to k and σ^2 is possible for a Gaussian design. In fixed design, no procedure can be simultaneously adaptive to the sparsity k and the variance σ^2 (see the red curve in Figure 1 that corresponds to fixed design, σ and k unknown).

3.2. Testing

3.2.1. Definitions

Let us turn to the problem $(\mathbf{P_1})$ of testing $\mathbf{H_0}$: $\{\theta_0 = 0_p\}$ against $\mathbf{H_1}$: $\{\theta_0 \in \Theta[k,p]\setminus\{0_p\}\}$. We fix a level $\alpha > 0$ and a type II error probability $\delta > 0$. Minimax lower and upper bounds for this problem are discussed in Section 4.

Suppose we are given a test procedure Φ_α of level α for fixed design $\mathbf{X}$ and known variance σ^2. The δ-separation distance of Φ_α over $\Theta[k,p]$, noted $\rho_F[\Phi_\alpha,k,\mathbf{X}]$ is the minimal number ρ, such that Φ_α rejects $\mathbf{H_0}$ with probability larger than $1-\delta$ if $\|\mathbf{X}\theta_0\|_n/\sqrt{n}\geq\rho\sigma$. Hence, $\rho_F[\Phi_\alpha,k,\mathbf{X}]$ corresponds to the minimal distance such that the hypotheses $\{\theta_0 = 0_p\}$ and $\{\theta_0\in\Theta[k,p],\ \|\mathbf{X}\theta_0\|_n^2\geq n\rho_F^2[\Phi_\alpha,k,\mathbf{X}]\sigma^2\}$ are well separated by the test Φ_α.

$$\rho_F[\Phi_\alpha,k,\mathbf{X}] := \inf\left\{\rho>0,\ \inf_{\theta_0\in\Theta[k,p],\ \|\mathbf{X}\theta_0\|_n\geq\sqrt{n}\rho\sigma}\mathbb{P}_{\theta_0,\sigma}[\Phi_\alpha=1]\geq 1-\delta\right\}\ .$$

Although the separation distance also depends on δ, n, and p, we only write $\rho_F[\Phi_\alpha,k,\mathbf{X}]$ for the sake of conciseness. By definition, the test Φ_α has a power larger than $1-\delta$ for $\theta_0\in\Theta[k,p]$ such that $\|\mathbf{X}\theta_0\|_n^2\geq\rho_F^2[\Phi_\alpha,k,\mathbf{X}]$. Then, we consider

$$\rho_F^*[k,\mathbf{X}] := \inf_{\Phi_\alpha}\rho[\Phi_\alpha,k,\mathbf{X}]\ . \tag{3.5}$$

The infimum runs over all level-α tests. We call this quantity the (α,δ)-minimax separation distance over $\Theta[k,p]$ with design $\mathbf{X}$ and variance σ^2. The minimax separation distance is a non-asymptotic counterpart of the detection boundaries studied in the Gaussian sequence model [20].

Similarly, we define the (α,δ)-minimax separation distance over $\Theta[k,p]$ with Gaussian design by replacing the distance $\|\mathbf{X}\theta_0\|_n/\sqrt{n}$ by the distance $\|\sqrt{\Sigma}\theta_0\|_p$:

$$\begin{aligned}\rho_R[\Phi_\alpha,k,\Sigma] &:= \inf\left\{\rho>0,\ \inf_{\theta_0\in\Theta[k,p],\ \|\sqrt{\Sigma}\theta_0\|_p\geq\rho\sigma}\mathbb{P}_{\theta_0,\sigma}[\Phi_\alpha=1]\geq 1-\delta\right\}\ ,\\ \rho^*_R[k,\Sigma] &:= \inf_{\Phi_\alpha}\rho_R[\Phi_\alpha,k,\Sigma]\ . \qquad (3.6)\end{aligned}$$

Various bounds on $\rho^*_F[k,\mathbf{X}]$, $\rho^*_R[k,\Sigma]$ are stated in Section 4. In this section, we only provide the orders of magnitude of the minimax separation distances in the "worst case" designs in order to emphasize the effect of dimensionality:

$$\rho^*_F[k]:=\sup_{\mathbf{X}}\rho^*_F[k,\mathbf{X}]\ ,\qquad \rho^*_R[k]:=\sup_{\Sigma}\rho^*_R[k,\Sigma]\ . \qquad (3.7)$$

This is the smallest separation distance that can be achieved by a procedure Φ_α *uniformly* over all designs $\mathbf{X}$ (resp. Σ). As for the prediction problem, it will be proved in Section 4, that the quantity $\rho^*_F[k]$ and $\rho^*_R[k]$ are achieved for well-balanced designs.

It is not always possible to achieve the minimax separation distances with a procedure Φ_α that *does not require* the knowledge of the variance σ^2. This is why we also consider $\rho^*_{F,U}[k]$ and $\rho^*_{R,U}[k]$ the minimax separation distance for fixed and Gaussian design when the variance is unknown. Roughly, $\rho^*_{F,U}[k]$ corresponds to the minimal distances ρ^2 that allows to separate well the hypotheses $\{\theta_0=0_p\ \text{ and }\ \sigma>0\}$ and $\{\theta_0\in\Theta[k,p]\text{ and }\sigma>0,\ \|\mathbf{X}\theta_0\|_n^2/\sigma^2\geq n\rho^2\}$ when σ is unknown. We shall provide a formal definition at the beginning of Section 4.

3.2.2. Results

In Table 2, we provide the orders of the minimax separation distances over $\Theta[k,p]$ for fixed and Gaussian designs, known and unknown variance (see also Figure 2).

In contrast to $(\mathbf{P_2})$, the minimax separation distances are of the same order for fixed and Gaussian design.

TABLE 2
Order of the minimax separation distances over $\Theta[k,p]$ for fixed and Gaussian design, known and unknown variance: $(\rho^*_F[k])^2$, $(\rho^*_R[k])^2$, $(\rho^*_{F,U}[k])^2$, and $(\rho^*_{R,U}[k])^2$

	Fixed and **Gaussian** Design
Known σ^2: $(\rho^*_F[k])^2$ and $(\rho^*_R[k])^2$	$C(\alpha,\delta)\frac{k\log(p)}{n}\wedge\frac{1}{\sqrt{n}}$
Unknown σ^2: $(\rho^*_{F,U}[k])^2$ and $(\rho^*_{R,U}[k])^2$	$C(\alpha,\delta)\frac{k\log(p)}{n}\exp\left[C_2(\alpha,\delta)\frac{k\log(p)}{n}\right]$

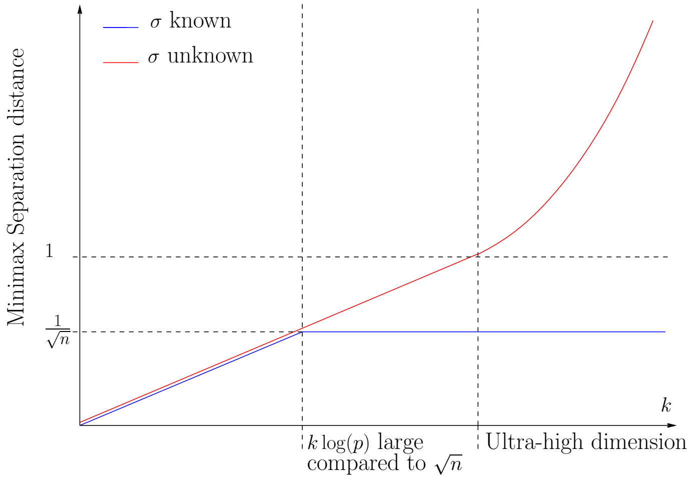

FIG 2. *Orders of magnitude of the minimax separation distances $(\rho^*_F[k])^2$, $(\rho^*_R[k])^2$, $(\rho^*_{F,U}[k])^2$ and $(\rho^*_{R,U}[k])^2$ over $\Theta[k,p]$ ($\mathbf{P_1}$) for fixed and random designs and known and unknown variances. Here, $\rho^*_F[k]$ and $\rho^*_R[k]$ behave similarly while $\rho^*_{F,U}[k]$ and $\rho^*_{R,U}[k]$ behave similarly. The corresponding bounds are stated in Section 4.*

1. When $k\log(p) \leq \sqrt{n}$, all the minimax separation distances are of order $k\log(p)/n$. This quantity also corresponds to the minimax risk of prediction ($\mathbf{P_2}$) stated in the previous subsection. This separation distance has already been proved in the specific case of the Gaussian sequence model [5, 20].
2. When $k\log(p) \geq \sqrt{n}$, the minimax separation distances are different under known and unknown variance. If the variance is known, the minimax separation distance over $\Theta[k,p]$ stays of order $1/\sqrt{n}$. Here, $1/\sqrt{n}$ corresponds in fixed design to the minimax separation distance of the hypotheses $\{\mathbb{E}[\mathbf{Y}] = 0_n\}$ against the general hypothesis $\{\mathbb{E}[\mathbf{Y}] \neq 0_n\}$ for known variance (see Baraud [5]).
3. If the variance is unknown, the minimax separation distance over $\Theta[k,p]$ is still of order $k\log(p)/n$ if $k\log(p)$ is small compared to n. In contrast, the minimax separation distance blows up to the order $C_1p^{C_2k/n}$ in a ultra-high dimensional setting. This blow up phenomenon has also been observed in the previous section for the problem of prediction ($\mathbf{P_2}$) in Gaussian design. In conclusion, the knowledge of the variance is of great importance for $k\log(p)$ larger than $\sqrt{n}$.

3.3. Inverse problem and support estimation

3.3.1. Definitions

In the inverse problem ($\mathbf{P_3}$), we are primarily interested in the estimation of θ_0 rather than $\mathbf{X}\theta_0$. This is why the loss function under study is $\|\theta_1 - \theta_2\|_p^2$.

Minimax lower and upper bounds for this loss function are discussed in Section 6. For a fixed design $\mathbf{X}$, the minimax risk of estimation is

$$\mathcal{RI}_F[k,\mathbf{X}] := \inf_{\widehat{\theta}} \sup_{\theta_0\in\Theta[k,p]} \mathbb{E}_{\theta_0,\sigma}[\|\theta_0-\widehat{\theta}\|_p^2/\sigma^2] \ . \tag{3.8}$$

If one transforms the design $\mathbf{X}$ by an homothety of factor $\lambda > 0$, then this multiplies the minimax risk for the inverse problem by a factor $1/\lambda^2$. For the sake of simplicity, we restrict ourselves to designs $\mathbf{X}$ such that each column has been normed to $\sqrt{n}$. The collection of such designs is noted $\mathcal{D}_{n,p}$. The supremum of the minimax risks over the designs $\mathcal{D}_{n,p}$ is $+\infty$. Take for instance a design where the two first columns are equal. In this section, we only present the infimum of the minimax risks over $\Theta[k,p]$ as $\mathbf{X}$ varies across $\mathcal{D}_{n,p}$:

$$\mathcal{RI}_F[k] := \inf_{\mathbf{X}\in\mathcal{D}_{n,p}} \mathcal{RI}_F[k,\mathbf{X}] \ .$$

The quantity $\mathcal{RI}_F[k]$ is interpreted the following way: given (k,n,p) what is the smallest risk we can hope if we use the best possible design? Alternatively, given n observations, what is the intrinsic difficulty of estimating a k-sparse vector of size p? We call this quantity the minimax risks for the inverse problem over $\Theta[k,p]$.

In Section 6, we also study the corresponding the minimax risks of the inverse problem in the random design case. Let $\mathcal{S}_p$ stand for the set of covariance matrices that contain only ones on the diagonal. We respectively define the minimax risk of estimation over $\Theta[k,p]$ for a covariance Σ and the minimax risk of estimation over $\Theta[k,p]$ as

$$\mathcal{RI}_R[k,\Sigma] := \inf_{\widehat{\theta}} \sup_{\theta_0\in\Theta[k,p]} \mathbb{E}_{\theta_0,\sigma}[\|\theta_0-\widehat{\theta}\|_p^2/\sigma^2] \quad \text{and} \quad \mathcal{RI}_R[k] := \inf_{\Sigma\in\mathcal{S}_p} \mathcal{RI}_R[k,\Sigma] \ . \tag{3.9}$$

3.3.2. Results

In Table 3, we provide the minimax risks in fixed design for different values of (k,n,p) (see also Figure 3).

If $k\log(p/k)$ remains smaller than n, it is possible to recover the risk $Ck\log(p/k)$ for "good" designs. This risk is for instance achieved by the Dantzig selector of Candès and Tao [15] for nearly-orthogonal designs, that roughly means that the restricted eigenvalues $\Phi_{3k,+}(\mathbf{X})$ and $\Phi_{3k,-}(\mathbf{X})$ of $\mathbf{X}^T\mathbf{X}$ are close to one. In

TABLE 3
Order of the minimax risks $\mathcal{RI}_F[k]$ for the inverse problem over $\Theta[k,p]$

$(\mathbf{k},\mathbf{n},\mathbf{p})$	$k\log(p)\leq Cn$	$k\log(p)\gg n\log(n)$
Minimax risk $\mathcal{RI}_F[k]$	$C\frac{k}{n}\log\left(\frac{p}{k}\right)$	$\exp\left[C'\frac{k}{n}\log\left(\frac{p}{k}\right)\right]$.

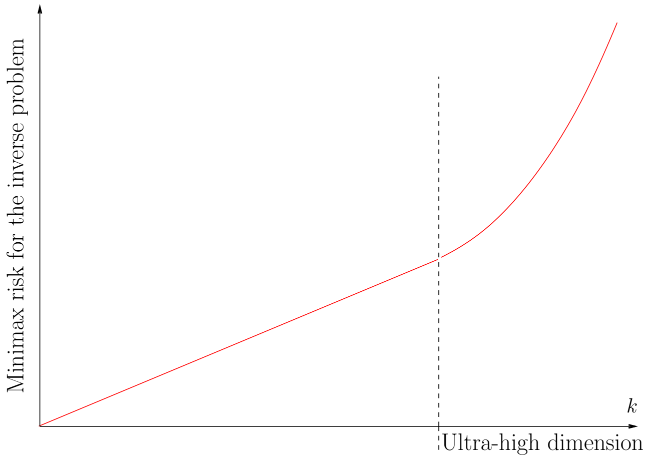

FIG 3. *Order of magnitude of the minimax risk $\mathcal{RI}_F[k]$ for the inverse problem ($\mathbf{P_3}$) over $\Theta[k,p]$ as a function of k. The corresponding bounds are stated in Section 6.*

an ultra high-dimensional setting, it is not anymore possible to build nearly-orthogonal designs $\mathbf{X}$ and the minimax risk of the inverse problem blows up as for testing problems ($\mathbf{P_1}$) or prediction problems in Gaussian design ($\mathbf{P_2}$). Moreover, adaptation to the sparsity k and to the variance σ^2 is possible for the inverse problem. As explained in Section 6, the quantities $\mathcal{RI}_R[k,\Sigma]$ and $\mathcal{RI}_R[k]$ behave somewhat similarly to their fixed design counterpart.

In Section 6, we also discuss the consequences of the minimax bounds on the problem of support estimation ($\mathbf{P_4}$). We prove that, in an ultra-high dimensional setting, it is not possible to estimate with high probability the support of θ_0 unless the ratio $\|\theta_0\|_p^2/\sigma^2$ is larger than $C_1(p/k)^{C_2k/n}$. In fact, even the problems of support estimation is almost hopeless in an ultra-high dimensional setting.

4. Hypothesis testing

We start by the testing problem ($\mathbf{P_1}$) because some minimax lower bounds in prediction and inverse estimation derive from testing considerations.

4.1. Known variance

4.1.1. Gaussian design

As mentioned in the introduction, the knowledge of $\sigma^2 = \mathrm{Var}(Y|X)$ is really unlikely in many practical applications. Nevertheless, we study this case to enhance the differences between known and unknown conditional variances. Furthermore, these results turn out to be useful for analyzing the minimax separation distances in fixed design problems. We recall that the notions of minimax separation distances $\rho^*_F[k,\mathbf{X}]$, $\rho^*_F[k]$, $\rho^*_R[k,\Sigma]$, and $\rho^*_R[k]$ have been defined in Section 3.2.

Theorem 4.1. *Assume that $\alpha+\delta\leq 53\%$, $p\geq n^2$, and that $n\geq 8\log(2/\delta)$. For any $1\leq k\leq n$, the (α,δ)-minimax separation distance (3.6) with covariance I_p is lower bounded by*

$$(\rho^*_R[k,I_p])^2\geq C\left[\frac{k}{n}\log(p)\wedge\frac{1}{\sqrt{n}}\right]\ . \tag{4.1}$$

For any $1\leq k\leq p$ and any covariance Σ, we have

$$(\rho^*_R[k,\Sigma])^2\leq C(\alpha,\delta)\left[\frac{k}{n}\log(p)\wedge\frac{1}{\sqrt{n}}\right]\ . \tag{4.2}$$

Furthermore, this upper bound is simultaneously achieved for all k and Σ by a procedure T^*_α (defined in Section 10.1.1).

Remark 4.1 (Adaptation to sparsity)**.** It follows from Theorem 4.1 that adaptation to the sparsity is possible and that the optimal optimal separation distance is of order

$$\frac{k}{n}\log(p)\wedge\frac{1}{\sqrt{n}}\ , \tag{4.3}$$

for all sparsities k between 1 and n.

Remark 4.2 (Correlated design)**.** The upper bound (4.2) is valid for any covariance matrix Σ. In contrast, the minimax lower bound (4.1) is restricted to the case $\Sigma=I_p$. This implies that there exists some constant $C(\alpha,\delta)$ such that,

$$\rho^*_R[k,I_p]\geq C(\alpha,\delta)\sup_{\Sigma}\rho^*_R[k,\Sigma]=C(\alpha,\delta)\rho^*_R[k]\ .$$

In other words, the testing problem is more complex (up to constants) for an independent design than for a correlated design.

Remark 4.3 (Which logarithmic term in the bound: $\log(p)$ or $\log(p/k)$?)**.** In the proof of Theorem 4.1, we derive the following bounds

$$\begin{aligned}(\rho^*_R[k,I_p])^2 &\geq C\left[\frac{k}{n}\log\left(1+\frac{p}{k^2}\right)\wedge\frac{1}{\sqrt{n}}\right]\ ,\\ (\rho^*_R[k,\Sigma])^2 &\leq C(\alpha,\delta)\left[\frac{k}{n}\log\left(\frac{ep}{k}\right)\wedge\frac{1}{\sqrt{n}}\right]\ .\end{aligned}$$

These two bounds are of order of (4.3) as it is assumed that $p\geq n^2$. However, the dependency of the logarithmic terms on k in the last bounds do not allow to provide the minimax separation distance when $p=n$ and k is close to $\sqrt{n}$. For instance, if $p=n$ and $k=\sqrt{n}/\log(n)$, the two bounds only match up to a factor $\log(n)/\log\log(n)$. The non-asymptotic minimax bounds of Baraud [5] in the Gaussian sequence model suffer the same weakness. Up to our knowledge the dependency on $\log(k)$ of the minimax separation distances has only been captured in an asymptotic setting [3, 34] ($(k,p,n)\to\infty$).

