Minimax risks for sparse regressions: Ultra-high-dimensional phenomenons

Nicolas Verzelen

Minimax risks for sparse regressions: Ultra-high-dimensional phenomenons

Nicolas Verzelen*

Abstract: Consider the standard Gaussian linear regression model $\mathbf{Y} = \mathbf{X}\theta + \boldsymbol{\epsilon}$, 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 when p is much larger than n. In such a situation, a classical approach amounts to assume that θ is approximately sparse. This paper studies the minimax risks of estimation and testing over classes of k-sparse vectors θ. These bounds shed light on the limitations due to high-dimensionality. The results encompass the problem of prediction (estimation of $\mathbf{X}\theta$), the inverse problem (estimation of θ) and linear testing (testing $\mathbf{X}\theta = 0$). 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 a 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.

AMS 2000 subject classifications: Primary 62J05; secondary 62F35, 62C20.
Keywords and phrases: High-dimensional regression, sparse vectors, minimax risk, minimax hypothesis testing, dimension reduction, adaptive estimation, model selection.

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. A typical example is the study of l_1-based penalization methods for the estimation of linear regression models.

In order to assess the qualities of statistical procedures, we 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 arbitrarely large p or are there theoretical limitations when p becomes "too large"? The knowledge of such limitations may drive the research towards areas where computationally efficient procedures are shown to be suboptimal. Furthermore, these limitations tell us what kind of data analysis problems are too complex so that no statistical procedure is able to provide reasonnable results.

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

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

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

where the vector θ 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 such as compressed sensing [20], 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 belongs to a linear subspace of $\mathbb{R}^p$. Here, we focus on testing the null hypothesis $\mathbf{H_0}$: "$\theta = 0$". In Gaussian design, this is equivalent to testing whether Y is independent from X.
($\mathbf{P_2}$) : **Prediction**. We focus on predicting the expectation $\mathbb{E}[\mathbf{Y}]$ in fixed design and the conditional expectation $\mathbb{E}[Y|X]$ in Gaussian design.
($\mathbf{P_3}$) : **Inverse problem**. The primary interest lies in estimating θ itself and the corresponding loss function is the $\|\widehat{\theta} - \theta\|_p^2$, where $\|.\|_p$ is the l_2 norm in $\mathbb{R}^p$.
($\mathbf{P_4}$) : **Support estimation** aims at recovering the support of θ, that is the set of indices corresponding to non-zero coefficients. The easier problem of **dimension reduction** amounts to estimate a set $\widehat{M} \subset \{1, \dots p\}$ of "reasonable" size that contains the support of θ with high probability.

Many work have been devoted to these statistical questions in a high dimensional setting ($p > n$). For the problem of prediction ($\mathbf{P_2}$), procedures based on complexity penalization are proved to provide good risk bounds for known variance [9] and unknown variance [4] but are computationally inefficient. In contrast, l_1-based 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 $\mathbf{X}$ [6, 11]. Exponential weighted aggregation methods [16, 35] are another examples of fast and efficient methods. Other popular methods include the elastic net [43]. The l_1 penalization methods have also been analyzed for the inverse problem ($\mathbf{P_3}$) [6] and for support estimation ($\mathbf{P_4}$) [31, 42]. Dimension reduction methods are often studied in more general settings than linear regression [15, 23]. In the linear regression model, [22] have introduced the SIS method based on the correlation between the response and the covariate. The problem of high-dimensional hypothesis testing ($\mathbf{P_1}$) has attracted less attention yet. Some testing procedures are discussed in [5] for fixed design and in [38] for Gaussian design.

1.2. Sparsity and ultra-high-dimensionality

We are primary interested 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 is *sparse*, in the sense that most of the components of θ are equal that to 0. Given a positive integer k, we say that the vector θ is k-sparse if θ 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. Such assumptions are performed for the analysis of the Lasso in prediction, inverse problems [6], and support estimation [33]. The exponential screening method of [35] is analyzed under the assumption $\log(p) \leq n$. In dimension reduction, the SIS method of [22] also requires this assumption. If the multiple testing procedure of [5] can be analyzed for $k[1+\log(p/k)]$ larger than n, it exhibits a much slower rate of testing in this case. In the sequel, we say that the problem is ultra-high dimensional when $k[1+\log(p/k)]$ is large compared to n. We prove in this paper that a new phenomenon occurs in a ultra-high dimensional setting and that most of the estimation and testing problems become much more difficult. If we take an asymptotic point of view ($(k_n, p_n) \to \infty$), an the ultra-high dimensionality does not necessary imply that p_n is exponential with respect to n. As an example, taking $p_n = n^2$ and $k_n = n/\log\log(n)$ asymptotically yields an ultra-high dimensional problem.

The study of ultra-high dimensional problems is partly motivated by the following question: in some gene network inference problems [14], the number p of genes can be as large as 5000 while the number n of microarray experiments is only of the order 50. Let us consider a gene A. How large can be its degree k in the network so that it is still "reasonnable" to estimate the set of genes that interact with the gene A from the microarray experiments? In statistical terms, inferring the set of genes interacting with A amounts to estimate the support of θ in a linear regression model (1.1). 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 $\mathbf{X}$ (resp. a covariance Σ) and a variance σ^2 is defined by $\sup_{\theta\in\Theta[k,p]} \mathbb{E}_{\theta,\sigma}[l(\widehat{\theta},\theta)]$. Taking the infimum of the maximal risk over all possible estimators $\widehat{\theta}$, we obtain the *minimax risk*

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

We say that an estimator $\widehat{\theta}$ is good 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 and we seldom know the value σ^2 of $\mathrm{Var}(\epsilon)$. 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, adaption 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 [19, 28] provide the asymptotic minimax risks of prediction ($\mathbf{P_2}$). Baraud [3] studies the optimal rate of testing from a non-asymptotic point of view while Donoho and Jin [18] provide the asymptotic optimal rate of testing with exact constants.

Much less results exist for non-orthogonal designs and for p larger than n. Wainwright [39, 40] has provided minimax lower bounds for the problem of support estimation ($\mathbf{P_4}$). Some minimax lower bounds have also been stated for testing ($\mathbf{P_1}$) and prediction ($\mathbf{P_2}$) problems with Gaussian design [37, 38].

This paper provides a general study of the minimax risks for the problems ($\mathbf{P_1}$), ($\mathbf{P_2}$), ($\mathbf{P_3}$) when the regression vector θ is k-sparse. The main discoveries are the following:

1. **High-dimensional and ultra-high dimensional problems**. Our results cover both the high-dimensional and ultra-high dimensional setting. Previous work do not cover the ultra-high dimensional setting or do not exhibit its specificity. We establish that for each of the problems ($\mathbf{P_1}$), ($\mathbf{P_2}$) and ($\mathbf{P_3}$), an elbow effect occurs when $k[1 + \log(p/k)]$ becomes large compared to n. This has some consequence on support estimation ($\mathbf{P_4}$): in a ultra-high dimensional setting, it is impossible to recover the support of θ except if the signal to noise ratio is exponentially large with respect to $k \log(p)/n$. It even becomes almost impossible to reduce efficiently the dimension of the problem. This phenomenon is 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 the 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 a ultra-high dimensional setting.
3. **Effect of the design**. Lastly, the minimax bounds of ($\mathbf{P_1}$) and ($\mathbf{P_2}$) are established for fixed and Gaussian designs. Except for the problem of prediction ($\mathbf{P_2}$), the minimax risks are of similar nature for both forms of the design. Furthermore, we investigate the dependency of the minimax risks on the design $\mathbf{X}$ (resp. Σ).

The minimax bounds stated in this paper are non asymptotic. Most of them rely on Fano's and Le Cam's methods [41] and on geometric considerations. In each case, near optimal procedures are exhibited.

While we were writing this paper we became aware of the work of Raskutti et al. [34] and of Rigollet and Tsybakov [35]. Raskutti et al. provide minimax upper bounds and lower bounds for ($\mathbf{P_2}$) and ($\mathbf{P_3}$) over l_q balls for general fixed designs $\mathbf{X}$. In the specific case of $q = 0$, this corresponds to studying minimax risks over sparse vectors θ and their bounds agree with our results (Propositions 5.3 and 6.1). Rigollet and Tsybakov [35] also provide a minimax lower bound the problem of prediction ($\mathbf{P_2}$). Except for the case of degenerate designs $\mathbf{X}$, their bound agrees with Propositions 5.3. Contrary to our results, these two work do not encompass the case of ultra-high dimensionality and require that the variance σ^2 is known.

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 studies illustrates this answer. Section 8 contains the final discussion. Section 9 is devoted to the proof of the minimax lower bounds, while the last section contains the remaining proofs.

2. Notations and preliminaries

Gaussian design and conditional distribution. For a Gaussian design the rows of $\mathbf{X}$ correspond to a n-sample of a random vector $X \sim \mathcal{N}(0_p, \Sigma)$. Then, $(\mathbf{Y}, \mathbf{X})$ can be interpreted a n-sample of the random vector $(Y, X^*) \in \mathbb{R}^{p+1}$ defined by

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

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 $\mathbb{E}[Y|X] = X\theta$ and $\mathrm{Var}(Y|X) = \sigma^2$. 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 [33, 38].

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 $\theta \in \mathbb{R}^p$ and $\sigma > 0$, $\mathbb{P}_{\theta,\sigma}$ refers to the joint distribution of $(\mathbf{Y}, \mathbf{X})$. In the sequel, we note $\mathrm{supp}(\theta)$ for the support of θ. 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.

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 $\mathbf{X}$.*

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

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 $\mathbf{X}^*\mathbf{X}$.

Given a symmetric real square matrix A, $\varphi_{\max}(A)$ stands for the largest eigenvalue of A. Finally, C, C_1,... 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 results will be 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 these results 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

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 classical loss $\|\mathbf{X}(\theta_1-\theta_2)\|_n^2/n$. For Gaussian design, we consider the integrated prediction loss function:

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

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

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

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

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

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

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

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 $\mathbf{X}$. In the sequel, we say that $\mathcal{R}_F[k]$ is *of order* $f(k, p, n, C)$ 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.

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 known for a long time in the specific case of the Gaussian sequence model [28]. Some procedures based on complexity penalization [9, 4] are proved to achieve these risks uniformly over all designs $\mathbf{X}$. Computationally efficient procedures like the Lasso or the Dantzig selector only achieve a $k \log(p)/n$ risk under assumption on the design $\mathbf{X}$ [6]. If the support of θ is known in advance, the parametric risk is of order k/n.

Fixed Design	**Gaussian** Design
$C\frac{k\log(p/k)}{n} \wedge 1$	$C\frac{k\log(p/k)}{n}\exp\left[C\frac{k\log(p/k)}{n}\right]$

Table 1: Minimax risks of prediction $\mathcal{R}_F[k]$ and $\mathcal{R}_R[k]$ over $\Theta[k,p]$.

Thus, the price to pay for not knowing the support of θ is only logarithmic in p.

In a 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 $C(p/k)^{Ck/n}$: it increases exponentially with respect to k and polynomially 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 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. In short, 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.

3.2. Testing

Let us turn to the problem $(\mathbf{P_1})$ of testing $\mathbf{H_0}$: "$\theta = 0$" against $\mathbf{H_1}$: "$\theta \in \Theta[k,p] \setminus \{0\}$". We fix a level $\alpha > 0$ and and a type II error probability $\delta > 0$. Minimax lower and upper bounds for this loss function 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 $\|\mathbf{X}\theta\|_n/\sqrt{n} \geq \rho\sigma$. Hence, $\rho_F[\Phi_\alpha, k, \mathbf{X}]$ corresponds to the minimal distance such that the hypotheses "$\theta = 0$" and "$\|\mathbf{X}\theta\|_n^2 \geq n\rho_F^2[\Phi_\alpha, k, \mathbf{X}]\sigma^2$, $\theta \in \Theta[k,p]$" are well separated by the test Φ_α.

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

Then, we consider

$$\rho_F^*[k, \mathbf{X}] := \inf_{\Phi_\alpha} \rho[\Phi_\alpha, k, \mathbf{X}] . \tag{3.4}$$

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 are a non-asymptotic counterpart of the detection boundaries studied in the Gaussian sequence model [18].

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

$$\rho_R[\Phi_\alpha, k, \Sigma] := \inf\left\{\rho > 0, \inf_{\theta\in\Theta[k,p],\ \|\sqrt{\Sigma}\theta\|_p \geq \rho\sigma} \mathbb{P}_{\theta,\sigma}[\Phi_\alpha = 1] \geq 1-\delta\right\} , \quad \rho_R^*[k,\Sigma] := \inf_{\Phi_\alpha} \rho_R[\Phi_\alpha, k, \Sigma] \tag{3.5}$$

As for the problem of prediction ($\mathbf{P_2}$), we state the orders of the minimax separation distances in the "worst case" designs:

$$\rho_F^*[k] := \sup_{\mathbf{X}} \rho_F^*[k, \mathbf{X}] \ , \qquad \rho_R^*[k] := \sup_{\Sigma} \rho_R^*[k, \Sigma] \ . \tag{3.6}$$

This is the smallest separation distance that can be achieved by a procedure Φ_α *uniformly* over all designs $\mathbf{X}$ (resp. Σ). Contrary to the problem of prediction ($\mathbf{P_2}$), 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 "$\theta = 0$" and "$\|\mathbf{X}\theta\|_n^2 \geq n\rho^2\sigma^2$" when σ is unknown. We shall provide a formal definition at the beginning of Section 4.

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.

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

Table 2: 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 $(\rho_{R,U}^*[k])^2$.

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

When $k\log(p/k) \leq \sqrt{n}$, all the minimax separation distances are of order $k\log(ep/k)/n$. 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 [3, 18].

When $k\log(p/k) \geq \sqrt{n}$, 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 "$\mathbb{E}[\mathbf{Y}] = 0$" against the general hypothesis "$\mathbb{E}[\mathbf{Y}] \neq 0$" for known variance (see Baraud [3]).

If the variance is unknown, the minimax separation distance over $\Theta[k, p]$ is still of order $k\log(ep/k)/n$ if $k\log(p/k)$ is small compared to n. Moreover, the minimax separation distance blows up to the order $C(p/k)^{Ck/n}$ 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 $k\log(p/k)$ larger than $\sqrt{n}$.

3.3. Inverse problem and support estimation

In the inverse problem ($\mathbf{P_3}$), we are primarily interested in the estimation of θ rather than $\mathbf{X}\theta$. 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 prediction is

$$\mathcal{RI}[k, \mathbf{X}] := \inf_{\widehat{\theta}} \sup_{\theta\in\Theta[k,p]} \mathbb{E}_{\theta,\sigma}[\|\theta - \widehat{\theta}\|_p^2/\sigma^2] \ . \tag{3.7}$$

If one transforms the design $\mathbf{X}$ by an homothety of factor $\lambda > 0$, then this multiplies the minimax risk of the inverse problem by a factor $1/\lambda^2$. For the sake of simplicity, we restrict ourselves to designs $\mathbf{X}$ such that each columns has a unit norm. 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. We are rather interested in the infimum of the minimax risks over $\Theta[k,p]$ as $\mathbf{X}$ varies across $\mathcal{D}_{n,p}$:

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

The quantity $\mathcal{RI}[k]$ is interpreted the following way: given (k,n,p) what is the smallest risk we can hope if we use the easiest possible design? We call this quantity the minimax risks of the inverse problem over $\Theta[k,p]$. In Table 3, we provide the minimax risks in fixed design for different values of (k,n,p).

$(\mathbf{k},\mathbf{n},\mathbf{p})$	$k\log(p/k) \le Cn$	$k\log(p/k) \gg n\log(n)$
Minimax risk $\mathcal{RI}[k]$	$Ck\log(p/k)$	$\exp[C'k\log(p/k)/n]$.

Table 3: Minimax risks of the inverse problem $\mathcal{RI}[k]$ over $\Theta[k,p]$

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 [13] for nearly-orthogonal designs, that roughly means that the restricted eigenvalues $\Phi_{3k,+}(\mathbf{X})$ and $\Phi_{3k,-}(\mathbf{X})$ of $\mathbf{X}^*\mathbf{X}$ are close to one. In a 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 problems of prediction in Gaussian design $(\mathbf{P_2})$.

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 a ultra-high dimensional setting, it is not possible to estimate with high probability the support of θ unless the ratio $\|\theta\|_p^2/\sigma^2$ is larger than $C(p/k)^{2k/n}$. Moreover, it is not possible to select of subset of $\{1,\dots,p\}$ of size n that contains the support of θ unless the ratio $\|\theta\|_p^2/\sigma^2$ is larger than $(p/k)^{Ck/n}$. Observe that the quantity $(p/k)^{Ck/n}$ is precisely huge in a ultra-high dimensional setting. In practice, this means that the problems of support estimation and dimension reduction are almost hopeless in a 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. Test T_α^*

In order to obtain the minimax upper bounds for known variance, we consider the following testing procedure. It is taken from Baraud [3] who applies it in the Gaussian sequence model.

Definition 4.1. [**Procedure** T^*_α] *Define k^* as the smallest integer such that $k^*[1+\log(p/k^*)] \geq \sqrt{n}$. Given a subset m of $\{1,\dots,p\}$, Π_m refers to the orthogonal projection onto the space generated by the vectors $(\mathbf{X}_i)_{i\in m}$. For any $1 \leq k < k^*$, we define the statistics $T^*_{\alpha,k}$ by*

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

where $\mathcal{M}(k,p)$ is defined in Section 2 and $\bar{\chi}_k(u)$ denotes the probability for a χ^2 distribution to be larger than u. We also consider

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

The procedure T^*_α is defined by

$$T^*_\alpha = \left[\vee_{1\leq k<k^*} T^*_{\alpha/(2k^*),k}\right] \vee T^*_{\alpha/2,n} . \tag{4.1}$$

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

T^*_α is a Bonferroni multiple testing procedure based on a large number of parametric tests of the hypothesis $\mathbf{H_0}$: "$\theta = 0$" against $\mathbf{H_{1,m}}$ "$\theta \neq 0$ and $\mathrm{supp}(\theta) \subset m$".