4.1.2. Fixed design

The separation distances are similar to the Gaussian design case.

Theorem 4.2. *Assume that $\alpha + \delta \leq 33\%$, $p \geq n^2 \geq C(\alpha, \delta)$, and that $n \geq 8\log(2/\delta)$. For any $1 \leq k \leq n$, there exist some $n \times p$ designs $\mathbf{X}$ such that*

$$(\rho^*_F[k, \mathbf{X}])^2 \geq C \left[\frac{k}{n} \log(p) \wedge \frac{1}{\sqrt{n}}\right] . \tag{4.4}$$

For any $1 \leq k \leq p$ and any design $\mathbf{X}$, we have

$$(\rho^*_F[k, \mathbf{X}])^2 \leq C(\alpha, \delta) \left[\frac{k}{n} \log(p) \wedge \frac{1}{\sqrt{n}}\right] . \tag{4.5}$$

Furthermore, this upper bound is simultaneously achieved for all k and $\mathbf{X}$ by a procedure T^*_α (defined in Section 10.1.1).

As for the random design case, we conclude that adaptation to the sparsity is possible and that $(\rho^*_F[k])^2$ is of order $\frac{k}{n} \log(p) \wedge \frac{1}{\sqrt{n}}$. In fact, the proof shows that, with large probability, designs $\mathbf{X}$ whose components are independently sampled from a standard normal variable satisfy (4.4).

Arias-Castro et al. [3] and Ingster et al. [34] have recently provided the asymptotic minimax separation distance with exact constant for known variance when the design satisfies very specific conditions. Theorem 4.2 provides the non-asymptotic counterpart of their result, but the constants in (4.4) and (4.5) are not optimal.

4.2. Unknown variance

4.2.1. Preliminaries

We now turn to the study of the minimax separation distances when the variance σ^2 is unknown. In Section 3.2, we have introduced the notions of δ-separation distances and (α, δ)-minimax separation distances when the variance σ^2. We now define their counterpart for an unknown variance σ^2.

Let us consider a test Φ_α of the hypothesis $\mathbf{H_0}$ for the linear regression model with fixed design $\mathbf{X}$. We say that Φ_α has a level α under unknown variance if

$$\sup_{\sigma > 0} \mathbb{P}_{0_p, \sigma}[\Phi_\alpha(\mathbf{Y}, \mathbf{X}) > 0] \leq \alpha \ .$$

This means that the type I error probability is controlled uniformly over all variance σ^2. Similarly, we want to control the type II error probabilities uniformly over all variances. The δ-separation distance $\rho_{F,U}[\Phi_\alpha, k, \mathbf{X}]$ of Φ_α over $\Theta[k, p]$ for unknown variance is defined by

$$\rho_{F,U}[\Phi_\alpha, k, \mathbf{X}] := \inf \left\{ \rho > 0, \inf_{\substack{\sigma > 0,\ \theta_0 \in \Theta[k,p], \\ \|\mathbf{X}\theta_0\|_n \geq \sqrt{n}\rho\sigma}} \mathbb{P}_{\theta_0, \sigma}[\Phi_\alpha = 1] \geq 1 - \delta \right\} . \tag{4.6}$$

Hence, $\rho_{F,U}[\Phi_\alpha, k, \mathbf{X}]$ corresponds to the minimal distance such that the hypotheses $\{\theta_0 = 0_p \text{ and } \sigma > 0\}$ and $\{\theta_0 \in \Theta[k,p] \text{ and } \sigma > 0\ ,\ \|\mathbf{X}\theta_0\|_n^2 \geq n\rho^2_{F,U}[\Phi_\alpha, k, \mathbf{X}]\sigma^2\}$ are well separated by the test Φ_α. Taking the infimum over all level α tests, we get the (α, δ) minimax separation distance over $\Theta[k,p]$ with design $\mathbf{X}$ and unknown variance is

$$\rho^*_{F,U}[k, \mathbf{X}] := \inf_{\Phi_\alpha} \rho_{F,U}[\Phi_\alpha, k, \mathbf{X}]\ . \tag{4.7}$$

Finally, $\rho^*_{F,U}[k] := \sup_{\mathbf{X}} \rho^*_{F,U}[k, \mathbf{X}]$ corresponds to the (α, δ)-minimax separation distance over $\Theta[k,p]$ with the "worst-case designs".

In the Gaussian design, we define $\rho_{R,U}[\Phi_\alpha, k, \Sigma]$, $\rho^*_{R,U}[k, \Sigma]$, and $\rho^*_{R,U}[k]$ analogously to (4.6) and (4.7) by replacing the norm $\|\mathbf{X}\theta_0\|_n/\sqrt{n}$ by $\|\sqrt{\Sigma}\theta_0\|_p$.

4.2.2. Gaussian design

Minimax bounds have been proved in [44] in the non ultra-high dimensional setting. The next theorem encompasses high dimensional and ultra-high dimensional settings.

Theorem 4.3. *Suppose that $\alpha + \delta \leq 53\%$ and that $p \geq n \geq 8\log(2/\delta)$. For any $1 \leq k \leq \lfloor p^{1/3} \rfloor$, the (α, δ)-minimax separation distance over $\Theta[k,p]$ with covariance I_p and unknown variance satisfies*

$$(\rho^*_{R,U}[k, I_p])^2 \geq C_1 \frac{k}{n} \log(p) \exp\left[C_2 \frac{k}{n} \log(p)\right]\ . \tag{4.8}$$

For any $1 \leq k \leq n/2$ and any covariance Σ, we have

$$(\rho^*_{R,U})^2[k, \Sigma] \leq C_1(\alpha, \delta) \frac{k}{n} \log\left(\frac{ep}{k}\right) \exp\left[C_2(\alpha, \delta) \frac{k}{n} \log\left(\frac{ep}{k}\right)\right]\ . \tag{4.9}$$

Furthermore, this upper bound is simultaneously achieved for all k and Σ by a procedure T_α (defined in Section 10.1.2).

Remark 4.4 (Minimax adaptation)**.** It follows from Theorem 4.3 that, under unknown variance, adaptation to the sparsity is possible and that the minimax separation distance $(\rho^*_{RU}[k])^2$ over $\Theta[k,p]$ is of order

$$C_1(\alpha, \delta) \frac{k}{n} \log(p) \exp\left[C_2(\alpha, \delta) \frac{k}{n} \log(p)\right]\ . \tag{4.10}$$

Remark 4.5. The condition $k \leq p^{1/3}$ can be replaced by $k \leq p^{1/2-\gamma}$ with $\gamma > 0$, the only difference being that the constants involved in (4.8) would depend on γ. These conditions are not really restrictive for a sparse high-dimensional regression since the usual setting is $k \leq n \ll p$.

Note $k \leq p^3$ implies that $\log(p) \leq 3/2 \log(p/k) \leq 3\log(p/k^2)$ so that we cannot distinguish terms $C_1 \log(p)$ from $C_2 \log(p/k^2)$ or $C_3 \log(p/k)$. As a consequence (4.10) does not necessarily capture the right dependency on k in the logarithmic terms. This observation also holds for all the next results that require $k \leq p^{1/3}$.

Remark 4.6 (Dependent design)**.** As for the known variance case, we have $\rho^*_{R,U}[k, I_p] \geq C(\alpha,\delta)\rho^*_{R,U}[k]$, that is the testing problem is more complex for an independent design than for a correlated design. For some covariance matrices Σ, the minimax separation distance with covariance Σ is much smaller than $\rho^*_{R,U}[k, I_p]$. Verzelen and Villers [44] provide such an example of a matrix Σ in (see Propositions 8 and 9). However, the arguments used in the proof of their example are not generalizable to other covariances. In fact, the computation of sharp minimax bounds that capture the dependency of $\rho^*_{R,U}[k,\Sigma]$ on Σ remains an open problem.

4.2.3. Fixed design

Ingster et al. [34] derive the asymptotic minimax separation distance for some specific design when $k\log(p)/n$ goes to 0. Here, we provide the non asymptotic counterpart that encompass all the regimes.

Proposition 4.4. *Assume that $\alpha + \delta \leq 26\%$ and that $p \geq n \geq C(\alpha,\delta)$. For any $1 \leq k \leq \lfloor p^{1/3} \rfloor$, there exist some $n \times p$ designs $\mathbf{X}$ such that*

$$(\rho^*_{F,U}[k,\mathbf{X}])^2 \geq C_1 \frac{k}{n}\log(p)\exp\left[C_2\frac{k}{n}\log(p)\right] \ . \tag{4.11}$$

For any $1 \leq k \leq n/2$ and any $n \times p$ design $\mathbf{X}$, we have

$$(\rho^*_{F,U}[k,\mathbf{X}])^2 \leq C_1(\alpha,\delta)\frac{k}{n}\log\left(\frac{ep}{k}\right)\exp\left[C_2(\alpha,\delta)\frac{k}{n}\log\left(\frac{ep}{k}\right)\right] \ . \tag{4.12}$$

Furthermore, this upper bound is simultaneously achieved for all k and $\mathbf{X}$ by a procedure T_α (defined in Section 10.1.2).

Again, we observe a phenomenon analogous to the random design case.

4.3. Comparison between known and unknown variance

There are three regimes depending on (k,p,n). They are depicted on Figure 2:

1. $\mathbf{k\log(p)} \leq \sqrt{\mathbf{n}}$. The minimax separation distances are of the same order for known and unknown σ^2. The minimax distance $k\log(p)/n$ is also of the same order as the minimax risk of prediction.
2. $\sqrt{\boldsymbol{n}} \leq \mathbf{k\log(p)} \leq \mathbf{n}$. If σ^2 is known, the minimax separation distance is always of order $1/\sqrt{n}$. In such a case, an optimal procedure amounts to test the hypothesis $\{\mathbb{E}[\|\mathbf{Y}\|_n^2] = n\sigma^2\}$ against $\{\mathbb{E}[\|\mathbf{Y}\|_n^2] > n\sigma^2\}$ using the statistic $\|\mathbf{Y}\|_n^2/\sigma^2$. If σ^2 is unknown, the statistic $\|\mathbf{Y}\|_n^2/\sigma^2$ is not available and the minimax separation distance behaves like $k\log(p)/n$.
3. $\mathbf{k\log(p)} \geq \mathbf{n}$. If σ^2 is unknown, the minimax separation distance blows up. It is of order $(p/k)^{Ck/n}$. Consequently, the problem of testing $\{\theta_0 = 0_p\}$ becomes extremely difficult in this setting.

5. Prediction

In contrast to the testing problem, the minimax risks of prediction ($\mathbf{P_2}$) exhibit really different behaviors in fixed and in random design. The big picture is summarized in Figure 1. We recall that the minimax risks $\mathcal{R}_F[k,\mathbf{X}]$, $\mathcal{R}_F[k]$, $\mathcal{R}_R[k,\Sigma]$, and $\mathcal{R}_R[k]$ are defined in Section 3.1.

5.1. Gaussian design

Proposition 5.1 (Minimax lower bound for prediction)**.** *Assume that $p\geq C$. For any $1\leq k\leq \lfloor p^{1/3}\rfloor$, we have*

$$\mathcal{R}_R[k,I_p]\geq C\frac{k}{n}\log\left(\frac{ep}{k}\right)\exp\left\{C_2\frac{k}{n}\log\left(\frac{ep}{k}\right)\right\}\ . \tag{5.1}$$

Remark 5.1 (General covariances Σ)**.** The lower bound (5.1) is only stated for the identity covariance $\Sigma=I_p$. For general covariance matrices Σ, we have

$$\mathcal{R}_R[k,\Sigma]\geq C\frac{\Phi_{2k,-}(\sqrt{\Sigma})}{\Phi_{2k,+}(\sqrt{\Sigma})}\times\frac{k}{n}\log\left(\frac{ep}{k}\right)\ , \tag{5.2}$$

for any $k\leq n\leq p/2$. This statement has been proved in [42] (Proposition 4.5) in the special case of restricted isometry, but the proof straightforwardly extends to restricted eigenvalue conditions. For $\Sigma=I_p$, the lower bound (5.2) does not capture the elbow effect in an ultra-high dimensional setting (compare with (5.1)).

Theorem 5.2 (Minimax upper bound)**.** *Assume that $n\geq C$. There exists an estimator $\widetilde{\theta}^V$ (defined in Section 10.2.1) such that the following holds:*

1. *The computation of $\widetilde{\theta}^V$ does not require the knowledge of σ^2 or k.*
2. *For any covariance Σ, any $\sigma>0$, any $1\leq k\leq\lfloor(n-1)/4\rfloor$, and any $\theta_0\in\Theta[k,p]$ we have*

$$\mathbb{E}_{\theta_0,\sigma}\left[\|\sqrt{\Sigma}(\widetilde{\theta}^V-\theta_0)\|_p^2\right]\leq C_1\frac{k}{n}\log\left(\frac{ep}{k}\right)\exp\left\{C_2\frac{k}{n}\log\left(\frac{ep}{k}\right)\right\}\sigma^2\ . \tag{5.3}$$

In contrast to similar results such as Theorem 1 in Giraud [27] or Theorem 3.4 in Verzelen [42], we do not restrict k to be smaller than $n/(2\log p)$, that is we encompass both high-dimensional and ultra-high dimensional setting. The proof of the theorem is based on a new deviation inequality for the spectrum of Wishart matrices stated in Lemma 11.2.

Remark 5.2 (Minimax risk)**.** We derive from Theorem 5.2 and Proposition 5.1 that the minimax risk $\mathcal{R}_R[k]$ is of order

$$C_1\frac{k}{n}\log\left(\frac{ep}{k}\right)\exp\left\{C_2\frac{k}{n}\log\left(\frac{ep}{k}\right)\right\}\ .$$

If $k\log(p/k)$ is small compared to n, the minimax risk of estimation is of order $Ck\log(p/k)/n$. In an ultra-high dimensional setting, we again observe a blow up.

Remark 5.3 (Adaptation to sparsity and the variance)**.** The estimator $\widetilde{\theta}^V$ does not requires the knowledge of k and of the variance $\sigma^2 = \mathrm{Var}(Y|X)$. It follows that $\widetilde{\theta}^V$ is minimax adaptive to all $1 \leq k \leq p^{1/3} \wedge [(n-1)/4]$ and to all $\sigma^2 > 0$. As a consequence, adaptation to the sparsity and to the variance is possible for this problem.

Remark 5.4 (Dependent design)**.** The risk upper bound of $\widetilde{\theta}^V$ stated in Theorem 5.2 is valid for any covariance matrix Σ of the covariance X. In contrast, the minimax lower bound of Theorem 4.3 is restricted to the identity covariance. This implies that the minimax prediction risk for a general matrix Σ is at worst of the same order as in the independent case: there exists a universal constant $C > 0$ such that for all covariance Σ,

$$\mathcal{R}_R[k, I_p] \geq C\mathcal{R}_R[k] \ .$$

In Remark 5.1, we have stated a minimax lower bound for prediction that depends on the restricted eigenvalues of Σ. Fix some $0 < \gamma < 1$. If we consider some covariance matrices Σ such that $\Phi_{2k,-}(\sqrt{\Sigma})/\Phi_{2k,+}(\sqrt{\Sigma}) \geq 1-\gamma$, the minimax lower bound (5.2) and the upper bound (5.3) match up to a constant $C(\gamma)$. In general, the lower bound (5.2) and the upper bound (5.3) do not exhibit the same dependency with respect to Σ, especially when $\Phi_{2k,-}(\sqrt{\Sigma})/\Phi_{2k,+}(\sqrt{\Sigma})$ is close to zero.

5.2. *Fixed design*

5.2.1. *Known variance*

The minimax prediction risk with known variance has been studied in Raskutti et al. [39] and Rigollet and Tsybakov [40] (see also [1, 47]). For any design $\mathbf{X}$ and any $1 \leq k \leq n$, these authors have proved that the minimax risk $\mathcal{R}_F[k, \mathbf{X}]$ satisfies

$$C_1 \inf_{s\leq k} \frac{\Phi_{2s,-}(\mathbf{X})}{\Phi_{2s,+}(\mathbf{X})} \frac{s}{n} \log\left(\frac{ep}{s}\right) \leq \mathcal{R}_F[k, \mathbf{X}] \leq C_2 \frac{k}{n} \log\left(\frac{ep}{k}\right) \ . \tag{5.4}$$

Next, we bound the supremum $\sup_{\mathbf{X}} R_F[k, \mathbf{X}]$ and we study the possibility of adaptation to the sparsity.

Proposition 5.3. *For any $1 \leq k \leq n$, the supremum $\sup_{\mathbf{X}} \mathcal{R}_F[k, \mathbf{X}]$ is lower bounded as follows*

$$\mathcal{R}_F[k] \geq C\left[\frac{k}{n}\log\left(\frac{ep}{k}\right) \wedge 1\right] \ . \tag{5.5}$$

Assume that $p \geq n$. There exists an estimator $\tilde{\theta}^{BM}$ (defined in Section 10.2.2) which satisfies

$$\sup_{\mathbf{X}} \sup_{\theta_0 \in \Theta[k,p]} \mathbb{E}_{\theta_0,\sigma}\left[\|\mathbf{X}(\widehat{\theta}^{BM} - \theta_0)\|_n^2\right]/(n\sigma^2) \leq C\left[\frac{k}{n}\log\left(\frac{ep}{k}\right) \wedge 1\right] \ , \qquad (5.6)$$

for any $1 \leq k \leq n$.

This upper bound (5.6) is a consequence of Birgé and Massart [11].

Remark 5.5. If $k\log(p/k)$ is small compared to n, the minimax risk is of order $Ck\log(p/k)/n$. In an ultra-high dimensional setting, this minimax risk remains close to one. This corresponds (up to renormalization) to the minimax risk of estimation of the vector $\mathbb{E}[\mathbf{Y}]$ of size n . As a consequence, the sparsity assumption does not play anymore a role in a ultra-high dimensional setting. From (5.6), we derive that adaptation to the sparsity is possible when the variance σ^2 is known.

Remark 5.6 (Dependency of $\mathcal{R}_F[k,\mathbf{X}]$ on $\mathbf{X}$)**.** For designs $\mathbf{X}$, such that the ratio $\Phi_{2k,-}(\mathbf{X})/\Phi_{2k,+}(\mathbf{X})$ is close to one, the lower bounds and upper bounds of (5.4) agree with each other. This is for instance the case of the realizations (with high probability) of a Gaussian standard independent design (see the proof of Proposition 5.3 for more details).

However, the dependency of the minimax lower bound in (5.4) on $\mathbf{X}$ is not sharp when the ratio $\Phi_{2k,-}(\mathbf{X})/\Phi_{2k,+}(\mathbf{X})$ is away from one. Take for instance an orthogonal design with $p = n$ and duplicate the last column. Then, the lower bound (5.4) for this new design $\mathbf{X}$ is 0 while the minimax risk is of order $k\log(p/k)/n$.