4.1.2. 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 distance have been defined in Section 3.2.

Theorem 4.1. [**Minimax lower bounds**] *Assume that $\alpha + \delta \leq 53\%$ and that $p \geq n$. For any $k \in \{1,\dots n\}$, the (α,δ)-minimax separation distance (3.5) with covariance I_p is lower bounded by*

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

This lower bound also implies that $(\rho^*_R[k])^2 \geq C(\alpha,\delta)\left[\frac{k}{n}\log\left(1 + \frac{p}{k^2} \vee \sqrt{\frac{p}{k^2}}\right) \wedge \frac{1}{\sqrt{n}}\right]$. Next, we state that T^*_α achieves this rate of testing for any covariance Σ.

Proposition 4.2. [**Power of** T^*_α] *For any covariance Σ, the size of the procedure T^*_α is smaller than α. Consider some $\delta > 0$ and assume that $n \geq 8\log(2/\delta)$. For any $k \in \{1,\dots,p\}$ and any covariance Σ, we have*

$$\rho^2_R[T^*_\alpha, k, \Sigma] \leq C(\alpha,\delta)\left[\frac{k}{n}\log\left(\frac{ep}{k}\right) \wedge \frac{1}{\sqrt{n}}\right] . \tag{4.3}$$

Remark 4.1. [**Minimax adaptation**]

1. *If $p \geq n^{1+\gamma}$ with $\gamma > 0$, then the procedure T_α^* defined in (4.1) simultaneously achieves up to a constant $C(\alpha, \delta, \gamma)$ the optimal separation distance*

$$\frac{k}{n} \log\left(\frac{ep}{k}\right) \wedge \frac{1}{\sqrt{n}} \ ,$$

for all sparsities k between 1 and n. The separation distance of T_α^* proved in Proposition 4.6 is valid for any covariance matrix Σ of the vector X. In contrast, the minimax lower bound of Theorem 4.5 is restricted to the case $\Sigma = I_p$. This implies that the minimax separation distance for a general matrix Σ is (up to a positive constant that does not depend on Σ) smaller than the minimax separation distance for $\Sigma = I_p$. In other words, there exists a positive constant $C(\alpha, \delta)$ such that for all covariance matrices Σ,

$$\rho_R^*[k, I_p] \geq C(\alpha, \delta) \rho_R^*[k, \Sigma] \ .$$

2. *When p is close to n and k is close to $\sqrt{n}$, the minimax lower bound (4.2) and the upper bound (4.3) only match up to a possible $\log(n)$ factor. Such a difficulty has already been observed by Baraud [3] in the case of Gaussian sequence model which corresponds to $p = n$ and a fixed design $\mathbf{X} = I_p$.*

4.1.3. Fixed design

The separation distances share similar behaviors with the Gaussian design case.

Theorem 4.3. [Minimax lower bound] *Assume that $\alpha + \delta \leq 53\%$ and that $p \geq n \geq C(\alpha, \delta)$. For any $k \in \{1, \ldots, n\}$, there exist some $n \times p$ designs $\mathbf{X}$ such that*

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

More specifically, we consider designs $\mathbf{X}$ that are realisations of a standard Gaussian design: all $\mathbf{X}_{i,j}$ follow independent standard normal distribution. Then, with large probability, the design $\mathbf{X}$ satisfies (4.4). See the proof for more details. Theorem 4.3 implies that

$$(\rho_F^*[k])^2 \geq C(\alpha, \delta) \left[\frac{k}{n} \log\left(1 + \frac{p}{k^2} \vee \sqrt{\frac{p}{k^2}}\right) \wedge \frac{1}{\sqrt{n}}\right] \ .$$

In order to get the minimax upper bounds, we still use the procedure T_α^*.

Proposition 4.4. [Power of T_α^* in fixed design] *For any design $\mathbf{X}$, the size of the procedure T_α is smaller than α. Consider some $\delta > 0$ and assume that $n \geq 8 \log(2/\delta)$. For any design $\mathbf{X}$ and any $k \in \{1, \ldots, n\}$, we have*

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

Remark 4.2. *The minimax lower bounds and the upper bounds are analogous to the random design case studied in Section 4.1.2 and the same comments apply. If $p \geq n^{1+\gamma}$ with $\gamma > 0$, T_α^* is minimax adaptive to the sparsity k. Moreover, Theorem 4.3 tells us that, with high probability, realisations $\mathbf{X}$ of a standard Gaussian design almost yield the largest minimax separation distance, that is $\rho_F^*[k, \mathbf{X}] \geq \rho_F^*[k]$.*

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

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

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 variance is defined by

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

Hence, $\rho_{F,U}[\Phi_\alpha, k, \mathbf{X}]$ corresponds to the minimal distance such that the hypotheses "$\theta = 0$" and "$\|\mathbf{X}\theta\|_n^2 \geq n\rho_{F,U}^2[\Phi_\alpha, k, \mathbf{X}]\sigma^2$, $\theta \in \Theta[k,p]$ and $\sigma > 0$" 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 $\|\mathbf{X}\theta\|_n/\sqrt{n}$ by $\|\sqrt{\Sigma}\theta\|_p$.

4.2.2. Test T_α

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

Definition 4.2. **[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*

$$\phi_m(\mathbf{Y}, \mathbf{X}) := \frac{[n - d_m(\mathbf{X})]\|\Pi_m \mathbf{Y}\|_n^2}{d_m(\mathbf{X})\|\mathbf{Y} - \Pi_m \mathbf{Y}\|_n^2} \ . \tag{4.8}$$

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

$$T_{\alpha,k} := \sup_{m \in \mathcal{M}(k,p)} \phi_m(\mathbf{Y}, \mathbf{X}) - \bar{F}^{-1}_{d_m(\mathbf{X}), n-d_m(\mathbf{X})}\left[\alpha / \tbinom{k}{p}\right] \ , \tag{4.9}$$

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

$$T_\alpha := \sup_{k=1,\dots,\lfloor n/2 \rfloor} T_{\alpha/\lfloor n/2 \rfloor, k} \ . \tag{4.10}$$

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

As T^*_α, the T_α is a a Bonferroni multiple testing procedure. Contrary to T^*_α, it is based on Fisher tests to handle the unknown variance. The ideas underlying this statistic have been introduced in [5] in the context of fixed design regression.

4.2.3. Gaussian design

Theorem 4.5. [Minimax lower bound] *Suppose that $\alpha + \delta \leq 53\%$ and that $p \geq C$. For any $k \in \{1, \dots, \lfloor p^{1/3} \rfloor\}$, the (α, δ)-minimax separation distance over $\Theta[k,p]$ with covariance I_p and unknown variance satisfies*

$$(\rho^*_U[k, I_p])^2 \geq C \frac{k}{n} \log\left(\frac{ep}{k}\right) \exp\left[C \frac{k}{n} \log\left(\frac{ep}{k}\right)\right] \ . \tag{4.11}$$

Remark 4.3. *This theorem extends a result of [38] to encompass both the high-dimensional and the ultra-high dimensional setting. The condition $p \leq k^{1/3}$ can be replaced by $k \leq p^{1/2-\gamma}$ with $\gamma > 0$. This condition is not really restrictive for a sparse high-dimensional regression since the usual setting is $k \leq n \ll p$.*

Proposition 4.6. [Power of T_α] *For any covariance Σ, the size of T_α is smaller than α. Consider some $\delta > 0$ and assume that $p \geq n \geq 8 \log(2/\delta)$. For all $1 \leq k \leq n/2$ and all covariance matrices Σ, we have*

$$\rho^2_{R,U}[T_\alpha, k, \Sigma] \leq C(\alpha, \delta) \frac{k \log(ep/k)}{n} \exp\left[C_2(\alpha, \delta) \frac{k \log(ep/k)}{n}\right] \ . \tag{4.12}$$

This proposition extends a result of [38] to encompass both the high-dimensional and the ultra-high dimensional setting.

Remark 4.4.

1. **[Minimax adaptation]** *If $k \leq p^{1/3} \wedge n/2$, the upper bound (4.12) agrees with the minimax lower bound (4.11). Consequently, the test T_α simultaneously achieves the optimal rate of testing over all $\Theta[k,p]$ with $k \leq n/2 \wedge p^{1/3}$. The minimax separation distance is of order $k \log(p)/n$ when $k \log(p/k)$ remains small compared to n. In a ultra-high dimensional setting it blows up to the order of $\exp[C(\alpha, \delta) k \log(p)/n]$.*

2. **[Dependent design]** *As for known conditional variance, the separation distance of T_α proved in Proposition 4.6 is valid for any Σ, while the minimax lower bound of Theorem 4.5 has only been proved for $\Sigma = I_p$. This implies that there exists a constant $C(\alpha, \delta)$ such that for all covariance matrices Σ,*

$$\rho^*_{R,U}[k, I_p] \geq C(\alpha, \delta) \rho^*_{R,U}[k, \Sigma] \ .$$

For some covariance matrices Σ, the minimax separation distance with covariance Σ is much smaller than $\rho^*_{R,U}[k, I_p]$. Next, we provide an example of such a matrix Σ.

Example 4.1. *Consider a covariance Σ_c such that $\Sigma_c[i,i]=1$ and $\Sigma_c[i,j]=c>0$ for any $i\neq j$. Let us introduce $X_{p+1}=\sum_{i=1}^p X_i/\sqrt{p}$. For $nc\geq C(\alpha,\delta)$ the test T'_α defined by*

$$T'_\alpha := \frac{(n-1)\|\Pi_{\{p+1\}}\mathbf{Y}\|_n^2}{\|\mathbf{Y}-\Pi_{\{p+1\}}\mathbf{Y}\|_n^2} - \bar{F}^{-1}_{k,n-k}(\alpha)\ ,$$

satisfies $\mathbb{P}_0\left(T'_\alpha>0\right)=\alpha$ and $\rho^2_{R,U}[T_\alpha,1,\Sigma_c]\leq C(\alpha,\delta)(cn)^{-1}$. When $c\geq 1/\log p$, this separation distance is minimax.

This example is derived from Propositions 8 and 9 in [38] and its proof is analogous. Observe that $\rho_{R,U}[T_\alpha,1,\Sigma_c]$ does not depend on p. Thus, for large p, we get $\rho^*_{R,U}[1,\Sigma_c]\ll\rho^*_{R,U}[1,I_p]$. We cannot easily generalize the computations of this example to other covariances. The computation of sharp minimax bounds that capture the dependency of $\rho^*_{R,U}[k,\Sigma]$ on Σ remains an open problem.

Remark 4.5. [Comparison between known and unknown variance] *There are three regimes depending on (k,p,n):*

1. $\mathbf{klog(p/k)\leq\sqrt{n}}$. *The minimax separation distances are of the same order for known and unknown σ^2. The minimax distance $k\log(p/k)/n$ is also of the same order as the minimax risk of prediction.*
2. $\mathbf{\sqrt{n}\leq klog(p/k)\leq n}$. *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 "$Var(Y)=\sigma^2$" against "$Var(Y)\neq\sigma^2$" using the statistic $T^*_{\alpha,n}$ introduced in Definition 4.1. If σ^2 is unknown, we cannot use the statistic $T^*_{\alpha,n}$ and the minimax separation distance behaves like $k\log(p/k)/n$.*
3. $\mathbf{klog(p/k)\geq n}$. *If σ^2 is unknown, the minimax separation distance blows up. It is of order $(p/k)^{Ck/n}$. Consequently, the problem of testing "$\theta=0$" becomes extremely difficult in this setting.*

4.2.4. Fixed design

Proposition 4.7. [Minimax lower bound] *Assume that $\alpha+\delta\leq 53\%$ and that $p\geq n\geq C(\alpha,\delta)$. For any $k\in\{1,\dots,\lfloor p^{1/3}\rfloor\}$, there exist some $n\times p$ designs $\mathbf{X}$ such that*

$$(\rho^*_{F,U}[k,\mathbf{X}])^2\geq C\frac{k}{n}\log\left(\frac{ep}{k}\right)\exp\left[C\frac{k}{n}\log\left(\frac{ep}{k}\right)\right]\ . \tag{4.13}$$

As for Theorem 4.5, the assumption $k\leq p^{1/3}$ can be replaced by $k\leq p^{1/2-\gamma}$ with $\gamma>0$. Proposition 4.7 implies that

$$(\rho^*_{F,U}[k,\mathbf{X}])^2\geq C\frac{k}{n}\log\left(\frac{ep}{k}\right)\exp\left[C\frac{k}{n}\log\left(\frac{ep}{k}\right)\right]\ .$$

To assess the optimality of the lower bound, we use the procedure T_α. The following result is a consequence of Theorem 1 in Baraud et al [5].

Proposition 4.8. [Power of T_α in fixed design] *Assume that $n\geq 4$. For any design $\mathbf{X}$, the size of T_α is less than α. Consider some $\delta>0$. For any $k\leq n/2$ and any $n\times p$ designs $\mathbf{X}$ we have*

$$\rho^2_{F,U}[T_\alpha,k,\mathbf{X}]\leq C(\alpha,\delta)\frac{k}{n}\log\left(\frac{ep}{k}\right)\exp\left[C_2(\alpha,\delta)\frac{k\log(ep/k)}{n}\right]\ . \tag{4.14}$$

Again, we observe an analogous phenomenon to the random design: the procedure T_α is minimax adaptive to the sparsity. Moreover, the minimax separation hypotheses grow exponentially fast with k in a ultra-high dimensional setting.

5. Prediction

In contrast to the testing problem, the minimax risks of predictions ($\mathbf{P_2}$) exhibit really different behaviors in fixed and in random design.

5.1. Gaussian design

Proposition 5.1. [Minimax lower bound for prediction] *Assume that $p \geq C$. For any $k \in \{1, \dots, \lfloor p^{1/3} \rfloor\}$, 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*

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

This statement has been proved in [37] (Proposition 4.5) in the special case of restricted isometry, but the proof straightforwardly extends to any restricted eigenvalue. 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)).

Next, we build an estimation procedure $\widetilde{\theta}^V$ that achieves the lower bound (5.1).

Definition 5.1. [Estimator $\widetilde{\theta}^V$] *For any integer $k \in \{1, \dots, p\}$, we consider a least-squares estimator $\widehat{\theta}_k$ defined by*

$$\widehat{\theta}_k \in \arg\min_{\theta \in \Theta'[k,p]} \|\mathbf{Y} - \mathbf{X}\theta'\|_n^2 . \tag{5.3}$$

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

$$\mathrm{pen}(k) = K \frac{k}{n} \log\left(\frac{ep}{k}\right) , \tag{5.4}$$

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 $\mathrm{pen}(k)$ 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, σ^2.

Theorem 5.2. **[Risk bound for $\widetilde{\theta}^V$]** *Assume that $n \geq C$. There exists a universal choice of K in the penalty (5.4) such that the following holds. For any covariance Σ ,any $k \in \{1, \dots, \lfloor (n-1)/4 \rfloor\}$ and any $\theta \in \Theta[k,p]$ we have*

$$\mathbb{E}\left[\|\sqrt{\Sigma}(\widetilde{\theta}^V - \theta)\|_p^2\right] \leq C(K)\frac{k}{n}\log\left(\frac{ep}{k}\right)\exp\left\{C\frac{k}{n}\log\left(\frac{ep}{k}\right)\right\}\sigma^2 \ . \tag{5.5}$$

In contrast to similar results such as Theorem 1 in Giraud [24] or Theorem 3.4 in Verzelen [37], we do not restrict the size of the models $|m|$ to be smaller than $n/(2\log p)$. The proof of the theorem is based on a new concentration inequality for the spectrum of Wishart matrices stated in Lemma A.2.

Remark 5.2.

1. **[Minimax risk]** *We derive from Theorem 5.2 and Proposition 5.1 that the minimax risk $\mathcal{R}_R[k]$ is of order*
$$C\frac{k}{n}\log\left(\frac{ep}{k}\right)\exp\left\{C\frac{k}{n}\log\left(\frac{ep}{k}\right)\right\} \ .$$
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.
2. **[Adaptation to sparsity]** *It also follows from Theorem 5.2 and Proposition 5.1 that $\widetilde{\theta}^V$ is minimax adaptive to all $1 \leq k \leq p^{1/3} \wedge (n-1)/4$. As a consequence, adaptation is possible for this problem.*
3. **[Unknown Variance]** *The estimator $\widetilde{\theta}^V$ does not requires the knowledge of the variance $\sigma^2 = Var(Y|X)$. Consequently, the minimax risk of prediction is of the same order for known and unknown variance.*
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 4.5 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 an universal constant $C > 0$ such that for all covariance Σ,*
$$\mathcal{R}_R[k, I_p] \geq C\mathcal{R}_R[k, \Sigma] \ .$$
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 (5.5) match up to constant $C(\gamma)$. However, the lower and the upper bounds do not exhibit the same dependency with respect to Σ, especially when $\Phi_{2k,-}(\sqrt{\Sigma})/\Phi_{2k,+}(\sqrt{\Sigma})$ is away from one.

5.2. Fixed design

5.2.1. Known variance

Proposition 5.3. **[Minimax lower bound]** *For any design $\mathbf{X}$ and any $1 \leq k \leq n$, the minimax risk $\mathcal{R}_F[k, \mathbf{X}]$ is lower bounded as follows*

$$\mathcal{R}_F[k, \mathbf{X}] \geq C\frac{\Phi_{2k,-}(\mathbf{X})}{\Phi_{2k,+}(\mathbf{X})}\frac{k}{n}\log\left(\frac{ep}{k}\right) \ . \tag{5.6}$$

For any $1 \leq k \leq n$, *we also have*

$$\mathcal{R}_F[k] \geq C \left[\frac{k}{n} \log\left(\frac{ep}{k}\right) \wedge 1 \right] . \tag{5.7}$$

The minimax lower bound (5.6) has been independently proved by Raskutti et al. [34]. Rigollet and Tsybakov [35] have also independently derived a bound similar to (5.6) which better handles the case where the rank of $\mathbf{X}$ is smaller than k. The bound (5.6) is useful to derive the second lower bound (5.7). The designs $\mathbf{X}$ that allow to prove this second lower bound when $k \log(p/k) \leq n/32$ correspond to realizations of a Gaussian standard independent design. See the proof for more details.