Similarly, the dependency of the minimax upper bound in (5.4) on $\mathbf{X}$ is not sharp. For very specific design, it is possible to obtain a minimax risk $\mathcal{R}_F[k,\mathbf{X}]$ that is much smaller than $k/n\log(p/k)\wedge 1$ (see Abramovich and Grinshtein [1]).

Remark 5.7 (Comparison with l_1 procedures)**.** The designs $\mathbf{X}$ for which l_1 procedures such as the Lasso or the Dantzig selector are proved to perform well require that $\Phi_{2k,-}(\mathbf{X})/\Phi_{2k,+}(\mathbf{X})$ is close to one. It is interesting to notice that these designs $\mathbf{X}$ precisely correspond to situations where the minimax risk is close to its maximum $k\log(p/k)/n$ (see Equation (5.4)). We refer to [39] for a more complete discussion.

Remark 5.8. We easily retrieve from (5.4) a result of asymptotic geometry first observed by Baraniuk et al. [4] in the special of restricted isometry property [14]. For any $0 < \delta \leq 1$, there exists a constant $C(\delta) > 0$ such that no $n \times p$ matrix $\mathbf{X}$ can fulfill $\Phi_{k,-}(\mathbf{X})/\Phi_{k,+}(\mathbf{X}) \geq \delta$ if $k(1+\log(p/k)) \geq C(\delta)n$.

Proof. If $\Phi_{2k,-}(\mathbf{X})/\Phi_{2k,+}(\mathbf{X}) \geq \delta$, then $\mathcal{R}_F[k,\mathbf{X}] \geq C\delta k\log(ep/k)/n$.

We also have $\mathcal{R}_F[k,\mathbf{X}] \leq \mathcal{R}_F[p,\mathbf{X}] \leq 1$. The last inequality follows from the risk of an estimator $\widehat{\theta}_n \in \arg\min_{\theta\in\mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\theta\|_n^2$. Gathering these two bounds allows to conclude. □

5.2.2. Unknown variance

We now consider the problem of prediction when the variance σ^2 is unknown.

Proposition 5.4. *For any $1 \leq k \leq n$, there exists an estimator $\widehat{\theta}^{(k)}$ that does not require the knowledge of σ^2 such that*

$$\sup_{\mathbf{X}} \sup_{\sigma>0} \sup_{\theta_0 \in \Theta[k,p]} \mathbb{E}_{\theta_0,\sigma}\left[\frac{\|\mathbf{X}(\widehat{\theta}^{(k)}-\theta_0)\|_n^2}{n\sigma^2}\right] \leq C\left[\frac{k}{n}\log\left(\frac{ep}{k}\right) \wedge 1\right] \ . \tag{5.7}$$

Thus, the optimal risk of prediction over $\Theta[k,p]$ remains of the same order for known and unknown σ^2.

Let us now study to what extent adaptation to the sparsity is possible when the variance σ^2 is unknown. In order to get some ideas let us provide risk bounds for two procedures that do not require the knowledge of σ: the estimator $\widetilde{\theta}^V$ already studied for Gaussian design (defined in Section 10.2.1) and the estimator $\widehat{\theta}_n$ defined by $\widehat{\theta}_n \in \arg\min_{\theta\in\mathbb{R}^p} \|\mathbf{Y}-\mathbf{X}\theta\|_n^2$.

Proposition 5.5 (Risk bound for $\widetilde{\theta}^V$ and $\widehat{\theta}_n$)**.** *Assume that $n \geq 14$. For any $1 \leq k \leq \lfloor (n-1)/4 \rfloor$, the maximal risk of $\widehat{\theta}^V$ over $\Theta[k,p]$ is upper bounded as follows*

$$\sup_{\mathbf{X}} \sup_{\sigma>0,\theta_0\in\Theta[k,p]} \mathbb{E}_{\theta_0,\sigma}\left[\frac{\|\mathbf{X}(\tilde{\theta}^V-\theta_0)\|_n^2}{n\sigma^2}\right] \leq C_1\frac{k}{n}\log\left(\frac{ep}{k}\right)\exp\left[C_2\frac{k}{n}\log\left(\frac{ep}{k}\right)\right]\sigma^2. \tag{5.8}$$

For any $1 \leq k \leq n$, the maximal risk of $\widehat{\theta}_n$ over $\Theta[k,p]$ is upper bounded as follows

$$\sup_{\mathbf{X}} \sup_{\sigma>0} \sup_{\theta_0\in\Theta[k,p]} \mathbb{E}_{\theta_0,\sigma}\left[\|\mathbf{X}(\widehat{\theta}_n-\theta_0)\|_n^2\right]/(n\sigma^2) \leq 1 \ . \tag{5.9}$$

The risk bound (5.8) is also satisfied by the procedure of Baraud et al. [6]. The proof of (5.8) is a consequence of one of their results.

Remark 5.9. As a consequence, $\tilde{\theta}^V$ simultaneously achieves the minimax risk over all $\Theta[k,p]$ for all $k \leq \lfloor (n-1)/4 \rfloor$ such that $k(1+\log(p)/k) \leq n$. In an ultra-high dimensional setting, the maximum risk of $\tilde{\theta}^V$ over $\Theta[k,p]$ is controlled by $(ep/k)^{Ck/n}$ while the minimax risk is smaller than 1. If the upper bound (5.8) is sharp then this would imply that $\tilde{\theta}^V$ is not adaptive to the sparsity in an ultra-high dimensional setting.

In contrast, $\widehat{\theta}_n$ is minimax adaptive over all $\Theta[k,p]$ such that $k(1+\log(p)/k) \geq n$, but its behavior is suboptimal in a non-ultra-high dimensional setting.

In order to get an estimator that is adaptive to all indexes k, we would need to merge the properties of $\widetilde{\theta}^V$ (for non-ultra-high dimensional cases) and of $\widehat{\theta}_n$ (for ultra-high dimensional cases). The following proposition tells us that it is in fact impossible.

Proposition 5.6 (Adaptation to the sparsity is impossible under unknown variance)**.** *Consider any* $p \geq n \geq C_1$ *and* $1 \leq k \leq \lfloor p^{1/3} \rfloor$ *such that* $k \log(ep/k) \geq C_2 n$*. There exists a design* $\mathbf{X}$ *of size* $n \times p$ *such that for any estimator* $\widehat{\theta}$*, we have either*

$$\sup_{\sigma>0} \mathbb{E}_{0_p,\sigma}\left[\|\mathbf{X}(\widehat{\theta}-0_p)\|_n^2/(n\sigma^2)\right] > C \ ,$$

or

$$\sup_{\sigma>0} \sup_{\theta_0 \in \Theta[k,p]} \mathbb{E}_{\theta_0,\sigma}\left[\|\mathbf{X}(\widehat{\theta}-\theta_0)\|_n^2/(n\sigma^2)\right] > \exp\left[C\frac{k}{n}\log\left(\frac{ep}{k}\right)\right] \ .$$

As a benchmark, we recall the minimax upper bounds:

$$\mathcal{R}_F[1] \leq C_1 \frac{\log(p)}{n} \qquad \textit{and} \qquad \mathcal{R}_F[k] \leq C_2 \left[\frac{k}{n}\log\left(\frac{ep}{k}\right) \wedge 1\right] \ .$$

The proof of proposition 5.6 is based on the minimax lower bounds (4.11) for the testing problem ($\mathbf{P_1}$) under unknown variance. The proof uses designs $\mathbf{X}$ that are realizations of standard Gaussian designs.

Remark 5.10. In the setup of Proposition 5.6, any estimator $\widehat{\theta}$ that does not require the knowledge of k and σ^2 has to pay at least one of these two prices:

1. The estimator $\widehat{\theta}$ does not use the sparsity of the true parameter θ_0. Its risk for estimating 0_p is of the same order as the minimax risk over $\mathbb{R}^p$. The estimator $\widehat{\theta}_n$ has this drawback.
2. For any $1 \leq k \leq p^{1/3}$, we have

$$\sup_{\mathbf{X}} \sup_{\sigma>0, \theta_0 \in \Theta[k,p]} \mathbb{E}_{\theta_0,\sigma}\left[\frac{\|\mathbf{X}(\widehat{\theta}-\theta_0)\|_n^2}{n\sigma^2}\right] \geq C_1 \frac{k}{n}\log\left(\frac{ep}{k}\right) \exp\left[C_2\frac{k}{n}\log\left(\frac{ep}{k}\right)\right].$$

   This is the price for adaptation when σ^2 is unknown. The estimator $\widetilde{\theta}^V$ exhibits this behavior.

As a conclusion, it is impossible to merge the qualities of $\widetilde{\theta}^V$ and of $\widehat{\theta}_n$.

The best prediction risk that can be achieved by a procedure that aim to adaptation to the sparsity is of order

$$\frac{k}{n}\log\left(\frac{p}{k}\right)\exp\left[C\frac{k}{n}\log(p/k)\right] \ .$$

In other words, the unavoidable loss for adaptation for unknown variance is a factor $\exp[Ck/n\log(p/k)]$ In this sense, the estimator $\widetilde{\theta}^V$ (and as a byproduct the procedure of Baraud et al. [6]) achieves the optimal prediction risk under unknown variance and unknown sparsity.

In conclusion, the minimax risks of prediction are of the same order for fixed and Gaussian design and for known and unknown variance when $k\log(p/k)$ is small compared to n. In an ultra-high dimensional setting, the minimax risks behave differently. For Gaussian design, the minimax risk is of the order $(p/k)^{Ck/n}$. In contrast, the minimax risk of prediction remains smaller than one for fixed design regression with known variance. When the sparsity and the variance are unknown, there is a price to pay for adaptation under fixed design. All these behaviors are depicted on Figure 1.

6. Inverse problem and support estimation

6.1. Minimax risk of estimation

We recall that the minimax risks of estimation for the inverse problem $\mathcal{RI}_F[k,\mathbf{X}]$, $\mathcal{RI}_F[k]$, $\mathcal{RI}_R[k,\Sigma]$, and $\mathcal{RI}_R[k]$ have been defined in Section 3.3.

6.1.1. Fixed design

First, we consider the problem $(\mathbf{P_3})$ for a fixed design regression model. The minimax risk of estimation over $\Theta[k,p]$ with a design $\mathbf{X}$ is noted $\mathcal{RI}_F[k,\mathbf{X}]$ and is defined in (3.8). Raskutti et al. [39] have recently provided the following bounds

$$C_1\left[\frac{k\log(ep/k)}{\Phi_{2k\wedge p,+}(\mathbf{X})}\right]\leq \mathcal{RI}_F[k,\mathbf{X}]\leq C_2\frac{k\log(ep/k)}{\Phi_{2k\wedge p,-}(\mathbf{X})}\ , \tag{6.1}$$

that holds for any fixed design $\mathbf{X}$ and any $1\leq k\leq n$. The lower and upper bounds match up to the factor $\Phi_{2k\wedge p,+}(\mathbf{X})/\Phi_{2k\wedge p,-}(\mathbf{X})$. The upper bound is achieved by least-squares estimator over $\Theta[k,p]$ [39]. If the restricted eigenvalues of $\mathbf{X}$ are close to one, then the minimax risk is of order $k\log(ep/k)$. Next, we improve the lower bound in (6.1) in order to grasp the behavior of the minimax risk for non orthogonal design.

Proposition 6.1. *For any design $\mathbf{X}$ and any $1\leq k\leq n$, we have*

$$\mathcal{RI}_F[k,\mathbf{X}]\geq C\left[\frac{1}{\Phi_{2k\wedge p,-}(\mathbf{X})}\vee\frac{k\log(ep/k)}{\Phi_{1,+}(\mathbf{X})}\right]\ . \tag{6.2}$$

In order to interpret these bounds let us restrict ourselves to design $\mathbf{X}$ such that each column has $\sqrt{n}$ norm, as justified in Section 3.3. The collection of such designs is noted $\mathcal{D}_{n,p}$. Observe that $\mathbf{X}\in\mathcal{D}_{n,p}$ enforces $\Phi_{1,+}(\mathbf{X})=n$.

In the sequel, we are interested in the smallest minimax risk $\mathcal{RI}_F[k,\mathbf{X}]$ that is achievable if we can choose the $n\times p$ design $\mathbf{X}\in\mathcal{D}_{n,p}$, that is we want to bound $\mathcal{RI}_F[k]=\inf_{\mathbf{X}\in\mathcal{D}_{n,p}}\mathcal{RI}_F[k,\mathbf{X}]$. The minimax risk $\mathcal{RI}_F[k]$ tells us the intrinsic difficulty of estimating a k sparse vector of size p with n observations.

Proposition 6.2.

1. *Assume that $k[1+\log(p/k)]\leq Cn$. Then, we have*

$$C_1\frac{k}{n}\log\left(\frac{ep}{k}\right)\leq\mathcal{RI}_F[k]\leq C_2\frac{k}{n}\log\left(\frac{ep}{k}\right)\ . \tag{6.3}$$

This bound is for instance achieved for designs $\mathbf{X}$ that are realizations (with a high probability) of normalized standard Gaussian design.

2. *For any design* $\mathbf{X} \in \mathcal{D}_{n,p}$ *and any* $k \leq n \wedge p/2$*, we have*

$$\Phi_{2k,-}(\mathbf{X}) \leq C_1 n \left(\frac{k}{ep}\right)^{C_2 k/n} . \tag{6.4}$$

3. *For any* $k \leq n/4 \wedge p/2$*, we have*

$$C_1 \left[\frac{k}{n} \log\left(\frac{ep}{k}\right) \vee \frac{1}{n} \exp\left\{C_4 \frac{k}{n} \log\left(\frac{p}{k}\right)\right\}\right] \leq \mathcal{RI}_F[k] \tag{6.5}$$
$$\mathcal{RI}_F[k] \leq C_2 \frac{k}{n} \log\left(\frac{p}{k}\right) \exp\left[C_3 \frac{k}{n} \log\left(\frac{p}{k}\right)\right] .$$

Remark 6.1. The bound (6.3) tells us that the best minimax risk that is achievable in a non-ultra-high dimensional setting is of order $k \log(ep/k)/n$. The Lasso achieves the (almost optimal) risk bound $k \log(p)/n$ under some assumptions on the design matrix.

Remark 6.2. The lower bound (6.4) is of geometric nature. Combined with (6.2), it implies the lower bound of (6.5). In an ultra-high dimensional setting, it is not possible to build a design $\mathbf{X}$ such that $\Phi_{2k,+}(\mathbf{X})/\Phi_{2k,-}(\mathbf{X})$ is close to one (see Remark 5.8). In fact, the quantity $\Phi^{-1}_{2k,-}(\mathbf{X})$ blows up because of geometric constrains. When $k[1+\log(p/k)]$ is larger compared to $n\log(n)$, both bounds in (6.5) are comparable and the minimax risk is of order $\exp[Ck/n \log(p/k)]$. As a consequence, the inverse problem becomes extremely difficult in an ultra-high dimensional setting.

Remark 6.3. While the quantity $k\log(p/k)$ in (6.3) is due to the "size" of the parameter space $\Theta[k,p]$, the exponential term of the minimax risk in ultra-high dimension is essentially driven by geometrical constrains on the design $\mathbf{X}$.

Proposition 6.3 (Adaptation to the sparsity and the variance)**.** *As in the prediction case, we consider the estimator* $\widetilde{\theta}^V$ *(defined in Section 10.2.1). Assume that* $p \geq 2n$*. For any design* $\mathbf{X}$*, any* $\sigma > 0$*, any* $1 \leq k \leq \lfloor (n-1)/4 \rfloor$*, and any* $\theta_0 \in \Theta[k,p]$*, we have*

$$\frac{\|\widetilde{\theta}^V - \theta_0\|_p^2}{\sigma^2} \leq C_1 \frac{k}{\Phi_{3k,-}(\mathbf{X})} \log\left(\frac{ep}{k}\right) \exp\left[C_2 \frac{k}{n} \log\left(\frac{ep}{k}\right)\right] , \tag{6.6}$$

with probability larger than $1 - e^{-n} - C/p$.

Remark 6.4. Although the bound (6.6) is in probability and not in expectation, it suggests that adaptation to the sparsity and to the variance are possible.

6.1.2. Random design

Let us turn to the Gaussian design case. We are interested in bounding $\mathcal{RI}_R[k,\Sigma]$ and $\mathcal{RI}_R[k]$ as defined in (3.9).

Proposition 6.4. *For any $1 \leq k \leq (n-1)/4$ and any covariance Σ, we have*

$$C_1\left[\frac{1}{n\Phi_{2k\wedge p,-}(\sqrt{\Sigma})}\vee\frac{k\log(ep/k)}{n\Phi_{1,+}(\sqrt{\Sigma})}\right]\leq\mathcal{RI}_R[k,\Sigma] \tag{6.7}$$

$$\mathcal{RI}_R[k,\Sigma]\leq C_2\frac{k\log\left(ep/k\right)}{n\Phi_{2k\wedge p,-}(\sqrt{\Sigma})}\exp\left[C_3\frac{k}{n}\log\left(\frac{ep}{k}\right)\right]\ .$$

As long as $k[1\log(p/k)] \leq n$, we derive that $\mathcal{RI}_R[k] := \inf_{\Sigma\in\mathcal{S}_p}\mathcal{RI}_R[k,\Sigma]$ satisfies

$$C_1\frac{k}{n}\log\left(\frac{ep}{k}\right)\leq\mathcal{RI}_R[k]\leq C_2\frac{k}{n}\log\left(\frac{ep}{k}\right)\ . \tag{6.8}$$

We observe that $\mathcal{RI}_R[k]$ and $\mathcal{RI}_F[k]$ behave similarly in a non-ultra-high dimensional setting.

Remark 6.5 (Ultra-high dimensional case)**.** Proposition 6.4 does not allow to derive the order of magnitude of $\mathcal{RI}_R[k]$ in an ultra-high dimensional setting. While the upper bound in (6.7) is blowing up, the lower bound remains as small as $k\log(p/k)/n$. Nevertheless, we know from Proposition 5.1 that

$$\mathcal{RI}_R[k,I_p]=\mathcal{R}_R[k,I_p]\geq C_1\frac{k\log\left(ep/k\right)}{n}\exp\left[C_2\frac{k}{n}\log\left(\frac{ep}{k}\right)\right]\ .$$

This suggests that $\mathcal{RI}_R[k]$ is blowing up in an ultra-high dimensional setting but the problem remains open.

In the next proposition, we state the counterpart of Proposition 6.3 in the random design case.