Remark 5.3. *We easily retrieve from (5.6) a result of asymptotic geometry first observed by Baraniuk et al. [2] in the special of restricted isometry property [12]. 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 $\mathcal{R}_F[k, \mathbf{X}] \leq \mathcal{R}_F[p, \mathbf{X}] \leq \mathcal{R}_F[p] \leq 1$ Gathering these two bounds allows to conclude. □

Let us turn to the upper bound. For the sake of simplicity, we assume in the rest of the section that $p \geq n$. Let us define a specific version of an estimation procedure due to Birgé and Massart [9, 10].

Definition 5.2. [Procedure for fixed design regression] *Define* k^* *as the smallest integer* k *such that* $k[1 + \log(p/k)] \leq n$. *Let us consider the collection of dimensions* $\mathcal{K} := \{1, \ldots, k^*\} \cup \{n\}$. *Then, is defined by the penalty function* $pen : \mathcal{K} \mapsto \mathbb{R}^+$

$$pen(k) := \begin{cases} 2k \left[1 + \sqrt{2 \log\left(\frac{e^2 p}{k}\right)} \right]^2 & \text{if} \quad k \leq k^* \\ 2n & \text{if} \quad k = n \ , \end{cases}$$

The size $\widehat{k}^{BM}$ *is selected by minimizing the following penalized criterion*

$$\widehat{m}^{BM} := \arg \inf_{m \in \mathcal{M}} \|\mathbf{Y} - \mathbf{X}\widehat{\theta}_k\|_n^2 + \sigma^2 pen(k) \ , \tag{5.8}$$

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. The following risk bound is a special case of Theorem 1 in Birgé and Massart [10].

Proposition 5.4. [Risk bound for $\widetilde{\theta}^{BM}$ (Birgé and Massart)] *Assume that* $p \geq n$. *For any* $1 \leq k \leq n$, *we set* $k_1 = k$ *if* $k \leq k^*$ *and* $k_1 = n$ *else. For any design* $\mathbf{X}$, *we have*

$$\sup_{\theta \in \Theta[k,p]} \mathbb{E}\left[\|\mathbf{X}(\widehat{\theta}_{k_1} - \theta)\|_n^2 / n \right] \leq C \left[\frac{k}{n} \log\left(\frac{ep}{k}\right) \wedge 1 \right] \sigma^2 \ . \tag{5.9}$$

For any $1 \leq k \leq n$ *and any design* $\mathbf{X}$, *the estimator* $\tilde{\theta}^V$ *satisfies*

$$\sup_{\theta \in \Theta[k,p]} \mathbb{E}\left[\|\mathbf{X}(\tilde{\theta}^{BM} - \theta)\|_n^2 / n \right] \leq C \left[\frac{k}{n} \log\left(\frac{ep}{k}\right) \wedge 1 \right] \sigma^2 \ . \tag{5.10}$$

Remark 5.4.

1. [**Minimax risks**] *We derive from (5.7) and (5.4) that the minimax risk $\mathcal{R}_F[k]$ is of order*

$$C\left[\frac{k}{n}\log(ep/k)\right]\wedge 1\ .$$

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.
2. [**Adaptation to sparsity**] *For any design $\mathbf{X}$, $\tilde{\theta}^{BM}$ simultaneously achieves the minimax risk of estimation at the sense of (5.7) over all $\Theta[k,p]$ with $1\leq k\leq n$. Thus, adaptation to the sparsity is possible when σ^2 is known.*
3. [**Adaptation to the design**] *For designs $\mathbf{X}$, such that the ratio $\Phi_{2k,-}(\mathbf{X})/\Phi_{2k,+}(\mathbf{X})$ is close to one, the lower bounds (5.6) and the upper bounds (5.10) agree with each other. However, the dependency of (5.6) on $\mathbf{X}$ is not sharp when the ratio $\Phi_{2k,-}(\mathbf{X})/\Phi_{2k,+}(\mathbf{X})$ is away from one. Take for instance and orthogonal design with $p=n$ and duplicate the last column. Then, the lower bound (5.6) for this new design $\mathbf{X}$ is 0 while the minimax risk is of order $k\log(p/k)/n$.*

Remark 5.5. [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 (5.6)).*

5.2.2. Unknown variance

We now consider the problem of prediction when the variance σ^2 is unknown. The optimal risk of prediction remain of the same order when σ^2 is unknown. Indeed, the minimax upper bound (5.9) involves the estimator $\widehat{\theta}_{k_1}$ that do rely on the knowledge of σ^2. Let us now study to what extent adaptation is possible when the variance σ^2 is unknown.

The estimator $\tilde{\theta}^V$ introduced in the previous subsection does not rely on the knowledge of σ^2. As a benchmark, we first provide a risk bound for $\tilde{\theta}^V$. This risk bound derives from the work of Baraud et al. [4].

Proposition 5.5. [Risk bound for $\widetilde{\theta}^V$] *Assume that $n\geq 14$. There exists a universal choice of K in the penalty (5.4) such that the following holds. For any design $\mathbf{X}$ and any $1\leq k\leq \lfloor (n-1)/4\rfloor$, we have*

$$\sup_{\theta\in\Theta[k,p]}\mathbb{E}\left[\|\mathbf{X}(\tilde{\theta}^V-\theta)\|_n^2/n\right]\leq C\frac{k}{n}\log\left(\frac{ep}{k}\right)\exp\left[C\frac{k}{n}\log\left(\frac{ep}{k}\right)\right]\sigma^2\ . \tag{5.11}$$

Remark 5.6. *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 a 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 n. In contrast, the estimator $\widehat{\theta}_n$ is minimax adaptive over all $\Theta[k,p]$ such that $k(1+\log(p)/k)\geq n$. Can we merge the qualities of $\tilde{\theta}^V$ and of $\widehat{\theta}_n$? The following proposition tells us that it is impossible.*

Proposition 5.6. [Adaptation is impossible] *Consider any $p \geq n \geq C$ and $k \in \{1, \dots, p^{1/3}\}$ such that $k \log(ep/k) \geq Cn$. There exists a design $\mathbf{X}$ of size $n \times p$ such that for any estimator $\widehat{\theta}$, we have either*

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

or

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

As a benchmark, we recall the minimax upper bounds (5.9):

$$\mathcal{R}_F[1] \leq C\frac{\log(p)}{n} \quad \text{and} \quad \mathcal{R}_F[k] \leq C\frac{k}{n}\log\left(\frac{ep}{k}\right) \wedge 1$$

The proof of proposition 5.6 uses the minimax lower bounds (4.13) for the testing problem $(\mathbf{P_1})$ under unknown variance.

Remark 5.7. *In the setup of Proposition 5.6, any estimator $\widehat{\theta}$ that does not rely on 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. 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*

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

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

In short, it is impossible to merge the qualities of $\widetilde{\theta}^V$ and of $\widehat{\theta}_n$. Furthermore, $\widetilde{\theta}^V$ achieves the optimal adaptive rate of estimation under unknown variance.

In conclusion, the minimax risk 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 a ultra-high dimensional setting, the minimax risks behave differently. In 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. Finally, there is a price to pay for adaptation under unknown variance and fixed design.

6. Inverse problem and support estimation

6.1. Minimax risk of estimation

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 $\mathcal{RI}[k,\mathbf{X}]$ and is defined in (3.7).

Proposition 6.1. [Minimax lower bound in fixed design] *For any design $\mathbf{X}$ and any $1 \leq k \leq n$, we have*

$$\mathcal{RI}[k,\mathbf{X}] \geq C\left[\frac{1}{\Phi_{2k\wedge p,-}(\mathbf{X})} \vee \frac{k\log(ep/k)}{\Phi_{2k\wedge p,+}(\mathbf{X})}\right] \ . \tag{6.1}$$

Proposition 6.2. [Upper bound in fixed design] *Consider any $k \leq n \wedge p$. The least-squares estimator $\widehat{\theta}_k$ defined in (5.3) satisfies*

$$\sup_{\theta \in \Theta[k,p]} \mathbb{E}\left[\|\widehat{\theta}_k - \theta\|_p^2\right] \leq C \frac{k \log(ep/k)}{\Phi_{2k,-}(\mathbf{X})} \sigma^2 \ . \tag{6.2}$$

The minimax lower and upper bounds match up to the ratio $\Phi_{2k,+}(\mathbf{X})/\Phi_{2k,-}(\mathbf{X})$. If the restricted eigenvalues of $\mathbf{X}$ are close to one, then the minimax risk is of order $k\log(ep/k)\sigma^2$. Note that the Lasso is proved to achieve this optimal rate under stronger assumptions on the design matrix $\mathbf{X}$ (see [34] Section 3.1).