Proposition 6.5 (Adaptation to the sparsity and the variance)**.** *As in the prediction case, we consider the estimator $\widetilde{\theta}^V$ (defined in Section 10.2.1). Assume that $p \geq 2n$. For any covariance Σ, any $\sigma > 0$, any $1 \leq k \leq \lfloor (n-1)/12\rfloor$, and any $\theta_0 \in \Theta[k,p]$, we have*

$$\frac{\|\widetilde{\theta}^V-\theta_0\|_p^2}{\sigma^2}\leq C_1\frac{k}{n\Phi_{3k,-}(\sqrt{\Sigma})}\log\left(\frac{ep}{k}\right)\exp\left[C_2\frac{k}{n}\log\left(\frac{ep}{k}\right)\right]\ , \tag{6.9}$$

with probability larger than $1-e^{-n}-C/p$.

6.2. Consequences on support estimation

We deduce from the minimax lower bounds for the inverse problem ($\mathbf{P_3}$) some consequences for the support estimation problem ($\mathbf{P_4}$) in a ultra-high dimensional setting. The case $k[1+\log(p/k)]$ small compared to n has been studied in Wainwright [45].

Definition 6.1. For any $\rho > 0$ and any $k \leq p$, the set $\mathcal{C}_k^p(\rho)$ is made of all vectors θ in $\Theta[k,p]$ such that θ contains exactly k non-zero coefficients that are all equal to $\rho/\sqrt{k}$.

In a non-ultra high dimensional setting, Wainwright [46] has proved, that under suitable conditions on a design $\mathbf{X} \in \mathcal{D}_{n,p}$, it is possible to recover the support of any vector θ_0 that belong to $\mathcal{C}_k^p(\rho)$ with ρ of order of $\sqrt{k\log(p)/n}\sigma$. Here, we prove that ρ has to be much larger in an ultra-high dimensional setting.

Proposition 6.6 (Support recovery is almost impossible)**.** *For any* $\rho^2 \leq C_1/n\left(\frac{ep}{k}\right)^{C_2k/n}$ *and any* $k \leq n \wedge p/2$*, we have*

$$\inf_{\mathbf{X}\in\mathcal{D}_{n,p}} \inf_{\hat{m}} \sup_{\theta_0\in\mathcal{C}_k^p(\rho)} \mathbb{P}_{\theta_0,1}\left[\hat{m} \neq \mathrm{supp}(\theta_0)\right] \geq 1/(2e+1) \ .$$

For any design $\mathbf{X} \in \mathcal{D}_{n,p}$ it is not possible to recover the support of θ_0 with high probability, unless θ_0 satisfies:

$$\frac{\|\theta_0\|_p^2}{\sigma^2} \geq C_1/n\left(\frac{p}{k}\right)^{C_2k/n} \quad .$$

This quantity is blowing up in an ultra-high dimensional setting and it can be much larger than the usual $k\log(p)/n$ that can be achieved in a non-ultra high dimensional setting.

As it is almost impossible to estimate the support of θ_0 in an ultra-high dimensional setting, we may aim to an easier objective. Can we choose a subset $\widehat{M}$ of $\{1,\ldots,p\}$ of size $p_0 \leq p$ that contains the support of θ_0 with high probability? This would allow to reduce the dimension of the problem from p to p_0. Dimension reductions techniques are popular for analyzing high dimensional problems. We study here to what extent dimension reduction is a realistic objective: how large should be the non-zero components of θ_0? How small can we choose p_0?

Proposition 6.7. *Consider a Gaussian design regression with* $\Sigma = I_p$ *and* $\sigma^2 = 1$*. We assume that* $p \geq k^3 \vee C$ *and* $n \geq C$*. Set*

$$\rho^2 = C\frac{k}{n}\log\left(\frac{ep}{k}\right)\exp\left[C_2\frac{k}{n}\log\left(\frac{ep}{k}\right)\right] \ .$$

There exists a universal constant $0 < \delta < 1$ *such that for any measurable subset* $\widehat{M}$ *of* $\{1,\ldots,p\}$ *of size* $p_0 \leq p^\delta$*, we have*

$$\sup_{\theta_0\in\mathcal{C}_k^p(\rho)} \mathbb{P}_{\theta_0,1}\left[\mathrm{supp}(\theta_0) \nsubseteq \widehat{M}\right] \geq 1/8 \ . \tag{6.10}$$

In an ultra-high dimensional setting, it is therefore not possible to reduce the dimension of the problem to p^δ unless the square norm of θ_0 is of order $\exp[Ck/n\log(p)]\sigma^2$. In (6.10), the number 1/8 is of no particular significance. It can be replaced by any constant $c \in (0,1)$ if we take an asymptotic point of view ($(k,p,n) \to \infty$).

Remark 6.6. In Proposition 6.7, we have taken the maximal risk points of view. If we put an uniform prior π on $\mathcal{C}_k^p(\rho)$, it is possible to replace (6.10) by

$$\pi\left[\mathbb{P}_{\theta_0,1}\left\{\mathrm{supp}(\theta_0) \nsubseteq \widehat{M}\right\}\right] \geq C \ ,$$

where C is a positive constant.

Remark 6.7. In order to shed light on the problem of dimension reduction, let us consider a simple asymptotic example: $p_n = \exp(n^{\gamma_1})$ and $k_n = n^{1-(\gamma_1 \wedge 1)+\gamma_2}$ with $\gamma_1 > 0$ and $\gamma_2 > 0$. If we assume that $\theta_n \in \Theta[k_n, p_n]$ is such that $\|\theta_n\|_p^2 \leq \exp(Cn^{\gamma_2+(\gamma_1-1)_+})$, then it is not possible to find a subset $\widehat{M}_n$ of size $\exp(\delta n^{\gamma_1})$ that contains the support of θ_n with probability going to one, where δ is defined as in Proposition 6.7. Consequently, we still have to keep at least $\exp(\delta n^{\gamma_1})$ variables after the process of dimension reduction if we do not want to forget relevant variables!

7. What is an ultra-high dimensional problem?

Until now, we have stated that a problem is ultra-high dimensional when $k\log(p/k)$ is large compared to n. It has been proved that in such a setting, estimation of θ_0, support estimation and even dimension reduction become almost impossible. In this section, we numerically illustrate this phase transition phenomenon. This allows us to quantify on specific examples how large should be $k\log(p/k)/n$ for the phase transition to occur.

First simulation setting. Following the example described in the introduction, we consider a Gaussian design linear regression model with $p = 5000$ and $p = 200$, $n = 50$, $\Sigma = I_p$, and $\sigma = 1$. We set the number of non zero components k ranging from 1 to 15. k being fixed, we take θ_0 such that $(\theta_0)_1 = \cdots = (\theta_0)_k = 4\sqrt{\log(p)/n} \approx 1.30$ (resp. 1.65) for $p = 200$ (resp. $p = 5000$) and $(\theta_0)_{k+1} = \cdots = (\theta_0)_p = 0$. As a consequence, we have $\|\theta_0\|^2 = 16k\log(p)/n$. The non-zero coefficients of θ_0 are chosen large enough so that the support of θ_0 is recoverable when the problem is not ultra-high dimensional. Each experiment is repeated $N = 100$ times.

Dimension reduction procedures. We apply the SIS method [25] to reduce the dimension to a set $\widehat{M}^S$ of size $p_0 = 50$. We then compute the Power of the procedure,

$$\text{Power} := \frac{\text{Card}[\widehat{M}^S \cap \{1,\ldots,k\}]}{k} .$$

The power measures whether the dimension reduction has been performed efficiently.

We also compute the regularization path of the Lasso using the LARS [24] algorithm. Before applying the Lasso, each column of $\mathbf{X}$ is normalized. We consider the set $\widehat{M}^L$ made of the p_0 covariates occurring first in the regularization path. We do not argue that SIS and the Lasso are the best methods here. We have chosen them because they are classical and easy to implement.

Results. The results are presented on Figure 4. When k is small, the dimension reduction problem is not ultra-high dimensional and the Lasso and the SIS methods keep all the relevant covariates. For large k, the both methods miss some of the relevant covariates. For $p = 5000$, there is a clear decrease in the

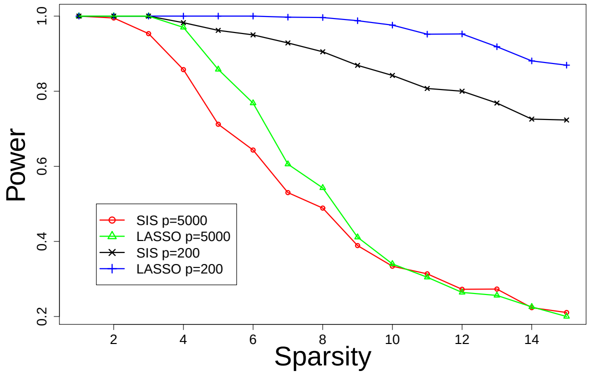

Fig 4. *Power of the dimension reduction procedures (SIS and Lasso) as a function of k.*

power beyond $k = 4$. For $p = 5000$ and $k = 8$, both methods only have a power close to 0.5. In expectation, only four covariates belong to the sets $\widehat{M}^S$ and $\widehat{M}^L$ of size 50. For $p = 200$, there is not a so clear transition, but the power decreases slowly for $k > 8$. If there was no elbow effect in the minimax risk of estimation, then it would still be possible to recover the support of θ_0 with high probability. Indeed, each non-zero component of θ_0 is larger than $4\sqrt{\log(p)/n}$ which is detectable in a reasonable setting (see e.g. [46]). For instance, for $k = 6$ and $p = 5000$, $\|\theta_0\|_p^2/\sigma^2 = 16k\log(p)/n \approx 16.4$. Here, the elbow effect implies that even for a huge signal over noise ratio, it is impossible to reduce the dimension of the problem without forgetting relevant variables.

Second simulation setting. We still take $p = 5000$, $n = 50$, $\Sigma = I_p$, $\sigma = 1$, and k ranging from 1 to 5. k being fixed, we take θ_0 such that $(\theta_0)_1 = \cdots = (\theta_0)_k = u\sqrt{\log(p)/n}$ and $(\theta_0)_{k+1} = \cdots = (\theta_0)_p = 0$. Relying on $N = 100$ experiments, we estimate u_k^* the smallest u such that $\widehat{M}^L$ has a power larger than 0.9. u_k^* corresponds (up to the renormalization $\sqrt{\log(p)/n}$) to the minimal intensity of the signal so that the dimension reduction method does not forget relevant covariates.

Results. The results are presented on Figure 5. For small k, u_k^* remains close to $\sqrt{2}$. In contrast, we observe that u_k^* blows up at $k = 5$. We have not depicted u_6^*, but we have $u_6^* \geq 100$. These two simulation studies confirm that when k becomes large (in comparison to p and n), the dimension reduction problem becomes extremely difficult.

Remark 7.1 (Rule of thumb)**.** From these simulations and from other theoretical arguments (e.g. [27, 22, 45]), we derive a simple rule of thumb. We say that a problem is ultra-high dimensional if

$$\boxed{\frac{k\log(p/k)}{n} \geq 1/2.} \tag{7.1}$$

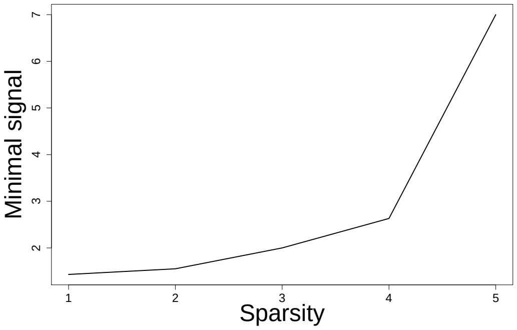

FIG 5. *Minimal signal u_k^* as a function of k.*

For $p = 5000$ and $n = 50$, this corresponds to $k \geq 4$. Setting $p = 200$ and $n = 50$ yields $k \geq 8$. In practice, we do not know k in advance. Nevertheless, this criterion (7.1) helps us to know what is the largest sparsity index such that the statistical problem remains reasonably difficult in the minimax sense.

8. Discussion

As stated in Sections 4–6, the behaviors of the minimax separation distances and of the minimax risks become really different in an ultra-high dimensional setting. Apart from the test problem ($\mathbf{P_1}$) with known variance and the problem of prediction ($\mathbf{P_2}$) with fixed design, all the other separations distances and minimax risks blow up when $k\log(p/k)$ becomes larger than n.

This elbow effect has important practical implications: there is no hope of selecting the relevant covariates in an ultra-high dimensional setting, except if signal over noise ratio is exponentially large. Moreover, even dimension reduction techniques cannot work well in such a setting.

In linear testing ($\mathbf{P_1}$), we have proved that the optimal separation distances highly depend on the knowledge of the variance. Most of the testing procedures in the literature rely on the knowledge of σ^2. Some specific work is therefore needed to derive fast and efficient procedures under unknown variance (but see [34] for a procedure in a specific situation).

We have not discussed so far the problem of variance estimation. From the minimax lower bounds of testing, we deduce the following lower bound.

Proposition 8.1. *Assume that $p \geq n \geq C$. For any $1 \leq k \leq p^{1/3}$, there exist designs $\mathbf{X}$ such that*

$$\inf_{\widehat{\sigma}} \sup_{\sigma>0,\ \theta_0\in\Theta[k,p]} \mathbb{E}_{\theta_0,\sigma}\left[\left|\frac{\widehat{\sigma}^2}{\sigma^2}-\frac{\sigma^2}{\widehat{\sigma}^2}\right|\right] \geq C_1\frac{k}{n}\log\left(\frac{p}{k}\right)\exp\left[C_2\frac{k}{n}\log\left(\frac{p}{k}\right)\right] .$$

As a consequence, the problem of variance estimation becomes extremely difficult in an ultra-high dimensional setting.

In Propositions 5.3 and 6.1, we have provided minimax lower bounds for $(\mathbf{P_2})$ and $(\mathbf{P_3})$ over $\Theta[k,p]$ for arbitrary designs $\mathbf{X}$. Our corresponding upper bounds match these lower bounds when the restricted eigenvalues of $\mathbf{X}^T\mathbf{X}$ are close to each other. However, these bounds do not agree anymore when these restricted eigenvalues are away from each other. Deriving the exact dependency of the minimax risks on $\mathbf{X}$ would require sharper lower bounds and the analysis of new estimation procedures.

Our minimax results use the Gaussianity of the noise $\boldsymbol{\epsilon}$ and the Gaussianity of the design $\mathbf{X}$ in the random design setting. In an ultra-high dimensional setting, the minimax upper bounds do not seem to be robust with respect to the Gaussianity. In smaller dimensions $(k[1+\log(p/k)] < n)$, the Gaussian distribution of the design is less critical. For instance, consider a design $\mathbf{X}$ where all the components are independent and follow a subgaussian distribution. By a result of Rudelson and Vershynin [41], the restricted eigenvalues of $\mathbf{X}^T\mathbf{X}$ remain away from 0 with high probability. Consequently, some of the minimax bounds should still hold for subgaussian designs. Nevertheless, the derivation of sharp minimax bounds for non-Gaussian designs and noises remains an open problem

9. Proofs of the minimax lower bounds

Some propositions contain both minimax lower bounds and upper bounds. This section is devoted to the proof of all the lower bounds, while the upper bounds are proved in Appendix B in [43]. In order to keep our notations as short as possible, we set

$$\eta = 2(1-\alpha-\delta)\ .$$

We also note $\|.\|_{TV}$ for the total variation norm. For any subset $\mathcal{T} \subset \mathbb{R}^p$, $\alpha \in (0,1)$, covariance matrix Σ, and any variance σ^2, we denote $\beta^R_{\Sigma,\sigma,\alpha}(\mathcal{T})$ the quantity

$$\beta^R_{\Sigma,\sigma,\alpha}(\mathcal{T}) := \inf_{\Phi_\alpha} \sup_{\theta_0\in\mathcal{T}} \mathbb{P}_{\sigma\theta_0,\sigma}[\Phi_\alpha = 0]\ ,$$

the infimum being taken over all tests Φ_α satisfying $\mathbb{P}_{0_p,\sigma}[\Phi_\alpha = 0] \leq \alpha$. Its counterpart for unknown variance is defined by

$$\beta^R_{\Sigma,\alpha}(\mathcal{T}) := \inf_{\Phi_\alpha} \sup_{\sigma>0,\ \theta_0\in\mathcal{T}} \mathbb{P}_{\sigma\theta_0,\sigma}[\Phi_\alpha = 0]\ ,$$

the infimum being taken over all tests Φ_α satisfying $\sup_{\sigma>0}\mathbb{P}_{0_p,\sigma}[\Phi_\alpha = 0] \leq \alpha$. Similarly, we define $\beta^F_{\mathbf{X},\sigma,\alpha}(\mathcal{T})$ for fixed design and $\beta^F_{\mathbf{X},\alpha}(\mathcal{T})$ for fixed design and unknown variance.

Most of the minimax lower bounds in this paper are based on an approach which goes back to Ingster [28, 29, 30]. The following lemma encompasses fixed and random design and known and unknown variance.

Lemma 9.1. *Let $\mathcal{T}$ be a subset of $\mathbb{R}^p \setminus \{0_p\}$ and let σ and σ_0 be two positive integers. Consider μ a probability measure on $\sigma\mathcal{T} := \{\sigma\theta,\ \theta\in\mathcal{T}\}$. We note*

$\mathbb{P}_{\mu,\sigma} = \int_{\sigma\mathcal{T}} \mathbb{P}_{\theta,\sigma} d\mu$ and $L_\mu = d\mathbb{P}_{\mu,\sigma}/d\mathbb{P}_{0_p,\sigma_0}$. Then,

$$\begin{aligned} \beta_\alpha(\mathcal{T}) &\geq 1-\alpha-\frac{1}{2}\|\mathbb{P}_{\mu,\sigma}-\mathbb{P}_{0_p,\sigma_0}\|_{TV}. \\ &\geq 1-\alpha-\frac{1}{2}\left(\mathbb{E}_{0_p,\sigma_0}\left[L_\mu^2(\mathbf{Y},\mathbf{X})\right]-1\right)^{1/2} . \qquad (9.1) \end{aligned}$$

Here, $\beta_\alpha(\mathcal{T})$ can be replaced by $\beta^F_{\mathbf{X},\alpha}(\mathcal{T})$ or $\beta^R_{\Sigma,\alpha}(\mathcal{T})$. If we also have $\sigma = \sigma_0$, then $\beta_\alpha(\mathcal{T})$ can be replaced by $\beta^R_{\Sigma,\sigma_0,\alpha}(\mathcal{T})$ or $\beta^F_{\mathbf{X},\sigma_0,\alpha}(\mathcal{T})$.

We refer to Baraud [5] Section 7.1 for a proof and further explanations in a close framework. The main idea is to find a prior probability on $\mathcal{T}$ so that the total variation distance between $\mathbb{P}_{\mu,\sigma}$ and $\mathbb{P}_{0_p,\sigma_0}$ is as large as possible. We derive from Lemma 9.1 that $\beta_\alpha(\mathcal{T}) \geq \delta$ if $\mathbb{E}_{0_p,\sigma_0}[L^2_\mu(\mathbf{Y},\mathbf{X})] \leq 1+\eta^2$.

9.1. Proof of the lower bound (4.1) in Theorem 4.1

Proof of Theorem 4.1. By homogeneity, we can assume that $\sigma^2 = \mathrm{Var}(Y|X) = 1$. We first build a suitable prior probability μ_ρ in order to apply Lemma 9.1.

Let us take a set $\hat{m}$ of size k uniformly in $\mathcal{M}(k,p)$ (defined in Section 2). Let $\xi = (\xi_j)_{1\leq j\leq p}$ be a sequence of independent Rademacher random variables. Consider some $\rho > 0$. Define $\lambda = \rho/\sqrt{k}$ and consider μ_ρ the distribution of the random variable $\theta_{\hat{m},\xi} = \sum_{j\in\hat{m}} \lambda\xi_j e_j$. $\mathbb{P}_{\mu_\rho,1}$ stands for the distribution of $(\mathbf{Y},\mathbf{X})$ with $\theta_0 \sim \mu_\rho$ and $\sigma = 1$. Here, $(e_j)_{1\leq j\leq p}$ is the orthonormal family of vectors of $\mathbb{R}^p$ defined by

$$(e_j)_i = 1 \text{ if } i=j \text{ and } (e_i)_j = 0 \text{ otherwise.}$$

The likelihood ratio $L_{\mu_\rho}(\mathbf{X},\mathbf{Y}) = \mathbb{P}_{\mu_\rho,1}/\mathbb{P}_{0_p,1}$ writes

$$L_{\mu_\rho}(\mathbf{X},\mathbf{Y}) = \mathbb{E}_{\xi,\hat{m}}\left[\exp\left(-\frac{\|\mathbf{Y}-\mathbf{X}\theta_{\hat{m},\xi}\|_n^2-\|\mathbf{Y}\|_n^2}{2}\right)\right] ,$$

where $\mathbb{E}_{\xi,\hat{m}}$ stands for the expectation with respect to the distribution of ξ and $\widehat{m}$.

In order to apply Lemma 9.1, we need to upper bound the expectation of $L^2_{\mu_\rho}(\mathbf{X},\mathbf{Y})$. Let us first take the expectation of $L^2_{\mu_\rho}(\mathbf{X},\mathbf{Y})$ with respect to $\mathbf{Y}$.

$$\begin{aligned} &\mathbb{E}_{0_p,1}\left[L^2_{\mu_\rho}(\mathbf{X},\mathbf{Y})\right] \\ &\quad= 2^{-2k}\binom{p}{k}^{-2} \\ &\qquad\times \sum_{m_1,m_2,\xi^{(1)},\xi^{(2)}} \mathbb{E}_{0_p,1}\left[e^{-(\|\mathbf{X}\theta_{m_1,\xi^{(1)}}\|_n^2+\|\mathbf{X}\theta_{m_2,\xi^{(2)}}\|_n^2)/2+\langle \mathbf{Y},\mathbf{X}(\theta_{m_1,\xi^{(1)}}+\theta_{m_2,\xi^{(2)}})\rangle_n}\right] \\ &\quad= \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{\hat{m}_1,\hat{m}_2,\xi^{(1)},\xi^{(2)}}\left\{\exp\left(\langle\mathbf{X}\theta_{\hat{m}_1,\xi^{(1)}},\mathbf{X}\theta_{\hat{m}_2,\xi^{(2)}}\rangle_n\right)\right\}\right] , \qquad (9.2) \end{aligned}$$

where $\mathbb{E}_{\mathbf{X}}$ stands for the expectation with respect to $\mathbf{X}$ while $\mathbb{E}_{\hat{m}_1,\hat{m}_2,\xi^{(1)},\xi^{(2)}}$ refers to the expectation with respect to the independent variables $\xi^{(1)}$, $\xi^{(2)}$, m_1 and m_2.

Lemma 9.2. *If we assume that* $\rho^2 \leq C\left[\frac{k}{n}\log\left(1+\frac{p}{k^2}\right)\wedge\frac{1}{\sqrt{n}}\right]$, *then we have*

$$\mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{0_p,1}\left\{L^2_{\mu_\rho}\left(\mathbf{Y},\mathbf{X}\right)\Big|\mathbf{X}\right\}\right] \leq 1+\eta^2\ .$$

In this lemma, we have specifically distinguished the integration with respect to $\mathbf{X}$ from the integration with respect to $\mathbf{Y}$. This will be useful for deriving minimax lower bound in fixed design (Proposition 4.2). Gathering Lemmas 9.1 and 9.2 allows to derive that

$$(\rho^*_R[k,I_p])^2 \geq C\left[\frac{k}{n}\log\left(1+\frac{p}{k^2}\right)\wedge\frac{1}{\sqrt{n}}\right]\ .$$

This last bound allows to conclude since $p \geq n^2$. □

Proof of Lemma 9.2. Let us fix m_1, m_2, $\xi^{(1)}$ and $\xi^{(2)}$. First, we shall compute the expectation $\mathbb{E}_{\mathbf{X}}[\exp(\langle\mathbf{X}\theta_{m_1,\xi^{(1)}},\mathbf{X}\theta_{m_2,\xi^{(2)}}\rangle_n)]$.

Let us decompose the set $m_1\cup m_2$ into four sets (which possibly are empty): $m_1\setminus m_2$, $m_2\setminus m_1$, m_3, and m_4, where m_3 and m_4 are defined by $m_3 := \{j\in m_1\cap m_2|\xi^{(1)}_j=\xi^{(2)}_j\}$ and $m_4 := \{j\in m_1\cap m_2|\xi^{(1)}_j=-\xi^{(2)}_j\}$. For the sake of simplicity, we reorder the elements of $m_1\cup m_2$ from 1 to $|m_1\cup m_2|$ such that the first elements belong to $m_1\setminus m_2$, then to $m_2\setminus m_1$ and so on.

$$\begin{aligned}
&\mathbb{E}_{\mathbf{X}}\left[\exp\left(\langle\mathbf{X}\theta_{m_1,\xi^{(1)}},\mathbf{X}\theta_{m_2,\xi^{(2)}}\rangle_n\right)\right]\\
&\quad= \left[\int_{\mathbb{R}^p}(2\pi)^{-p/2}\exp\left(-\sum_{i=1}^p t_i^2/2+\sum_{1\leq i,j\leq p}[\theta_{m_1,\xi^{(1)}}]_i[\theta_{m_2,\xi^{(2)}}]_j t_it_j\right)\prod_{i=1}^p dt_i\right]^n\\
&\quad= \left|I_{|m_1\cup m_2|}-\lambda^2C\right|^{-n/2}\ ,
\end{aligned}$$

where $I_{|m_1\cup m_2|}$ is the identity matrix of size $|m_1\cup m_2|$ and C is block symmetric matrix of size $|m_1\cup m_2|$ defined by

$$C := \begin{bmatrix} 1 & 0 & 1 & 1\\ 0 & 1 & 1 & 1\\ 1 & 1 & 2 & 0\\ 1 & 1 & 0 & -2\end{bmatrix}\ .$$

Each block corresponds to one of the four previously defined subsets of $m_1\cup m_2$ (i.e. $m_1\setminus m_2$, $m_2\setminus m_1$, m_3, and m_4). The matrix C is of rank at most four. Hence, $I_{|m_1\cup m_2|}-\lambda^2C$ has the same determinant as the matrix D of size 4 defined by:

$$D := \begin{bmatrix} 1-\frac{\lambda^2}{n}|m_1\setminus m_2| & 0 & -\frac{\lambda^2}{n}|m_3| & -\frac{\lambda^2}{n}|m_4|\\ 0 & 1-\frac{\lambda^2}{n}|m_2\setminus m_1| & -\frac{\lambda^2}{n}|m_3| & -\frac{\lambda^2}{n}|m_4|\\ -\frac{\lambda^2}{n}|m_1\setminus m_2| & -\frac{\lambda^2}{n}|m_2\setminus m_1| & 1-2\frac{\lambda^2}{n}|m_3| & 0\\ -\frac{\lambda^2}{n}|m_1\setminus m_2| & -\frac{\lambda^2}{n}|m_2\setminus m_1| & 0 & 1+2\frac{\lambda^2}{n}|m_4|\end{bmatrix}\ .$$

After some computations, we lower bound the determinant of D

$$|D| \quad \geq \quad 1-2(2|m_3|-|m_1\cap m_2|)\lambda^2-8\rho^4 \quad .$$

From now on, we assume that $\rho^2\leq 1/20$ so that $|D|\geq 1/2$. Hence, we get

$$\begin{aligned}&\mathbb{E}_{\mathbf{X}}[\exp(\langle \mathbf{X}\theta_{m_1,\xi^{(1)}},\mathbf{X}\theta_{m_2,\xi^{(2)}}\rangle_n)]\\ \leq\quad & \left[1-2(2|m_3|-|m_1\cap m_2|)\lambda^2-8\rho^4\right]^{-n/2}\\ \leq\quad & \exp\left(8n\rho^4\right)\exp\left[2n\lambda^2(2|m_3|-|m_1\cap m_2|)\right] \quad . \qquad (9.3)\end{aligned}$$

Then, we take the expectation with respect to $\xi^{(1)}$, $\xi^{(2)}$, m_1 and m_2. When m_1 and m_2 are fixed the expression (9.3) depends on $\xi^{(1)}$ and $\xi^{(2)}$ only through the cardinality of m_3. As $\xi^{(1)}$ and $\xi^{(2)}$ follow independent Rademacher distributions, the random variable $2|m_3|-|m_1\cap m_2|$ follows the distribution of Z, a sum of $|m_1\cap m_2|$ independent Rademacher variables and

$$\mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{0_p,1}\left\{L^2_{\mu_\rho}\left(\mathbf{Y},\mathbf{X}\right)\Big|\mathbf{X}\right\}\right]\leq \exp\left(8n\rho^4\right)\mathbb{E}_Z\left[\exp\left(2n\lambda^2Z\right)\right] \ , \qquad (9.4)$$

where $\mathbb{E}_Z$ stands for the expectation with respect to Z. We now proceed as in the proof of Theorem 1 in Baraud [5] in order to upper bound the term

$$\mathbb{E}_Z\left[\exp\left(2n\lambda^2Z\right)\right] \quad = \quad \binom{p}{k}^{-2}\sum_{m_1,m_2\in\mathcal{M}(k,p)}\cosh\left(2n\lambda^2\right)^{|m_1\cap m_2|} \quad .$$

Following Baraud's arguments, we get that $\mathbb{E}_Z\left[\exp\left(2n\lambda^2Z\right)\right]\leq\sqrt{1+\eta^2}$ when

$$\rho^2\leq C\frac{k}{n}\log\left(1+\frac{p}{k^2}\vee\sqrt{\frac{p}{k^2}}\right) \quad .$$

Moreover, we have $\exp(8\rho^4 n)\leq\sqrt{1+\eta^2}$ as soon as $\rho^2\leq C/\sqrt{n}$ since $\eta\geq 0.94$. Gathering these observations with (9.4), we conclude that

$$\mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{0_p,1}\{L^2_{\mu_\rho}\left(\mathbf{Y},\mathbf{X}\right)\Big|\mathbf{X}\}\right]\leq 1+\eta^2$$

as soon as

$$\rho^2\leq C\left[\frac{k}{n}\log\left(1+\frac{p}{k^2}\vee\sqrt{\frac{p}{k^2}}\right)\wedge\frac{1}{\sqrt{n}}\right] \quad .$$

□

9.2. Proof of the lower bound (4.8) in Theorem 4.3

Proof of (4.8) in Theorem 4.3. Consider the Condition

$$(\mathbf{A}.\mathbf{1}) \qquad \frac{k}{n}\log\left(\frac{p}{e^3k^2}\right)\geq 2\,.$$

We deduce Theorem 4.3 from the following result.

Lemma 9.3. *Suppose that $\alpha + \delta \leq 53\%$. We have*

$$\beta^R_{I_p,\alpha}\left(\left\{\theta_0 \in \Theta[k,p], \|\theta_0\|_p^2 = \rho^2\right\}\right) \geq \delta \ , \tag{9.5}$$

for any $\rho^2 > 0$ such that

$$\rho^2 \leq \frac{k}{2n}\log\left(1+\frac{p}{k^2}\right) \ . \tag{9.6}$$

*If we assume that Condition (**A.1**) holds, (9.5) holds for any $\rho > 0$ such that*

$$\rho^2 \leq -1 + \left(\frac{p}{2ek}\right)^{\frac{k}{n}} (4k)^{-2/n} \ . \tag{9.7}$$

If $p \geq k^3 \vee C$ and $k\log(p)/n \geq C_1$ with C and C_1 large enough, then Assumption (**A.1**) is satisfied. For C large enough, the quantity $k\log(p)/\log(k)$ is large enough so that the lower bound (9.7) satisfies

$$\begin{aligned} -1 + \left(\frac{p}{2ek}\right)^{\frac{k}{n}} (4k)^{-2/n} &\geq -1 + \exp\left[C\frac{k}{n}\log(p)\right] \\ &\geq C_1\frac{k}{n}\log(p)\exp\left[C_2\frac{k}{n}\log(p)\right] \ . \end{aligned}$$

Let us now assume that $p \geq k^3 \vee C$ and $k\log(p)/n \leq C_1$ where C_1 has been previously fixed. Then, the first lower bound (9.6) satisfies:

$$\frac{k}{2n}\log\left(1+\frac{p}{k^2}\right) \geq C_1\frac{k}{n}\log(p)\exp\left[C_2\frac{k}{n}\log(p)\right] \ .$$

Gathering the two previous lower bounds with Lemma 9.3 allows to conclude. □

Proof of Lemma 9.3. Consider some $\rho > 0$. To apply Lemma 9.1, we first have to define a suitable prior μ_ρ on θ_0 and a suitable σ^2. More specifically, we set $\sigma^2 = (1+\rho^2)^{-1}$ and the distribution μ_ρ is supported by $\Theta[k,p,\rho]$ defined by

$$\Theta[k,p,\rho] \quad := \quad \left\{\theta_0 \in \Theta[k,p] \ , \ \|\theta_0\|_p^2 = \frac{\rho^2}{1+\rho^2}\right\} \ .$$

Let $\hat{m}$ be a random variable uniformly distributed over $\mathcal{M}(k,p)$. Let μ_ρ be the distribution of the random variable $\widehat{\theta} = \sum_{j\in\hat{m}} \lambda e_j$ where

$$\lambda^2 := \frac{\rho^2}{k(1+\rho^2)} \ ,$$

and where $(e_j)_{1\leq j\leq p}$ is the orthonormal family of vectors of $\mathbb{R}^p$ defined by $(e_j)_i = 1$ if $i = j$ and $(e_i)_j = 0$ otherwise. By Lemma 9.1, we only have to prove under conditions (9.6) or (9.7) with (**A.1**), we have

$$\mathbb{E}_{0_p,1}(L^2_{\mu_\rho}(\mathbf{Y},\mathbf{X})) \leq 1 + \eta^2 \ . \tag{9.8}$$

Observe here that we use a variance 1 for $\mathbf{H_0}$ and a variance $1-\|\theta_0\|_p^2$ for the hypothesis $\mathbf{H_1}$. Using these two different variances allows us to take advantage of the fact that we work under unknown variance.

As a specific case of [44] Eq.(8.5), we have

$$\begin{aligned}\mathbb{E}_{0_p,1}(L^2_{\mu_\rho}(\mathbf{Y},\mathbf{X})) &= \binom{p}{k}^{-2}\sum_{m_1,m_2\in\mathcal{M}(k,p)}\left(1-\frac{\rho^2|m_1\cap m_2|}{(1+\rho^2)k}\right)^{-n}\\ &= \mathbb{E}_Z\left[\left(1-\frac{\rho^2 Z}{(1+\rho^2)k}\right)^{-n}\right],\end{aligned}$$

where Z follows an hypergeometric distribution with parameters p, k, and k/p. We know from Aldous (p.173) [2] that Z follows the same distribution as the random variable $\mathbb{E}(W|\mathcal{B}_p)$ where W is a binomial random variable of parameters k, k/p and $\mathcal{B}_p$ some suitable σ-algebra. By a convexity argument, we get

$$\mathbb{E}_Z\left[\left(1-\frac{\rho^2 Z}{(1+\rho^2)k}\right)^{-n}\right]\leq \mathbb{E}_W\left[\left(1-\frac{\rho^2 W}{(1+\rho^2)k}\right)^{-n}\right]. \tag{9.9}$$

Hence, we only need to upper bound the expectation of the second random variable.

CASE 1: Proof of Equation (9.6) Since $\log(1+x)\leq x$ and since $W\leq k$, we have

$$\begin{aligned}\mathbb{E}_W\left[\left(1-\frac{\rho^2 W}{(1+\rho^2)k}\right)^{-n}\right] &\leq \mathbb{E}_W\left[\exp\left(\frac{n\rho^2W/k}{1+\rho^2-\rho^2W/k}\right)\right]\\ &\leq \mathbb{E}_W\left[\exp\left(n\rho^2W/k\right)\right]\\ &\leq \left[1+\frac{k}{p}\left(e^{n\rho^2/k}-1\right)\right]^k\\ &\leq \exp\left[\frac{k^2}{p}(e^{n\rho^2/k}-1)\right].\end{aligned}$$

As a consequence, the condition (9.8) holds if $\rho^2\leq\frac{k}{n}\log\left[1+\frac{p}{k^2}\log(1+\eta^2)\right]$. Observe that $\log(1+\eta^2)\geq 0.6$. Since $\log(1+ux)\geq u\log(1+x)$ for any $0<u<1$ and any $x>0$, the last condition is enforced by $\rho^2\leq\frac{k}{2n}\log\left[1+\frac{p}{k^2}\right]$.

CASE 2: Proof of Equation (9.7). Here, we bound (9.9) under condition (**A**.**1**). We have

$$\mathbb{E}_W\left[\left(1-\frac{\rho^2 W}{(1+\rho^2)k}\right)^{-n}-1\right] \leq \sum_{i=1}^{k}\mathbb{P}\left[W\geq i\right]\left(1-\frac{\rho^2 i}{(1+\rho^2)k}\right)^{-n}.$$

Since we need to ensure that $\mathbb{E}_W[\{1-\rho^2 W/((1+\rho^2)k)\}^{-n}-1] \leq \eta^2$, it is sufficient to prove that

$$\mathbb{P}\left[W \geq i\right]\left(1-\frac{i}{k}\right)^{-n} \leq \frac{\eta^2 i^{-i}}{4} \text{ for any } 1 \leq i \leq \lfloor k/2 \rfloor \ , \tag{9.10}$$

$$\mathbb{P}\left[W \geq i\right]\left(1-\frac{\rho^2 i}{(1+\rho^2)k}\right)^{-n} \leq \frac{\eta^2}{2k} \text{ for any } \lfloor k/2 \rfloor + 1 \leq i \leq k \ . \tag{9.11}$$

In order to prove these bounds, we shall use a deviation inequality of the random variable W/k.