Let us now study to what extent we can build designs that constrain the ratio $\Phi_{2k,+}(\mathbf{X})/\Phi_{2k,-}(\mathbf{X})$ to be close to one. We restrict ourselves to designs $\mathbf{X}$ such that each column has a unit norm, as justified in Section 3.3. The collection of such designs is noted $\mathcal{D}_{n,p}$. We recall that $\mathcal{RI}[k]$ is defined by $\inf_{\mathbf{X}\in\mathcal{D}_{\mathbf{n},\mathbf{p}}} \mathcal{RI}[k,\mathbf{X}]$.

Corollary 6.3. *Assume that $k\log(ep/k) \leq Cn$. Then, we have*

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

This bound is achieved by $\widehat{\theta}_k$ for some designs $\mathbf{X}$ that are realisations of a normalized Gaussian design.

This result is due to the following property: as soon as $k\log(ep/k)$ is small compared to n, there exists a design $\mathbf{X}$ such that the restricted eigenvalues of $\mathbf{X}^*\mathbf{X}$ of order $2k$ are close to one. For such designs, the minimax risk of estimation is of order $k\log(ep/k)$.

Proposition 6.4. *For any design $\mathbf{X} \in \mathcal{D}_{n,p}$ and any $k \leq n \wedge p/2$, we have*

$$\Phi_{2k,-}(\mathbf{X}) \leq Ck^2 \left(\frac{k}{p}\right)^{2k/n} \vee 1 \ . \tag{6.3}$$

Consider a sequence (k_n, p_n) such that $[k_n \log(p_n/k_n)]/\{n\log(n)\}$ goes to infinity and $k_n = o(n \vee p_n)$. Then, we have

$$\left(\frac{p_n}{k_n}\right)^{4k_n/n} \overset{\log}{\gtrsim} \mathcal{RI}[k_n] \overset{\log}{\sim} \inf_{\mathbf{X}\in D_{n,p_n}} \Phi^{-1}_{2k_n,-}(\mathbf{X}) \overset{\log}{\gtrsim} \left(\frac{p_n}{k_n}\right)^{2k_n/n} \ , \tag{6.4}$$

where $\overset{\log}{\sim}$ stands for log *equivalent and $\overset{\log}{\gtrsim}$ stands for* log *dominance.*

In a ultra-high dimensional setting, this is not anymore possible to build a design $\mathbf{X}$ such that the restricted eigenvalues of $\mathbf{X}^*\mathbf{X}$ of order $2k$ are close to one. In fact, the term $1/\Phi^{-1}_{2k_n,-}(\mathbf{X})$ blows up and becomes preponderant in the lower bound (6.1). As a consequence, it is not possible to achieve a smaller risk than $C(p/k)^{2k/n}$. While the quantity $k\log(p/k)$ in Corollary 6.3 is due to the size $\Theta[k,p]$, the minimax risk in ultra-high dimension is essentially driven by geometrical constrains on the design $\mathbf{X}$. Whatever design $\mathbf{X} \in \mathcal{D}_{n,p}$ we have, the minimax risk blows up to the order of $(p/k)^{Ck/n}$.

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.

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

Proposition 6.5. [Support recovery is almost impossible] *For any $\rho^2 \leq Ck^{-1}\left(\frac{p}{k}\right)^{2k/n}$, we have*

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

Assume that $k\log(ep/k)/[n\log(n)]$ is larger than 4 (ultra-high dimensional setting). Then, for any design $\mathbf{X} \in \mathcal{D}_{n,p}$ it is not possible to recover the support of θ with high probability, unless θ satisfies:

$$\frac{\|\theta\|_p^2}{\sigma^2} \geq C\left(\frac{p}{k}\right)^{k/n} \ .$$

As it is almost impossible to estimate the support of θ in a ultra-high dimensional setting, we may aim to an easier objective. Can we choose a subset $\widehat{M}$ of $\{1,\dots,p\}$ of size $p_0 \leq p$ that contains the support of θ 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? How small can we choose p_0?

Proposition 6.6. *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*

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

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

$$\sup_{\theta\in\mathcal{C}_k^p(\rho)} \mathbb{P}_{\theta,1}\left[\mathrm{supp}(\theta) \not\subseteq \widehat{M}\right] \geq 1/8 \ . \tag{6.5}$$

In a 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 is of order $\exp[Ck/n\log(p/k)]$. In (6.5), 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 ($(k,p,n) \to \infty$).

In order to shed light on the phenomenon, 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 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 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 6.6. Consequently, we still have to keep at least $\exp(\delta n^{\gamma_1})$ variables after the process of dimension reduction!

7. What is a ultra-high dimensional problem?

Until now, we have stated that a problem is ultra-high dimensional when $k\log(ep/k)$ is large compared to n. At the end of this section, we provide a simple rule of thumb to decide whether a problem should be considered as ultra-high dimensional. This claim is supported by a simulation study.

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 such that $\theta_1 = \ldots = \theta_k = 4\sqrt{\log(p)/n} \approx 1.30$ (resp. 1.65) for $p = 200$ (resp. $p = 5000$) and $\theta_{k+1} = \ldots = \theta_p = 0$. As a consequence, we have $\|\theta\|^2 = 16k\log(p)/n$. The non-zero coefficients of θ are chosen large enough so that the support of θ 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 [22] 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 [21] 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 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.

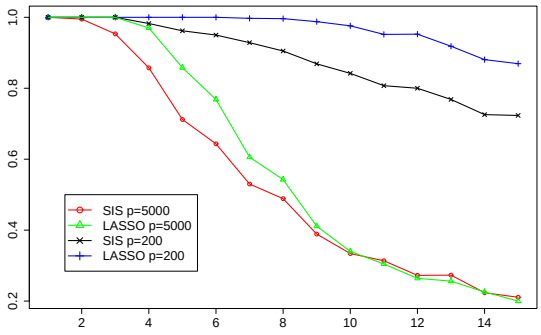

FIGURE 1. *Power of the dimension reduction procedures (SIS and LASSO)*

Results. The results are presented on Figure 1. 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 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 $k > 8$.

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 we take θ such that $\theta_1 = \ldots = \theta_k = u\sqrt{\log(p)/n}$ and $\theta_{k+1} = \ldots = \theta_p = 0$. Thanks to $N = 100$ 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.

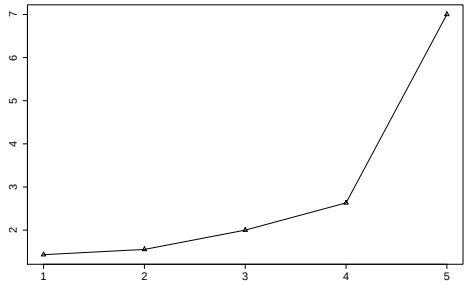

FIGURE 2. *Minimal signal u_k^* as a function of k.*

Results. The results are presented on Figure 2. 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. This phenomenon is particularly striking when p is much larger than n. From these simulations and from other theoretical arguments (e.g. [24]), we derive a simple rule of thumb. We say that a problem is ultra-high dimensional if

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

For $p = 5000$ and $n = 50$, this corresponds to $k \geq 4$. Setting $p = 200$ and $n = 50$ yields $k \geq 8$.

8. Discussion

As proved in Sections 3–6, the behaviors of the minimax separation distances and of the minimax risks become really different in a 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 a ultra-high dimensional setting, except if signal over noise ratio is exponentially large. Moreover, even dimension reduction techniques like correlation screening 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 σ^2. Some specific work is therefore needed to derive fast and efficient procedures under unknown variance.

We have not discussed so far the problem of variance estimation. From the testing minimax lower bounds, we can deduce that the minimax square risks of estimation of $\widehat{\sigma}$ is at least of order $\exp(Ck/n\log(p/k))$ in a ultra-high dimensional setting when $\theta \in \Theta[k,p]$ is unknown.

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 $\mathbf{X}^*\mathbf{X}$ are close to one. However, these bounds do not agree anymore when these restricted eigenvalues are away from one. 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 a 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 seems 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 [36], the restricted eigenvalues of $\mathbf{X}^*\mathbf{X}$ 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

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 $\beta_{\Sigma,\sigma,\alpha}(\mathcal{T})$ the quantity

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

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

Most of the minimax lower bounds in this paper are based on an approach which comes back to Ingster [25, 26, 27]. The following lemma encompasses fixed and random design and fixed and random variance.

Lemma 9.1. *Let $\mathcal{T}$ be a subset of $\mathbb{R}^p \setminus \{0\} \times \mathbb{R}_+^*$ and let μ a probability measure on $\mathcal{F}$. We note $\mathbb{P}_\mu = \int_{\mathcal{T}} \mathbb{P}_{\theta,\sigma} d\mu$ and $L_\mu = d\mathbb{P}_\mu / d\mathbb{P}_{0,\sigma_0}$. Then,*

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

Here, β_α should be replaced by $\beta_{\Sigma,\sigma,\alpha}$, $\beta_{\Sigma,\alpha}$, $\beta_{\mathbf{X},\sigma,\alpha}$, or $\beta_{\mathbf{X},\alpha}$ depending on the design (fix or random) or the variance (known or unknown). We refer to Baraud [3] 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 $\mathbb{P}_\mu$ and $\mathbb{P}_{0,\sigma_0}$ is as large as possible. We have $\beta_\alpha(\mathcal{T}) \geq \delta$ if $\mathbb{E}_{0,\sigma_0}[L_\mu^2(\mathbf{Y},\mathbf{X})] \leq 1+\eta^2$.

9.1. Proof of 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 on $\Theta[k,p]$ in order to apply Lemma 9.1.

Let us take the set $\widehat{m}$ of size k uniformly in $\mathcal{M}(k,p)$. We recall that $\mathcal{M}(k,p)$ is the collection of all subsets of $\{1,\dots,p\}$ of size k. For each $m \in \mathcal{M}(k,p)$, let $\xi^m = (\xi_j^m)_{j\in m}$ 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 $\theta_{\hat{m},\xi} = \sum_{j\in\hat{m}} \lambda \xi_j^{\widehat{m}} e_j$. Here, $(e_j)_{1\le j\le p}$ 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 $L_{\mu_\rho}(\mathbf{X},\mathbf{Y}) = P_{\mu_\rho}/P_0$ writes:

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

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

$$\begin{aligned}
&\mathbb{E}_0\left[L_{\mu_\rho}^2(\mathbf{X},\mathbf{Y})\right] \\
&= 2^{-2k}\binom{p}{k}^{-2} \sum_{m,m',\xi_1,\xi_2} \mathbb{E}_0\left[e^{-\left(\|\mathbf{X}\theta_{m_1,\xi_1}\|_n^2 + \|\mathbf{X}\theta_{m_2,\xi_2}\|_n^2\right)/2 + \left\langle \mathbf{Y}, \mathbf{X}\left(\theta_{m_1,\xi_1}+\theta_{m_2,\xi_2}\right)\right\rangle_n}\right] \\
&= \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{m_1,m_2,\xi_1,\xi_2}\left\{\exp\left(\langle \mathbf{X}\theta_{m_1,\xi_1}, \mathbf{X}\theta_{m_2,\xi_2}\rangle\right)\right\}\right] . \qquad (9.2)
\end{aligned}$$

Lemma 9.2. *If we assume that*

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

then, we have

$$\mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{0,\mathbf{Y}}\left\{L_{\mu_\rho}^2(\mathbf{Y},\mathbf{X})\right\}\right] \le 1+\eta^2 .$$

In this lemma, we have specifically distinguished the integration with respect to $\mathbf{X}$ from the integration with respect with respect to $\mathbf{Y}$. This will be useful for deriving minimax lower bound in fixed design (Proposition 4.3). Gathering Lemmas 9.1 and 9.2 allow to derive that

$$(\rho_R^*[k,I_p])^2 \ge C(\alpha,\delta)\left[\frac{k}{n}\log\left(1+\frac{p}{k^2}\right) \wedge \frac{1}{\sqrt{n}}\right] .$$

□

Proof of Lemma 9.2. Let us fix m_1, m_2, ξ_1 and ξ_2. First, we shall compute the expectation $\mathbb{E}[\exp(\langle \mathbf{X}\theta_{m_1,\xi_1}, \mathbf{X}\theta_{m_2,\xi_2}\rangle)]$.

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:

$$\begin{aligned} m_3 &:= \left\{ j \in m_1 \cap m_2 | \xi_j^1 = \xi_j^2 \right\} \\ m_4 &:= \left\{ j \in m_1 \cap m_2 | \xi_j^1 = -\xi_j^2 \right\} . \end{aligned}$$

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.

$$\begin{aligned} &\mathbb{E}\left[\exp\left(\langle \mathbf{X}\theta_{m_1,\xi_1}, \mathbf{X}\theta_{m_2,\xi_2}\rangle\right)\right] \\ &= \left[\int_{\mathbb{R}^p} (2\pi)^{-p/2} \exp\left(-\sum_{i=1}^p t_i^2/2 + \sum_{1\le i,j\le p} [\theta_{m_1,\xi_1}]_i [\theta_{m_2,\xi_2}]_j t_i t_j \right) \prod_{i=1}^p dt_i \right]^n \\ &= \left| I_{|m_1\cup m_2|} - \lambda^2 C \right|^{-n/2} , \end{aligned}$$

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, $I_{|m_1\cup m_2|} - \lambda^2 C$ 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

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

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

$$\begin{aligned} \mathbb{E}[\exp(\langle \mathbf{X}\theta_{m_1,\xi_1}, \mathbf{X}\theta_{m_2,\xi_2}\rangle)] &\le \left[1 - 2(2|m_3| - |m_1 \cap m_2|)\lambda^2 - 8\rho^4\right]^{-n/2} \\ &\le \exp\left(8n\rho^4\right) \exp\left[2n\lambda^2(2|m_3| - |m_1 \cap m_2|)\right] . \end{aligned} \tag{9.3}$$

Then, we take the expectation with respect to ξ^1, ξ^2, m_1 and m_2. When m_1 and m_2 are fixed the expression (9.3) depends on ξ^1 and ξ^2 only through the cardinality of m_3. As ξ^1 and ξ^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

$$\mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{0,\mathbf{Y}}\left\{L_{\mu_\rho}^2(\mathbf{Y}, \mathbf{X})\right\}\right] \le \exp\left(8n\rho^4\right) \mathbb{E}\left[\exp\left(2n\lambda^2 Z\right)\right] . \tag{9.4}$$

We now proceed as in the proof of Theorem 1 in Baraud [3] in order to upper bound the term

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

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

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

Moreover, we have $\exp(8\rho^4 n) \le \sqrt{1+\eta^2}$ as soon as $\rho^2 \le C(\eta)/\sqrt{n}$. Gathering these observations with (9.4), we conclude that $\mathbb{E}_{\mathbf{X}}[\mathbb{E}_0\{L^2_{\mu_\rho}(\mathbf{Y},\mathbf{X})\}] \le 1+\eta^2$ as soon as

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

□

9.2. Proof of Theorem 4.5

Proof of Theorem 4.5. Consider the Condition

$$(\mathbf{A}.1) \qquad \frac{k}{n}\log\left(\frac{p}{e^4k^2}\right) \ge 2\,.$$

We deduce Theorem 4.5 from the following result.

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

$$\beta_{I_p,\alpha}\left(\left\{\theta\in\Theta[k,p],\ \sigma^2>0,\ \frac{\|\theta\|_p^2}{\sigma^2}=\rho^2\right\}\right) \ge \delta \;, \tag{9.5}$$

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

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

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

$$\rho^2 \le -1+\left(\frac{p}{ek}\right)^{\frac{k}{n}}(8k)^{-2/n} \;. \tag{9.7}$$

If $p \ge k^{1/3}\vee C$ and $k\log(p)/n \ge C_1$ 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

$$\begin{aligned} -1+\left(\frac{p}{ek}\right)^{\frac{k}{n}}(8k)^{-2/n} &\ge -1+\exp\left[C\frac{k}{n}\log\left(\frac{ep}{k}\right)\right] \\ &\ge C\frac{k}{n}\log\left(\frac{ep}{k}\right)\exp\left[C\frac{k}{n}\log\left(\frac{ep}{k}\right)\right] \;. \end{aligned}$$

Let us now assume that $p \geq k^{1/3} \vee C$ and $k\log(p)/n \leq C_1$ where C_1 has been previously fixed. Then, the first lower bound (9.6) satisfies:

$$\frac{k}{2n}\log\left(1+\frac{p}{k^2}\vee\sqrt{\frac{p}{k^2}}\right) \geq C\frac{k}{n}\log\left(\frac{ep}{k}\right)\exp\left[C\frac{k}{n}\log\left(\frac{ep}{k}\right)\right] .$$

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

$$(\rho_U^*[k,I_p])^2 \geq C\frac{k}{n}\log\left(\frac{ep}{k}\right)\exp\left[C\frac{k}{n}\log\left(\frac{ep}{k}\right)\right] .$$

□

Proof of Lemma 9.3. The minimax lower bound (9.6) has already been proved in Theorem 4.3 in [38]. We only have to prove the lower bound (9.7).

Consider some $\rho > 0$. To apply Lemma 9.1, we first have to define a suitable prior μ_ρ on θ and σ^2. More specifically, the distribution μ_ρ is support by $\Theta[k,p,\rho]\times\{\sigma^2(\rho)\}$ defined by

$$\begin{aligned}\Theta[k,p,\rho] &:= \left\{\theta\in\Theta[k,p]\ ,\ \frac{\|\theta\|_p^2}{1-\|\theta\|_p^2}=\rho^2\right\}\\ \sigma^2(\rho) &= 1-\|\theta\|_p^2\end{aligned}$$

Let $\widehat{m}$ be some random variable uniformly distributed over $\mathcal{M}(k,p)$. For each $m\in\mathcal{M}(k,p)$, let $\xi^m=(\xi_j^m)_{j\in m}$ be a sequence of independent Rademacher random variables. We assume that for all $m\in\mathcal{M}(k,p)$, ξ^m and $\widehat{m}$ are independent. Let ρ be given and μ_ρ be the distribution of the random variable $\widehat{\theta}=\sum_{j\in\widehat{m}}\lambda\xi_j^{\widehat{m}}e_j$ 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 that for any $\rho^2 \leq -1 + (p/(ek))^{k/n}(8k)^{-2/n}$, we have

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

Observe here that we use a variance 1 for $\mathbf{H_0}$ and a variance $1-\|\theta\|_p^2$ 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.