Lemma 9.4. *For any $k \geq 1$, $0 < x \leq 1$, it holds that*

$$\mathbb{P}\left[\frac{W}{k} \geq x\right] \leq \left[\left(\frac{k}{px}\right)^x \frac{1}{(1-x)^{1-x}}\right]^k \ . \tag{9.12}$$

FACT 1. For any $1 \leq i \leq \lfloor k/2 \rfloor$, the upper bounds (9.10) hold under Condition (**A.1**).

FACT 2. The upper bound (9.11) holds for any $\lfloor k/2 \rfloor + 1 \leq i \leq k$ as soon as

$$\rho^2 \leq -1 + \left(\frac{p}{2ek}\right)^{k/n}\left(\frac{\eta^2}{2k}\right)^{2/n} \ . \tag{9.13}$$

We derive that under (9.13), we have $\mathbb{E}_{0_p,1}[L^2_{\mu_\rho}(\mathbf{Y},\mathbf{X})] \leq 1+\eta^2$. The fact that $\eta^2 \geq 1/2$ allows to conclude. □

Proof of FACT 1. Since $\log(1-x) \geq -x/(1-x)$ for any $0 \leq x < 1$, we derive that $(1-x)^{1-x} \geq e^{-x}$. Gathering this bound with Lemma 9.4, we get a new deviation inequality for W.

$$\mathbb{P}\left[\frac{W}{k} \geq x\right] \leq \left(\frac{ke}{px}\right)^{xk} \ , \tag{9.14}$$

for any $x < 1$. We apply this bound with $x = i/k$. Then, Inequality (9.10) holds if

$$\left(\frac{k^2 e}{p}\right)^{i/n}\left(\frac{4}{\eta^2}\right)^{1/n} \leq 1 - \frac{i}{k} \ .$$

Taking the logarithm of this expression leads to

$$-\frac{i}{n}\log\left(\frac{p}{ek^2}\right) + \frac{1}{n}\log\left(4/\eta^2\right) + \frac{i/k}{1-i/k} \leq 0 \ ,$$

Since i is constrained to be smaller than $k/2$, we get

$$-\frac{ik}{n}\log\left(\frac{p}{ek^2}\right) + \frac{k}{n}\log\left(4/\eta^2\right) + 2i \leq 0 \ .$$

By Assumption (**A.1**), $k/n \log[p/(ek^2)]$ is larger than 2. Consequently, the worst case among all i between 1 and $k/2$ is $i = 1$. Hence, we only need to prove that

$$\frac{k}{n}\left[\log\left(\frac{p}{k^2}\right) - \log\left(\frac{4e}{\eta^2}\right)\right] \geq 2\ .$$

Since η is larger than 0.94, $\log(4e/\eta^2)$ is smaller than 3 and this last inequality is ensured by Assumption (**A.1**). □

Proof of FACT 2. We consider here the case $1/2 < i/k \leq 1$. We derive from (9.14) that

$$\mathbb{P}\left[W \geq i\right] \leq \left(\frac{2ek}{p}\right)^i\ .$$

Consequently, we want to ensure that

$$\left(\frac{2ek}{p}\right)^{i/n}\left(\frac{2k}{\eta^2}\right)^{1/n} \leq \left(1 - \frac{\rho^2 i}{(1+\rho^2)k}\right)\ ,$$

for any i between $\lfloor k/2 \rfloor$ and k. For any x and u between 0 and 1, $(1-x)^u \leq (1-xu)$. Setting $u = i/k$ and $x = \rho^2/(1+\rho^2)$, we obtain that the last inequality holds if

$$1 - \frac{\rho^2}{1+\rho^2} \geq \sup_{\lfloor k/2 \rfloor \leq i \leq k}\left(\frac{2ek}{p}\right)^{k/n}\left(\frac{2k}{\eta^2}\right)^{k/(in)}$$

Since $2k/\eta^2$ is positive, the largest term in the bound corresponds to $i = k/2$. Hence, it remains to prove that

$$\frac{1}{1+\rho^2} \geq \left(\frac{2ek}{p}\right)^{k/n}\left(\frac{2k}{\eta^2}\right)^{2/n}$$

We conclude that the upper bounds hold if

$$\rho^2 \leq -1 + \left(\frac{p}{2ek}\right)^{k/n}\left(\frac{\eta^2}{2k}\right)^{2/n}\ .$$

□

Proof of Lemma 9.4. We prove this deviation inequality using the Laplace transform of W/k. Consider some $x \in (0,1)$ and $\lambda > 0$.

$$\begin{aligned}
\log\left[\mathbb{P}\left\{\frac{W}{k} \geq x\right\}\right] &\leq -\lambda x + \log\left[\mathbb{E}_W\left\{\exp(\lambda W/k)\right\}\right] \\
&\leq -\lambda x + k\log\left[1 + \frac{k}{p}\left(\exp\left(\frac{\lambda}{k}\right) - 1\right)\right]\ .
\end{aligned}$$

Deriving with respect to λ an upper bound of the last expression leads to the following choice

$$e^{\lambda^*/k} = \frac{x}{1-x}\left(\frac{p}{k}-1\right) \ .$$

Hence, we get

$$\log\left[\mathbb{P}\left\{\frac{W}{k}\geq x\right\}\right] \leq -kx\log\left[\frac{x}{1-x}\left(\frac{p}{k}-1\right)\right] + k\log\left[\frac{1-k/p}{1-x}\right] \ .$$

Since we assume $x<1$, we conclude that

$$\mathbb{P}\left\{\frac{W}{k}\geq x\right\} \leq \left[\left(\frac{k}{px}\right)^x \frac{1}{(1-x)^{1-x}}\right]^k \ .$$

Since $\mathbb{P}(W=k) = [k/p]^k$, this upper bound is also valid when $x=1$. □

9.3. *Proof of Proposition 5.1*

We derive this minimax lower bound from the hypothesis testing problem $\{\theta_0 = 0_p\}$ studied in Section 4. Since the covariance $\Sigma = I_p$, the loss $\mathbb{E}\left[\{X(\theta_1-\theta_2)\}^2/\sigma^2\right]$ is simply $\|\theta_1-\theta_2\|_p^2/\sigma^2$. For the sake of simplicity, we assume that p is even. We split the p covariates into two groups M_1 and M_2 of size $p/2$. Given some $\rho>0$, we fix $\sigma^2=(1+\rho^2)^{-1}$ and we consider the two sets

$$\begin{aligned}\Theta_1[\rho] &= \Theta[k,p]\cap\left\{\theta : \mathrm{supp}(\theta)\subset M_1 \text{ and } \|\theta\|_p^2 = \frac{\rho^2}{1+\rho^2}\right\}\\ \Theta_2[\rho] &= \Theta[k,p]\cap\left\{\theta : \mathrm{supp}(\theta)\subset M_2 \text{ and } \|\theta\|_p^2 = \frac{\rho^2}{1+\rho^2}\right\} \ .\end{aligned}$$

Take any estimator $\widehat{\theta}$. We consider an estimator $\widetilde{\theta}\in\Theta_1[\rho]\cup\Theta_2[\rho]$ such that

$$\|\widetilde{\theta}-\widehat{\theta}\|_p = \min_{\theta\in\Theta_1[\rho]\cup\Theta_2[\rho]}\|\theta-\widehat{\theta}\|_p \ .$$

By the triangle inequality, we have $\|\widetilde{\theta}-\theta_0\|_p \leq 2\|\widehat{\theta}-\theta_0\|_p$, for any $\theta_0\in\Theta_1[\rho]\cup\Theta_2[\rho]$.

$$\sup_{i=1,2}\sup_{\theta_0\in\Theta_i[\rho]}\mathbb{E}_{\theta_0,\sigma}\left[\frac{\|\widehat{\theta}-\theta_0\|_p^2}{\sigma^2}\right] \geq \frac{\rho^2}{4}\sup_{i=1,2}\sup_{\theta_0\in\Theta_i[\rho]}\mathbb{P}_{\theta_0,\sigma}[\mathrm{supp}(\widetilde{\theta})\nsubseteq M_i] \ . \quad (9.15)$$

It is enough to prove that for $\rho^2 = C_1\frac{k}{n}\log(p)\exp\{C_2\frac{k}{n}\log(p)\}$, the supremum of the probabilities $\mathbb{P}_{\theta_0,\sigma}[\mathrm{supp}(\widetilde{\theta})\nsubseteq M_i]$ is lower bounded by a positive constant. This is equivalent to lower bounding the minimax separation distance for $\mathbf{H_0}$: $\{\theta_0\in\Theta_1[\rho]$ and $\sigma^2=(1+\rho^2)^{-1}\}$ against $\mathbf{H_1}$: $\{\theta_0\in\Theta_2[\rho]$ and $\sigma^2=(1+\rho^2)^{-1}\}$.

As in the proof of Theorem 4.3, we build a prior distribution $\mu_{1,\rho}$ on θ_0. Consider the collection $\mathcal{M}_1(k)$ of subsets of M_1 of size k. Let $\hat{m}$ be be some random variable uniformly distributed over $\mathcal{M}_1(k)$. Then, $\mu_{1,\rho}$ is the distribution of $\widehat{\theta} = \sum_{j\in\hat{m}} \rho/\sqrt{k(1+\rho^2)}e_j$. Similarly, we define the prior distribution $\mu_{2,\rho}$ on $\Theta_2[\rho]$. We note $\mathbb{P}_{\mu_{i,\rho},\sigma} = \int \mathbb{P}_{\theta_0,\sigma} d\mu_{i,\rho}$. We have

$$\begin{aligned}\sup_{i=1,2}\sup_{\theta_0\in\Theta_i[r]} \mathbb{P}_{\theta_0}[\operatorname{supp}(\widetilde{\theta}) \nsubseteq M_i] &\geq 1 - \frac{1}{2}\|\mathbb{P}_{\mu_{1,\rho},\sigma} - \mathbb{P}_{\mu_{2,\rho},\sigma}\|_{TV} \ . \\ &\geq 1 - \|\mathbb{P}_{\mu_{1,\rho},\sigma} - \mathbb{P}_{0_p,1}\|_{TV} \ , \qquad (9.16)\end{aligned}$$

by the triangle inequality. Lemma 9.1 states that

$$\|\mathbb{P}_{\mu_{1,\rho}} - \mathbb{P}_{0_p,1}\|_{TV} \leq \mathbb{E}_{0_p,1}\left[L^2_{\mu_{1,\rho}} - 1\right] \ ,$$

where $L_{\mu_{1,\rho}} = \mathrm{d}\mathbb{P}_{\mu_{1,\rho},\sigma}/\mathrm{d}\mathbb{P}_{0_p,1}$. In fact, the second moment of $L_{\mu_{1,\rho}}$ has been studied in the proof of Theorem 4.3. If we take $\alpha+\delta = 53\%$ in this proof, we derive $\mathbb{E}_0[L^2_{\mu_{1,\rho}}] \leq 1.9$ if

$$\frac{\rho^2}{1-\rho^2} \leq C_1 \frac{k\log(p)}{n}\exp\left(C_2\frac{k\log(p)}{n}\right) \text{ and if } p \geq k^3 \vee C_3 \ .$$

Gathering this result with Equations (9.15) and (9.16) allows to conclude.

9.4. Proof of Proposition 5.6

Let us set $\alpha = \delta = 0.01$. Consider a design $\mathbf{X}$ that achieves the bound (4.11) and take $\rho = \rho^*_{F,U}[k,\mathbf{X}]/2$. If $k\log(p)/n$ is large enough, then $\rho \geq \sqrt{2}$. Take any estimator $\widehat{\theta}$ that does not rely on the variance σ^2. Let us build a test T of the hypotheses $\mathbf{H_0}$: $\{\theta_0 = 0 \text{ and } \sigma > 0\}$ against $\mathbf{H_1}$: $\{\theta_0 \in \Theta[k,p] \text{ and } \sigma > 0, \|\mathbf{X}\theta_0\|_n^2/(n\sigma^2) \geq \rho^2\}$:

$$T = \begin{cases} 0 & \text{if} \quad 2\|\mathbf{X}\widehat{\theta}\|_n^2 < \|\mathbf{Y}\|_n^2 \\ 1 & \text{if} \quad 2\|\mathbf{X}\widehat{\theta}\|_n^2 \geq \|\mathbf{Y}\|_n^2 \end{cases}$$

By Proposition 4.4, we have at least one of the two following properties:

$$\sup_{\sigma>0}\mathbb{P}_{0_p,\sigma}(T=1) \geq \alpha \qquad (9.17)$$

$$\sup_{\sigma>0,\ \theta_0\in\Theta[k,p],\ \|\mathbf{X}\theta_0\|_n^2/(n\sigma^2)\geq\rho^2} \mathbb{P}_{\theta_0,\sigma}(T=0) \geq \delta \qquad (9.18)$$

CASE 1: (9.17) holds. We have $\mathbb{P}_{0_p,\sigma}[\|\mathbf{Y}\|_n^2 \geq n\sigma^2/2] \geq 1 - e^{-n/16}$ for any $\sigma > 0$. Thus, there exists $\sigma > 0$ such that $\|\mathbf{X}\widehat{\theta}\|_n^2 \geq n\sigma^2/4$ with probability larger than $\alpha/2 - e^{-n/16}$. As a consequence, we have

$$\sup_{\sigma>0}\mathbb{E}_{0_p,\sigma}\left[\|\mathbf{X}(\widehat{\theta}-\theta_0)\|_n^2/[n\sigma^2]\right] \geq C \ .$$

CASE 2: (9.18) holds. The random variable $\|\mathbf{Y}\|_n^2/\sigma^2$ follows a noncentral χ^2 distribution with n degrees of freedom and a non centrality parameter $\|\mathbf{X}\theta_0\|_n^2/\sigma^2$. By Lemma 1 in Birgé [9], we have $\|\mathbf{Y}\|_n^2 \leq 3/2\left[n\sigma^2 + \|\mathbf{X}\theta_0\|_n^2\right]$, with probability larger than $1 - e^{-Cn}$. Consequently, there exist $\sigma > 0$ and $\theta_0 \in \Theta[k,p]$ such that $\|\mathbf{X}\theta_0\|_n^2/(n\sigma^2) \geq \rho^2$ and

$$\|\mathbf{X}\widehat{\theta}\|_n^2/(n\sigma^2) \leq \frac{3}{4}\left[1 + \|\mathbf{X}\theta_0\|_n^2/(n\sigma^2)\right] \leq \frac{7}{8}\|\mathbf{X}\theta_0\|_n^2/(n\sigma^2),$$

with probability $\delta/2 - e^{-Cn}$, since $\rho^2 \geq 2$. Thus, we get

$$\begin{aligned}\mathbb{E}_{\theta_0,\sigma}\left[\|\mathbf{X}(\widehat{\theta}-\theta_0)\|_n^2/n\right] &\geq \mathbb{E}_{\theta_0,\sigma}\left[\left(\|\mathbf{X}\widehat{\theta}\|_n - \|\mathbf{X}\theta_0\|_n\right)^2/n\right] \\ &\geq C\|\mathbf{X}\theta_0\|_n^2/n \ \geq C\rho^2\sigma^2 \ .\end{aligned}$$

9.5. Proof of Proposition 8.1

For the sake of conciseness, we note $l(\widehat{\sigma},\sigma) = |\widehat{\sigma}^2/\sigma^2 - \sigma^2/\widehat{\sigma}^2|$. Given a positive number ρ, we note $\sigma_0 = (1+\rho^2)^{-1/2}$. As in the proof of Theorem 4.3, we consider the prior probability μ_ρ on $\Theta[k,p]$. For any estimator $\widehat{\sigma} > 0$, we define $\widetilde{\sigma}$ by $\widetilde{\sigma} \in \arg\min_{\sigma\in\{1,\sigma_0\}} l(\widehat{\sigma},\sigma)$. For any $\sigma \in \{1,\sigma_0\}$, the loss $l(\widehat{\sigma},\sigma)$ is controlled as follows:

$$l(\widehat{\sigma},\sigma) \geq \mathbf{1}_{\widetilde{\sigma}\neq\sigma} l(1,\sqrt{\sigma_0}) \ .$$

Thus, we get the minimax lower bound

$$\begin{aligned}&\inf_{\widehat{\sigma}} \sup_{\sigma>0,\ \theta_0\in\Theta[k,p]} \mathbb{E}_{\theta_0,\sigma}\left[l(\widehat{\sigma},\sigma)\right] \\ &\quad\geq \inf_{\widehat{\sigma}>0} \max\left[\mathbb{E}_{0_p,1}\left\{l(\widehat{\sigma},1)\right\}; \mathbb{E}_{\mu_\rho,\sigma_0}\left\{l(\widehat{\sigma},\sigma_0)\right\}\right] \\ &\quad\geq l(1,\sqrt{\sigma_0}) \inf_{\widetilde{\sigma}\in\{1,\sigma_0\}} \max\left[\mathbb{P}_{0_p,1}[\widetilde{\sigma}\neq 1]; \mathbb{P}_{\mu_\rho,\sigma_0}[\widetilde{\sigma}\neq\sigma_0]\right] \\ &\quad\geq \frac{l(1,\sqrt{\sigma_0})}{2}\left[1 - \frac{\|\mathbb{P}_{0_p,1} - \mathbb{P}_{\mu_\rho,1}\|_{TV}}{2}\right] \\ &\quad\geq \frac{\rho^2}{2\sqrt{1+\rho^2}}\left[1 - \frac{1}{2}\left(\mathbb{E}_{0_p,1}[L^2_{\mu_\rho}(\mathbf{Y},\mathbf{X})] - 1\right)^{1/2}\right] \ . \qquad (9.19)\end{aligned}$$

Let us note two numbers $\eta_1 = 1.5$ and $\eta_2 = 1.8$. If $\mathbf{X}$ is a standard Gaussian design and if $k \leq p^{1/3}$, then the proof of Theorem 4.3 states for

$$\rho^2 \leq C_1\frac{k}{n}\log\left(\frac{p}{k}\right)\exp\left[C_2\frac{k}{n}\log\left(\frac{p}{k}\right)\right] \ ,$$

we have $\mathbb{E}_{0_p,1}[L^2_{\mu_\rho}(\mathbf{Y},\mathbf{X})] \leq 1 + \eta_1^2$ where the expectation is taken both with respect to $\mathbf{Y}$ and $\mathbf{X}$. Applying Markov's inequality, we derive that with positive probability,

$$\mathbb{E}_{0_p,1}[L^2_{\mu_\rho}(\mathbf{Y},\mathbf{X})|\mathbf{X}] \leq 1 + \eta_2^2 \ .$$

For such designs $\mathbf{X}$ and such ρ we have

$$\inf_{\widehat{\sigma}} \sup_{\sigma>0,\ \theta_0\in\Theta[k,p]} \mathbb{E}_{\theta_0,\sigma}\left[l(\widehat{\sigma},\sigma)\right] \geq C\frac{\rho^2}{\sqrt{1+\rho^2}} \geq C'\left(\rho\wedge\rho^2\right) \ ,$$

since $\rho^2/\sqrt{1+\rho^2} \geq (\rho\wedge\rho^2)/\sqrt{2}$. We conclude that

$$\inf_{\widehat{\sigma}} \sup_{\sigma>0,\ \theta_0\in\Theta[k,p]} \mathbb{E}_{\theta_0,\sigma}\left[l(\widehat{\sigma},\sigma)\right] \geq C'_1\frac{k}{n}\log\left(\frac{p}{k}\right)\exp\left[C'_2\frac{k}{n}\log\left(\frac{p}{k}\right)\right] \ .$$

9.6. Fano's Lemma

The next lower bounds are established applying Birgé's version of Fano's Lemma [10]. More precisely, we shall use the following lemma, which is taken from Corollary 2.19 in [37],

Lemma 9.5. *Let (S,d) be some pseudo metric space, $\{\mathbb{P}_s, s\in\mathcal{S}\}$ be some statistical model. Let us note $\kappa = 2e/(2e+1)$. Then, for any estimator $\widehat{s}$ and any finite subset $\mathcal{C}$ of $\mathcal{S}$, setting $\delta = \min_{s,t\in\mathcal{C},\ s\neq t} d(s,t)$, provided that $\max_{s,t}\mathcal{K}(\mathbb{P}_s,\mathbb{P}_t)\leq\kappa\log|\mathcal{C}|$ the following lower bound holds for any $p\geq 1$:*

$$\sup_{s\in\mathcal{C}}\mathbb{E}_s\left[d^p(s,\widehat{s})\right] \geq 2^{-p}\delta^p(1-\kappa) \ .$$

9.7. Proof of the lower bounds of Propositions 6.1 and 6.4

This lower bound is based on Fano's lemma. For the sake of simplicity, we assume that $2k\leq p$ and that $\sigma^2=1$. First, we consider a unit vector $\theta\in\Theta[2k,p]$ such that $\|\mathbf{X}\theta\|_n^2 = \Phi_{2k,-}(\mathbf{X})$. Let us define $\kappa = 2e/(2e+1)$. It is possible to find two vectors $(\theta_1,\theta_2)\in\Theta[k,p]$ such that $\theta_1-\theta_2 = \theta\sqrt{2\kappa\log(2)/\Phi_{2k,-}(\mathbf{X})}$ and $\text{supp}(\theta_1)\cap\text{supp}(\theta_2)=\emptyset$. Consequently, the Kullback distance $\mathcal{K}(\theta_1,\theta_2)$ between the two distributions $\mathbb{P}_{\theta_1}$ and $\mathbb{P}_{\theta_2}$ is exactly $\kappa\log(2)$ and $\|\theta_1-\theta_2\|_p^2 = 2\kappa\log(2)/\Phi_{2k,-}(\mathbf{X})$. Applying Lemma 9.5, we derive the first part of the lower bound:

$$\mathcal{RI}_F[k,\mathbf{X}] \geq C\frac{1}{\Phi_{2k\wedge p,-}(\mathbf{X})} \ .$$

Let us turn to the second part of the lower bound. We consider $\mathcal{M}(k,p)$ the collections of subsets of $\{1,\ldots,p\}$ of size k. Applying combinatorial results such as Varshamov's lemma and Lemma 4.10 in [37], we derive that there exists $\mathcal{M}'(k,p)\subset\mathcal{M}(k,p)$ of size larger than $\exp[Ck\log(ep/k)]$ such that any pairs of distinct sets m_1, m_2 in $\mathcal{M}'(k,p)$, we have $|m_1\cap m_3|\leq 3k/4$.

For any $m\in\mathcal{M}'(k,p)$, we define a vector θ_m that satisfies:

- $|(\theta_m)_i| = 1/\sqrt{k}$ if $i\in m$ and 0 else.
- $\|\mathbf{X}\theta_m\|_n^2 \leq \Phi_{1,+}(\mathbf{X})$.

Let us prove that this construction is possible by induction on k. The construction is straightforward for $k = 1$. Assume that this construction is possible for $k-1$. Let us take some subset $m \in \mathcal{M}(k,p)$ and $m' \subset m$ such that $|m'| = k-1$. There exists a vector θ such that $\mathrm{supp}(\theta) = m'$, $|(\theta)_i| = 1/\sqrt{k}$ for any $i \in m'$ and $\|\mathbf{X}\theta\|_n^2 \leq \Phi_{1,+}(\mathbf{X})(k-1)/k$. Now consider the two vectors θ_1 and θ_2 such that $(\theta_1)_i = (\theta_2)_i = \theta_i$ if $i \in m'$, $(\theta_1)_i = -(\theta_2)_i = 1/\sqrt{k}$ if $i \in m \setminus m'$ and $(\theta_1)_i = -(\theta_2)_i = 0$ else. It follows that $\|\mathbf{X}\theta_1\|_n^2 \leq \Phi_{1,+}(\mathbf{X})$ or $\|\mathbf{X}\theta_2\|_n^2 \leq \Phi_{1,+}(\mathbf{X})$, which allows to conclude.

For any $r > 0$, we consider the set $\mathcal{C}'_k[r] := \{r\theta_m \ , \ m \in \mathcal{M}'(k,p)\}$. The Kullback distance between any two element $\theta_1 \neq \theta_2$ in $\mathcal{C}'_k[r]$ is upper bounded as follows:

$$\mathcal{K}(\theta_1,\theta_2) = \frac{\|\mathbf{X}(\theta_1-\theta_2)\|_n^2}{2\sigma^2} \leq 2\Phi_{1k,+}(\mathbf{X})\frac{r^2}{\sigma^2} \ ,$$

while we have $\|\theta_1-\theta_2\|_p^2 \geq r^2/2$. Applying Birgé's version of Fano's lemma [10] we conclude that:

$$\inf_{\widehat{\theta}} \sup_{\theta_0 \in \mathrm{Conv}[\mathcal{C}_k^p(\sqrt{k}r)]} \mathbb{E}_{\theta_0,\sigma}\left[\|\mathbf{X}(\widehat{\theta}-\theta_0)\|_n^2/n\right] \geq C\left[r^2 \wedge \frac{k(1+\log(p/k))}{\Phi_{1k,+}(\mathbf{X})}\sigma^2\right] \ ,$$

where Conv$[A]$ stands for the convex hull of A. Taking $r^2 = k[1+\log(p/k)]\sigma^2/\Phi_{2k,+}(\mathbf{X})$ allows to conclude.

The proof of the minimax lower bound (6.7) in Proposition 6.4 follows exactly the same steps. The minimax lower bound (6.8) is a consequence of (6.7) and the fact that $\Phi_{1,+}(\sqrt{\Sigma}) = 1$ for any $\Sigma \in \mathcal{S}_p$.

9.8. Proof of Proposition 6.2

Proof of the first result. First, the minimax lower bound is a straightforward consequence of (6.2), since $\Phi_{1,+}(\mathbf{X}) = n$ if $\mathbf{X} \in \mathcal{D}_{n,p}$. Let us turn to the upper bound. Thanks to the minimax upper bound (6.1), we only have to prove that there exists a design $\mathbf{X}$ such that its $2k$-restricted eigenvalues remain close from n.

Consider a standard Gaussian design $\mathbf{W}$ of size $n \times p$. Rescaling to a norm of $\sqrt{n}$ each column of $\mathbf{W}$, we get a design $\mathbf{X} \in \mathcal{D}_{n,p}$. Let us assume that $k[1+\log(p/k)] \leq \{4(1+\sqrt{2})\}^{-2}n$. Applying Lemma 11.2, we control the restricted eigenvalues of $\mathbf{W}$:

$$\Phi_{2k,+}(\mathbf{W}/\sqrt{n}) \leq (7/4)^2 \text{ and } \Phi_{2k,-}(\mathbf{W}/\sqrt{n}) \geq (1/4)^2 \ ,$$

with probability larger than $1-\exp(-n/32)$. Consider any $\theta \in \Theta[2k,p]$ such that $\|\theta\|_p = 1$. By definition of $\mathbf{X}$, there exists some $\theta' \in \Theta[2k,p]$ such that $\mathbf{X}\theta = \mathbf{W}\theta'$. Moreover we have

$$\|\theta'\|_p^2 \geq \Phi_{1,+}^{-1}(\mathbf{W}/\sqrt{n}) \ .$$

Hence, we derive that

$$\Phi_{2k,-}(\mathbf{X}) \geq \Phi_{2k,-}(\mathbf{W})\Phi_{1,+}^{-1}(\mathbf{W}/\sqrt{n}) \ .$$

Thus, we have $\Phi_{2k,-}(\mathbf{X}) \geq n/49$ with positive probability. □

Proof of the second result. Let $\mathbf{X}$ be a design in $\mathcal{D}_{n,p}$. Take $\delta \in (0,1]$. Let us consider the collection $\mathcal{M}(k,p)$ (defined in Section 2). As explained in the proof of Proposition 6.1, there exists $\mathcal{M}'(k,p) \subset \mathcal{M}(k,p)$ of size larger than $\exp[Ck\log(ep/k)]$ such that any pairs of distinct sets m_1, m_2 in $\mathcal{M}'(k,p)$, we have $|m_1 \cap m_3| \leq 3k/4$.

For any $m \in \mathcal{M}'(k,p)$, we define a vector θ_m such that $|(\theta_m)_i| = 1/\sqrt{k}$ if $i \in m$ and 0 else and that $\|\mathbf{X}\theta_m\|_n^2 \leq n$. Such a construction is justified in the proof of Proposition 6.1.

For any $m_1 \neq m_2$ in $\mathcal{M}'(k,p)$, we have $\|\theta_{m_1} - \theta_{m_2}\|_p^2 \geq 1/2$. If there exist two distinct sets $(m_1, m_2) \in \mathcal{M}'(k,p)$ such that $\|\mathbf{X}(\theta_{m_1} - \theta_{m_2})\|_n^2 \leq n\delta^2$, then the design $\mathbf{X}$ satisfies $\Phi_{2k,-}(\mathbf{X}) \leq 2n\delta^2$. A necessary condition for $\mathbf{X}$ to satisfy $\Phi_{2k,-}(\mathbf{X}) \geq 2n\delta^2$ is therefore that the vectors $\mathbf{X}\theta_m$ are $\sqrt{n}\delta$-separated.

If $\mathbf{X}$ satisfies $\Phi_{2k,-}(\mathbf{X}) \geq 2n\delta^2$, then the balls in $\mathbb{R}^n$ with radius $\sqrt{n}\delta$ centered at $\mathbf{X}\theta_m$ are all disjoint. Thus, the sum of their volumes, is smaller than the volume of a ball a radius $\sqrt{n}(1+\delta)$ in $\mathbb{R}^n$. This implies that $\delta \leq 2(k/ep)^{Ck/n}$. Hence, for any design $\mathbf{X}$ with unit columns, we have

$$\Phi_{2k,-}(\mathbf{X}) \leq C_1 \left(\frac{k}{ep}\right)^{C_2 k/n} ,$$

which allows to prove the second result. □

Proof of the third result. The minimax lower bound is direct consequence of (6.2) and (6.4). In order to finish the proof, we shall combine the minimax upper bound (6.1) with an upper bound of $\inf_{\mathbf{X}\in\mathcal{D}_{n,p}} \Phi_{2k,-}^{-1}(\mathbf{X})$. Consider a standard Gaussian design $\mathbf{X}$ with size $n \times p$. Applying the deviation inequality (11.3) of Lemma 11.2, we derive that with probability larger than $1 - 1/p$, we have

$$\Phi_{2k,-}^{-1}(\mathbf{X}) \leq nC_1 \left(\frac{p}{k}\right)^{C_2 k/n} \left[\frac{k}{n}\log\left(\frac{p}{k}\right) \vee 1\right] \ .$$

However, the design $\mathbf{X}$ does not belong to $\mathcal{D}_{n,p}$. This is why we consider $\mathbf{X}' = \mathbf{X}D^{-1}$, where D is a diagonal matrix of size p, whose l-th diagonal element corresponds to the norm of the l-th column of $\mathbf{X}/\sqrt{n}$. Obviously, $\mathbf{X}'$ belongs to $\mathcal{D}_{n,p}$.

$$\Phi_{2k,-}(\mathbf{X}') = \inf_{\theta\in\Theta[k,p]} \frac{\|\mathbf{X}'\theta\|_n^2}{\|\theta\|_p^2} = \inf_{\theta\in\Theta[k,p]} \frac{\|\mathbf{X}\theta\|_n^2}{\|D\theta\|_p^2} \geq \frac{\Phi_{2k,-}(\mathbf{X})}{\varphi_{\max}(D^2)} ,$$

Each diagonal element of nD^2 follows of χ^2 distribution with n degrees of freedom. Applying Lemma 11.1, we derive that $\varphi_{\max}(D) \leq C\sqrt{1 \vee \log(p)/n}$ with probability larger than $1-1/p$. We conclude that

$$\begin{aligned}\Phi_{2k,-}^{-1}(\mathbf{X}') &\leq C_1 n\left(\frac{p}{k}\right)^{C_2k/n)}\left[\frac{k}{n}\log\left(\frac{p}{k}\right)\vee 1\right]\left[1\vee\frac{\log(p)}{n}\right]\\ &\leq C_1' n\left(\frac{p}{k}\right)^{C_2'k/n)}\ .\end{aligned}$$

with probability larger than $1-2/p$. This allows to conclude. □

9.9. Proof of Proposition 6.6

For the sake of simplicity, we assume that $\sigma^2=1$. Consider a design $\mathbf{X}\in\mathcal{D}_{n,p}$. By the proof of Proposition 6.2, there exist two vectors θ_1 and θ_2 such that:

1. θ_1 and θ_2 contain exactly k non-zero components which are all equal to $1/\sqrt{k}$ in absolute value.
2. The Hamming distance between θ_1 and θ_2 is larger than $k/2$.
3. $\|\mathbf{X}(\theta_1-\theta_2)\|_n^2\leq C_1 n\exp\left[-C_2k/n\log(ep/k)\right]:=\rho^{*-2}$.

Let us set $\theta_1^*=C\rho^*\theta_1$ and $\theta_2^*=C\rho^*\theta_2$ with $C=4\log(2)e/(2e+1)$. Consequently, the Kullback discrepancy between $\mathbb{P}_{\theta_1^*}$ and $\mathbb{P}_{\theta_2^*}$ is smaller than $\log(2)2e/(2e+1)$. Consider an estimator $\widehat{\theta}$ taking its values in $\{\theta_1^*,\theta_2^*\}$. Applying Corollary 2.18 in [37] (which is another version of Fano's Lemma), we derive that $\inf_{\theta_0\in\{\theta_1^*,\theta_2^*\}}\mathbb{P}_{\theta_0,1}(\widehat{\theta}=\theta_0)\leq 2e/(2e+1)$. This allows to conclude.

9.10. Proof of Proposition 6.7

For the sake of simplicity, we assume that $\sigma^2=1$ and that p is even. Consider any estimator $\widehat{M}$ of size p_0. We set

$$\rho^2=\frac{C_1}{2}\frac{k}{n}\log(p)\exp\left[\frac{C_2}{2}\frac{k}{n}\log(p)\right] \tag{9.20}$$

where the constants C_1, C_2 correspond to the ones used at the end of the proof of Proposition 5.1. We also consider the set $\mathcal{C}_k^p(\rho)$. Suppose that we have

$$\sup_{\theta_0\in\mathcal{C}_k^p(\rho)}\mathbb{P}_{\theta_0,1}[\mathrm{supp}(\theta_0)\subset\widehat{M}]\geq 7/8\ . \tag{9.21}$$

Assume we are given a second n-sample of (Y,X) independent of the first one. We note $(\mathbf{Y}',\mathbf{X}')$ this new sample. We consider the estimator $\widetilde{\theta}_k$ defined by

$$\widetilde{\theta}_k:=\arg\min_{\theta\in\Theta[k,p]\text{ and }\mathrm{supp}(\theta)\subset\widehat{M}}\|\mathbf{Y}'-\mathbf{X}'\theta\|_n^2\ .$$

Since $\Sigma=I_p$, all the covariates that do not lie in the support of θ_0 play a symmetric role in the distribution of $(\mathbf{Y},\mathbf{X})$. This estimator $\tilde{\theta}_k$ has the same form as the estimator $\widehat{\theta}_k$ introduced in (10.5). Arguing as in the proof of Theorem 5.2, we derive that

$$\|\widetilde{\theta}_k - \theta_0\|_p^2 \mathbf{1}_{\mathrm{supp}(\theta_0)\subset \widehat{M}} \leq C_1' k \log\left(\frac{ep_0}{k}\right) \exp\left[C_2' \frac{k}{n} \log\left(\frac{ep_0}{k}\right)\right] \ ,$$

with probability larger than 7/8. Gathering this bound with (9.21), we derive that for any $\theta_0 \in \mathcal{C}_k^p(\rho)$, we have

$$\|\widehat{\theta}_k - \theta_0\|_p^2 \leq C_1' \frac{k}{n} \log\left(\frac{ep_0}{k}\right) \exp\left[C_2' \frac{k}{n} \log\left(\frac{ep_0}{k}\right)\right] , \tag{9.22}$$

with probability larger than 3/4.

We shall prove that (9.22) is impossible if p_0 is too large. Let us split the p covariates into two groups M_1 and M_2. We consider the subsets $\mathcal{C}_{k,1}^p(\rho)$ (resp. $\mathcal{C}_{k,2}^p(\rho)$) of $\mathcal{C}_k^p(\rho)$ whose elements have their support in M_1 (resp. M_2). Arguing as in (9.15) and (9.16), we derive that for any estimator $\widehat{\theta}$, there exists $\theta_0 \in \mathcal{C}_{k,1}^p(\rho) \cup \mathcal{C}_{k,2}^p(\rho)$ such that

$$\|\widehat{\theta} - \theta_0\|_p^2 \geq \frac{\rho^2}{4} = \frac{C_1}{8} \frac{k}{n} \log(p) \exp\left[\frac{C_2}{2} \frac{k}{n} \log(p)\right] \ ,$$

with probability larger than 1/4. Here, the constants C_1 and C_2 are the same as in (9.20).

The last lower bound contradicts (9.22) is $\log(p_0)/\log(p) \leq \delta$, where $\delta > 0$ depends on the relative values of C_1, C_2, C_1', and C_2' in (9.20) and (9.22).

10. Procedures involved in the proofs of the minimax upper bounds

10.1. Testing procedures

10.1.1. Known variance: Test T_α^*

In order to establish the minimax upper bounds for known variance, we consider the following testing procedure. It is taken from Baraud [5] who applies it in the Gaussian sequence model. In the sequel, $\bar{\chi}_k(u)$ denotes the probability for a χ^2 distribution with k degrees of freedom to be larger than u. Given a subset m of $\{1, \dots, p\}$, Π_m refers to the orthogonal projection onto the space generated by the vectors $(\mathbf{X}_i)_{i\in m}$.

Definition 10.1 (Procedure T_α^*)**.** Define k^* as the smallest integer such that $k^*[1 + \log(p/k^*)] \geq \sqrt{n}$. For any $1 \leq k < k^*$, we define the statistics $T_{\alpha,k}^*$ by

$$T_{\alpha,k}^* := \sup_{m\in\mathcal{M}(k,p)} \|\Pi_m \mathbf{Y}\|_n^2 - \sigma^2 \bar{\chi}_k^{-1}\left[\alpha/\tbinom{p}{k}\right] \ ,$$

where $\mathcal{M}(k,p)$ is defined in Section 2. We also consider

$$T_{\alpha,n}^* := \|\mathbf{Y}\|_n^2 - \sigma^2 \bar{\chi}_n^{-1}(\alpha) \ .$$

The procedure T^*_α is defined by

$$T^*_\alpha = \left[\vee_{1\leq k<k^*} T^*_{\alpha/(2k^*),k}\right] \vee T^*_{\alpha/2,n} \ . \tag{10.1}$$

The hypothesis $\mathbf{H_0}$ is rejected if T^*_α is positive.

$T^*_{\alpha,k}$ corresponds to a Bonferroni multiple testing procedure based on a large number of parametric tests of the hypothesis $\mathbf{H_0}$: $\{\theta_0 = 0_p\}$ against $\mathbf{H_{1,m}}$: $\{\theta_0 \neq 0 \text{ and } \mathrm{supp}(\theta_0) \subset m\}$ for any $m \in \mathcal{M}(k,p)$. As a consequence, $T^*_{\alpha,k}$ allows to test the hypothesis $\mathbf{H_0}$:$\{\theta_0 = 0\}$ against $\mathbf{H_{1,k}}$: $\{\theta_0 \in \Theta[k,p]\setminus\{0_p\}\}$. Then, T^*_α corresponds to a Bonferroni multiple testing procedures based on the statistics $T^*_{\alpha,k}$, $k \in \{1,\ldots k^*\} \cup \{n\}$. Obviously, the procedure T^*_α is computationally intensive. It is used here as a theoretical tool to derive minimax upper bounds.

10.1.2. Unknown variance: test T_α

We introduce a second testing procedure to handle the case of unknown variance σ^2.

Definition 10.2 (Procedure T_α)**.** Fixing some subset m of $\{1,\ldots,p\}$ such that $n - |m| > 0$, we note $d_m(\mathbf{X})$ the rank of the subdesign $\mathbf{X}_m$ of $\mathbf{X}$ of size $n \times |m|$. We define the Fisher statistic ϕ_m by

$$\phi_m(\mathbf{Y},\mathbf{X}) := \frac{[n - d_m(\mathbf{X})]\|\Pi_m \mathbf{Y}\|_n^2}{d_m(\mathbf{X})\|\mathbf{Y} - \Pi_m\mathbf{Y}\|_n^2} \ . \tag{10.2}$$

We build the statistic $T_{\alpha,k}(\mathbf{Y},\mathbf{X})$ as

$$T_{\alpha,k} := \sup_{m\in\mathcal{M}(k,p)} \phi_m(\mathbf{Y},\mathbf{X}) - \bar{F}^{-1}_{d_m(\mathbf{X}),n-d_m(\mathbf{X})}\left[\alpha/\tbinom{p}{k}\right] , \tag{10.3}$$

where $\bar{F}_{k,n-k}(u)$ denotes the probability for a Fisher variable with k and $n-k$ degrees of freedom to be larger than u. Finally, the statistic T_α is defined by

$$T_\alpha := \sup_{k=1,\ldots,\lfloor n/2\rfloor} T_{\alpha/\lfloor n/2\rfloor,k} \ . \tag{10.4}$$

The hypothesis $\mathbf{H_0}$ is rejected when T_α is positive.

In fact, T_α is a Bonferroni multiple testing procedure. Contrary to T^*_α, it is based on Fisher statistics to handle the unknown variance. The ideas underlying this statistic have been introduced in Baraud et al. [7] in the context of fixed design regression.

10.2. Estimation procedures

10.2.1. Definition of the estimator $\widetilde{\theta}^V$

Definition 10.3 (Estimator $\widetilde{\theta}^V$)**.** For any integer $k \in \{1,\ldots,p\}$, we consider a least-squares estimator $\widehat{\theta}_k$ defined by

$$\widehat{\theta}_k \in \arg\min_{\theta\in\Theta[k,p]} \|\mathbf{Y} - \mathbf{X}\theta\|_n^2 \ . \tag{10.5}$$

Let us define the penalty function pen : $\{1,\dots,\lfloor (n-1)/4\rfloor\} \mapsto \mathbb{R}^+$ by

$$\mathrm{pen}(k) = K\frac{k}{n}\log\left(\frac{ep}{k}\right)\ , \tag{10.6}$$

where $K > 0$ is a tuning parameter. The dimension $\widehat{k}^V$ is selected as follows

$$\widehat{k}^V \in \arg\min_{1\leq k\leq \lfloor (n-1)/4\rfloor} \log\left[\|\mathbf{Y}-\mathbf{X}\widehat{\theta}_k\|_n^2\right] + \mathrm{pen}(k)\ .$$

For short, we note $\widetilde{\theta}^V = \widehat{\theta}_{\widehat{k}^V}$.

This variable selection procedure relies on complexity penalization. The penalty pen(k) depends on the size of k and on the number $\binom{p}{k}$ of subsets of $\{1,\dots,p\}$ of size k. Observe that the estimator $\widetilde{\theta}^V$ does not require the knowledge of σ^2.

The choice of the tuning parameter K is universal: it neither depends on n, p, k, nor on Σ, θ_0, σ^2. It is only constrained to be larger than a positive numerical constant so that the equations (B.8), (B.24), (B.26), (B.31), and (B.34) in the proofs of Theorem 5.2, Propositions 5.5 and 6.3 in [43] hold.

10.2.2. Definition of the estimator $\widetilde{\theta}^{BM}$ and proof of (5.6) in Proposition 5.3

Definition 10.4 (Procedure for fixed design regression)**.** Define k^* as the smallest integer k such that $k[1+\log(p/k)] \geq n$. Let us consider the collection of dimensions $\mathcal{K} := \{1,\dots,k^*\}\cup\{n\}$. Then, the penalty function pen : $\mathcal{K} \mapsto \mathbb{R}^+$ is defined by

$$\mathrm{pen}(k) := \begin{cases} 4k\left[4+\log\left(\frac{p}{k}\right)\right] & \text{if} \quad k\leq k^* \\ 2n & \text{if} \quad k = n\ , \end{cases}$$

We recall that for $k\leq k^*$, the estimators $\widehat{\theta}_k$ are defined in (10.5) and that $\widehat{\theta}_n \in \arg\min_{\theta\in\mathbb{R}^p}\|\mathbf{Y}-\mathbf{X}\theta\|_n^2$. The size $\widehat{k}^{BM}$ is selected by minimizing the following penalized criterion

$$\widehat{k}^{BM} := \arg\inf_{k\in\{1,\dots k^*\}\cup\{n\}} \|\mathbf{Y}-\mathbf{X}\widehat{\theta}_k\|_n^2 + \sigma^2\mathrm{pen}(k)\ , \tag{10.7}$$

For short, we write $\widetilde{\theta}^{BM} = \widehat{\theta}_{\widehat{k}^{BM}}$.

Observe that the estimator $\widetilde{\theta}^{BM}$ requires the knowledge of the variance σ^2. Then, Eq. (5.6) is a special case of Theorem 1 in Birgé and Massart [12].

11. Deviation inequalities

The proofs of the deviation inequalities stated in this section are postponed to Appendix C in [43].

Lemma 11.1 (χ^2 distributions)**.** *For any integer $d > 0$ and any number $0 < x < 1$,*

$$\begin{aligned} \mathbb{P}\left(\chi^2(d) \geq d + 2\sqrt{d\log(1/x)} + 2\log(1/x)\right) &\leq x \ , \\ \mathbb{P}\left(\chi^2(d) \leq d - 2\sqrt{d\log(1/x)}\right) &\leq x \ . \end{aligned}$$

For any positive number $0 < x < 1$

$$\mathbb{P}\left[\chi^2(d) \leq dCx^{2/d}\right] \leq x \ , \qquad (11.1)$$

where the constant $C = \exp(-1)$.

Lemma 11.2 (Wishart distributions)**.** *Let Z^TZ be a standard Wishart matrix of parameters (n,d) with $n > d$. For any number $0 < x < 1$,*

$$\begin{aligned} \mathbb{P}\left[\varphi_{\max}\left(Z^TZ\right) \geq n\left(1 + \sqrt{d/n} + \sqrt{2\log(1/x)/n}\right)^2\right] &\leq x \ , \\ \mathbb{P}\left[\varphi_{\min}\left(Z^TZ\right) \leq n\left(1 - \sqrt{d/n} - \sqrt{2\log(1/x)/n}\right)_+^2\right] &\leq x \ . \qquad (11.2) \end{aligned}$$

For any (n,d) with $n \geq 4d+1$ and any number $0 < x < 1$,

$$\mathbb{P}\left[\varphi_{\min}\left(Z^TZ\right) \leq nCx^{\frac{2}{n-2d}}\left[1 \vee \frac{\log(2/x)}{n}\right]^{-1}\right] \leq x \ , \qquad (11.3)$$

where C is a numerical constant.

The two first deviation inequalities are taken from Theorem 2.13 in [19]. The bound (11.3) allows to control the tail distribution of the smallest eigenvalue of a Wishart distribution. Rudelson and Vershynin [41] have provided a control similar to (11.3) under subgaussian assumptions. However, their results only holds for events of probability smaller than $1 - e^{-n}$.

Acknowledgements

I am grateful to Yannick Baraud and Christophe Giraud, the Associate Editor, and two anonymous referees for suggestions that greatly improve the presentation of the paper.

Supplementary Material

Technical Appendix to "Minimax risks for sparse regressions: Ultra-high dimensional phenomenons"
(doi: 10.1214/12-EJS666SUPP; .pdf).

References

[1] Abramovich, F. and Grinshtein, V. (2010). MAP model selection in Gaussian regression. *Electron. J. Stat.* **4**, 932–949. http://dx.doi.org/10.1214/10-EJS573. MR2721039 (2011j:62028)

[2] Aldous, D. J. (1985). *Exchangeability and related topics*, École d'été de probabilités de Saint Flour XIII. Lecture Notes in Mathematics, Vol. **1117**. Springer-Verlag, Berlin.

[3] Arias-Castro, E., Candès, E. J., and Plan, Y. (2010). Global testing and sparse alternatives: Anova, multiple comparisons and the higher criticism. arXiv:1007.1434.

[4] Baraniuk, R., Davenport, M., DeVore, R., and Wakin, M. (2008). A simple proof of the restricted isometry property for random matrices. *Constr. Approx.* **28**, 3, 253–263. http://dx.doi.org/10.1007/s00365-007-9003-x. MR2453366

[5] Baraud, Y. (2002). Non-asymptotic rates of testing in signal detection. *Bernoulli* **8**, 5, 577–606.

[6] Baraud, Y., Giraud, C., and Huet, S. (2009). Gaussian model selection with an unknown variance. *Ann. Statist.* **37**, 2, 630–672.

[7] Baraud, Y., Huet, S., and Laurent, B. (2003). Adaptive tests of linear hypotheses by model selection. *Ann. Statist.* **31**, 1, 225–251. MR1962505 (2004a:62091)

[8] Bickel, P., Ritov, Y., and Tsybakov, A. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37**, 4, 1705–1732. http://dx.doi.org/10.1214/08-AOS620. MR2533469

[9] Birgé, L. (2001). An alternative point of view on Lepski's method. In *State of the art in probability and statistics (Leiden, 1999)*. IMS Lecture Notes Monogr. Ser., Vol. **36**. Inst. Math. Statist., Beachwood, OH, 113–133. http://dx.doi.org/10.1214/lnms/1215090065. MR1836557 (2002j:62049)

[10] Birgé, L. (2005). A new lower bound for multiple hypothesis testing. *IEEE Trans. Inform. Theory* **51**, 4, 1611–1615. MR2241522 (2007b:62024)

[11] Birgé, L. and Massart, P. (2001). Gaussian model selection. *J. Eur. Math. Soc. (JEMS)* **3**, 3, 203–268. MR1848946 (2002i:62072)

[12] Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields* **138**, 1-2, 33–73. http://dx.doi.org/10.1007/s00440-006-0011-8. MR2288064 (2008g:62070)

[13] Candès, E. and Plan, Y. (2009). Near-ideal model selection by ℓ_1 minimization. *Ann. Statist.* **37**, 5A, 2145–2177. http://dx.doi.org/10.1214/08-AOS653. MR2543688

[14] Candes, E. J. and Tao, T. (2005). Decoding by linear programming. *IEEE Trans. Inform. Theory* **51**, 12, 4203–4215. http://dx.doi.org/10.1109/TIT.2005.858979. MR2243152 (2007b:94313)

[15] Candes, E. J. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.* **35**, 6, 2313–2351. MR2382644

[16] Chu, J.-H., Weiss, S., Carey, V., and Rabyl, B. (2009). A graphical model approach for inferring large-scale networks integrating gene expression and genetic polymorphism. *BMC Systems Biology* **3**, 55.

[17] Cook, R. D. and Li, B. (2002). Dimension reduction for conditional mean in regression. *Ann. Statist.* **30**, 2, 455–474. http://dx.doi.org/10.1214/aos/1021379861. MR1902895 (2003c:62087)

[18] Dalalyan, A. and Tsybakov, A. (2008). Aggregation by exponential weighting, sharp oracle inequalities and sparsity. *Machine Learning* **72**, 1-2, 39–61.

[19] Davidson, K. R. and Szarek, S. J. (2001). Local operator theory, random matrices and Banach spaces. In *Handbook of the geometry of Banach spaces, Vol. I*. North-Holland, Amsterdam, 317–366. MR1863696 (2004f:47002a)

[20] Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32**, 3, 962–994. http://dx.doi.org/10.1214/009053604000000265. MR2065195 (2005e:62066)

[21] Donoho, D. and Johnstone, I. (1994). Minimax risk over l_p-balls for l_q-error. *Probab. Theory Related Fields* **99**, 2, 277–303. http://dx.doi.org/10.1007/BF01199026. MR1278886 (95g:62019)

[22] Donoho, D. and Tanner, J. (2009). Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **367**, 1906, 4273–4293. With electronic supplementary materials available online, http://dx.doi.org/10.1098/rsta.2009.0152. MR2546388 (2010k:62407)

[23] Donoho, D. L. (2006). Compressed sensing. *IEEE Trans. Inform. Theory* **52**, 4, 1289–1306. http://dx.doi.org/10.1109/TIT.2006.871582. MR2241189 (2007e:94013)

[24] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* **32**, 2, 407–499. With discussion, and a rejoinder by the authors. MR2060166 (2005d:62116)

[25] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Statist. Soc. B* **70**, 5, 849–911.

[26] Fukumizu, K., Bach, F., and Jordan, M. I. (2009). Kernel dimension reduction in regression. *Ann. Statist.* **37**, 4, 1871–1905. http://dx.doi.org/10.1214/08-AOS637. MR2533474

[27] Giraud, C. (2008). Estimation of Gaussian graphs by model selection. *Electron. J. Stat.* **2**, 542–563.

[28] Ingster, Y. (1993a). Asymptotically minimax hypothesis testing for nonparametric alternatives I. *Math. Methods Statist.* **2**, 85–114.

[29] Ingster, Y. (1993b). Asymptotically minimax hypothesis testing for nonparametric alternatives II. *Math. Methods Statist.* **3**, 171–189.

[30] Ingster, Y. (1993c). Asymptotically minimax hypothesis testing for nonparametric alternatives III. *Math. Methods Statist.* **4**, 249–268.

[31] Ingster, Y. (1998). Minimax detection of a signal for l^n-balls. *Math. Methods Statist.* **7**, 4, 401–428 (1999). MR1680087 (2000f:62012)

[32] Ingster, Y. (2001). Adaptive detection of a signal of growing dimension I. *Math. Methods Statist.* **10**, 4, 395–421.

[33] Ingster, Y. (2002). Adaptive detection of a signal of growing dimension II. *Math. Methods Statist.* **11**, 1, 37–68.

[34] Ingster, Y., Tsybakov, A., and Verzelen, N. (2010). Detection boundary in sparse regression. *Electron. J. Stat.* **4**, 1476–1526.

[35] Johnstone, I. (1994). On minimax estimation of a sparse normal mean vector. *Ann. Statist.* **22**, 1, 271–289. http://dx.doi.org/10.1214/aos/1176325368. MR1272083 (95g:62020)

[36] Lounici, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electron. J. Stat.* **2**, 90–102. http://dx.doi.org/10.1214/08-EJS177. MR2386087 (2009a:62287)

[37] Massart, P. (2007). *Concentration inequalities and model selection*. Lecture Notes in Mathematics, Vol. **1896**. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard. MR2319879 (2010a:62008)

[38] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34**, 3, 1436–1462. MR2278363 (2008b:62044)

[39] Raskutti, G., Wainwright, M., and Yu, B. (2009). Minimax rates of estimations for high-dimensional regression over l_q balls. Tech. rep., UC Berkeley.

[40] Rigollet, P. and Tsybakov, A. (2011). Exponential screening and optimal rates of sparse estimation. *Ann. Statist.* **39**, 2, 731–771.

[41] Rudelson, M. and Vershynin, R. (2009). Smallest singular value of a random rectangular matrix. *Comm. Pure Appl. Math.* **62**, 12, 1707–1739. http://dx.doi.org/10.1002/cpa.20294. MR2569075

[42] Verzelen, N. (2010a). High-dimensional gaussian model selection on a gaussian design. *Ann. Inst. H. Poincaré Probab. Statist.* **46**, 2, 480–524.

[43] Verzelen, N. (2010b). Technical Appendix to "Minimax risks for sparse regressions: Ultra-high-dimensional phenomenons.".

[44] Verzelen, N. and Villers, F. (2010). Goodness-of-fit tests for high-dimensional Gaussian linear models. *Ann. Statist.* **38**, 2, 704–752. MR2604699

[45] Wainwright, M. (2009a). Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Inform. Theory* **55**, 12, 5728–5741. http://dx.doi.org/10.1109/TIT.2009.2032816. MR2597190

[46] Wainwright, M. (2009b). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Trans. Inform. Theory* **55**, 5, 2183–2202.

[47] Ye, F. and Zhang, C.-H. (2010). Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *J. Mach. Learn. Res.* **11**, 3519–3540. MR2756192

[48] Yu, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*. Springer, New York, 423–435. MR1462963 (99c:62137)

[49] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7**, 2541–2563. MR2274449

[50] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67**, 2, 301–320. MR2137327