Let us define the random variable $Z=\sum_{i=1}^R\xi_i$, where R is distributed as a Hypergeometric distribution with parameters p, k, and k/p and the ξ_i are independent Rademacher random variables. In [38] Eq.(8.6) , It has been proved that

$$\mathbb{E}_{0,1}(L^2_{\mu_\rho}(\mathbf{Y},\mathbf{X})) = \mathbb{E}\left(1-\frac{\rho^2 Z}{(1+\rho^2)k}\right)^{-n} , \tag{9.8}$$

Let us define the random variable W by $W=\sum_{i=1}^k\xi_i a_i$ where $(\xi_i)_{i=1,\dots,k}$ are independent Rademacher variables and $(a_i)_{i=1,\dots,k}$ are independent Bernoulli random variables with parameters k/p. We first provide the main steps of the proof. These steps are proved afterwards.

FACT 1.

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

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

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

Since we need to ensure that $\mathbb{E}[\{1-\rho^2 W/((1+\rho^2)k)\}^{-n}-1]\leq \eta^2$, 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 concentration inequality of the random variable W/k.

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

$$\mathbb{P}\left[\frac{W}{k}\geq x\right]\leq\left[\left(\frac{k}{2px}\right)^{x}\frac{1}{(1-x)^{1-x}}\right]^{k}\ . \tag{9.12}$$

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

FACT 3. The upper bound (9.11) holds for any $\lfloor k/2\rfloor+1\leq i\leq k$ as soon as

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

We derive that under (9.13), we have $\mathbb{E}_{0,1}[L^2_{\mu_\rho}(\mathbf{Y},\mathbf{X})]\leq 1+\eta^2$. The fact that $\eta^2\geq 1/8$ allows to conclude.

□

Proof of Fact 1. Let us introduce the function $\psi(.)$ defined by

$$\psi(r)=\mathbb{E}_\xi\left[\left(1-\frac{\rho^2\sum_{i=1}^{r}\xi_i}{1+\rho^2 k}\right)^{-n}\right]\ .$$

FACT 4. $\psi(r)$ is a convex function with respect to r.

We deduce from the definition of ψ that

$$\mathbb{E}\left[\psi(R)\right] \ = \ \mathbb{E}\left[\left(1-\frac{\rho^2 Z}{(1+\rho^2)k}\right)^{-n}\right]\ ,$$

where Z is a random variable distributed according to a Hypergeometric distribution with parameters p, k and k/p. We know from Aldous (p.173) [1] that Z follows the same distribution as the random variable $\mathbb{E}(W|\mathcal{B}_p)$ where W is binomial random variable of parameters k, k/p and $\mathcal{B}_p$ some suitable σ-algebra. By a convexity argument, we get

$$\mathbb{E}\left[\left(1-\frac{\rho^2 Z}{(1+\rho^2)k}\right)^{-n}\right] \leq \mathbb{E}\left[\psi(W)\right] \leq \mathbb{E}\left[\left(1-\frac{\rho^2 W}{(1+\rho^2)k}\right)^{-n}\right] ,$$

which concludes the proof. □

Proof of Fact 4. We shall prove that for any $0 \leq r \leq k-2$, we have

$$\psi(r+2) - 2\psi(r+1) + \psi(r) \geq 0 .$$

Let us first express $\psi(r+1)$ in terms of $\sum_{i=1}^r \xi_i$.

$$\begin{aligned}\psi(r+1) &= \mathbb{E}_\xi\left[\left(1-\frac{\rho^2\sum_{i=1}^{r+1}\xi_i}{(1+\rho^2)k}\right)^{-n}\right] \\ &= \frac{1}{2}\mathbb{E}_\xi\left[\left(1-\frac{\rho^2(\sum_{i=1}^{r}\xi_i+1)}{(1+\rho^2)k}\right)^{-n}+\left(1-\frac{\rho^2(\sum_{i=1}^{r}\xi_i-1)}{(1+\rho^2)k}\right)^{-n}\right] .\end{aligned}$$

Similarly, we express $\psi(r+2)$ in terms of $\sum_{i=1}^r \xi_i$. If we define

$$V = \frac{(1+\rho^2)k}{\rho^2} - \sum_{i=1}^r \xi_i ,$$

then $\psi(r+2) - 2\psi(r+1) + \psi(r)$ decomposes as

$$\begin{aligned}\psi(r+2) - 2\psi(r+1) + \psi(r) &= \left[\frac{\rho^2}{(1+\rho^2)k}\right]^n \\ &\times \mathbb{E}_V\left[\frac{1}{4}(V-2)^{-n} - (V-1)^{-n} + \frac{3}{2}V^{-n} - (V+1)^{-n} + \frac{1}{4}(V+2)^{-n}\right] .\end{aligned}$$

We shall prove that the random variable inside the expectation is almost surely non-negative. For any $x \geq 2$, we consider the expression

$$\frac{1}{4}(x-2)^{-n} - (x-1)^{-n} + \frac{3}{2}x^{-n} - (x+1)^{-n} + \frac{1}{4}(x+2)^{-n} .$$

Let us define the function g by

$$g(u) = (x+\sqrt{u})^{-n} + (x-\sqrt{u})^{-n} = \frac{(x+\sqrt{u})^n + (x-\sqrt{u})^n}{(x^2-u)^n} .$$

The function $u \mapsto (x^2-u)^{-n}$ is positive, increasing and convex.

$$(x+\sqrt{u})^n + (x-\sqrt{u})^n = \sum_{i=0}^{\lfloor n/2 \rfloor}\binom{2i}{n}u^i x^{n-2i} .$$

Hence, the function $u \mapsto (x+\sqrt{u})^n + (x-\sqrt{u})^n$ positive, increasing and convex. It follows that the function g is convex. Consequently, we have

$$\frac{1}{4}(x-2)^{-n} + \frac{3}{2}x^{-n} + \frac{1}{4}(x+2)^{-n} - (x-1)^{-n} - (x+1)^{-n} \geq 0 \ ,$$

and we conclude that $\psi(.)$ is convex.

□

Proof of Fact 2. 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 concentration inequality for W.

$$\mathbb{P}\left[\frac{W}{k} \geq x\right] \leq \left(\frac{ke}{2px}\right)^{xk} \ , \tag{9.14}$$

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

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

Taking the logarithm of this expression leads to

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

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

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

By Assumption (**A.1**), $k/n\log[2p/(ek^2)]$ 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:

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

Since η is larger than 0.41, $\log(2e/\eta^2)$ is smaller than 4 and this last inequality is ensured by Assumption (**A.1**).

□

Proof of Fact 3. We consider here the case $1/2 < i/k \leq 1$. We derive from (9.14) that

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

Consequently, we want to ensure that

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

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

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

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

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

We conclude that the upper bounds hold if

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

□

Proof of Lemma 9.4. We prove this concentration inequality using the Laplace transform of W/k.

$$\begin{aligned} \log\left[\mathbb{E}\left\{\exp(\lambda W/k)\right\}\right] &= k \log\left[1 + \frac{k}{p}\left(\cosh\left(\frac{\lambda}{k}\right) - 1\right)\right] \\ &\leq k \log\left[1 + \frac{k}{2p}\left(\exp\left(\frac{\lambda}{k}\right) - 1\right)\right] , \end{aligned}$$

if λ is positive. Consider some $x \in (0,1)$.

$$\begin{aligned} \log\left[\mathbb{P}\left\{\frac{W}{k} \geq x\right\}\right] &\leq -\lambda x + \log\left[\mathbb{E}\left\{\exp(\lambda W/k)\right\}\right] \\ &\leq -\lambda x + k \log\left[1 + \frac{k}{2p}\left(\exp\left(\frac{\lambda}{k}\right) - 1\right)\right] . \end{aligned}$$

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

$$e^{\lambda^*/k} = \frac{x}{1-x}\left(\frac{2p}{k} - 1\right) .$$

Hence, we get

$$\log\left[\mathbb{P}\left\{\frac{W}{k} \geq x\right\}\right] \leq -kx \log\left[\frac{x}{1-x}\left(\frac{2p}{k} - 1\right)\right] + k \log\left[\frac{1 - k/2p}{1-x}\right] .$$

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

$$\mathbb{P}\left\{\frac{W}{k} \geq x\right\} \leq \left[\left(\frac{k}{2px}\right)^x \frac{1}{(1-x)^{1-x}}\right]^k .$$

Since $\mathbb{P}(W = k) = [k/(2p)]^k$, this upper bound is also valid when $x = 1$. □

9.3. Proof of Theorem 4.3

By homogeneity, we can assume that $\sigma^2 = 1$. The design $\mathbf{X}$ will be fixed later. Given $\rho > 0$, we take exactly the same prior probability μ_ρ on θ as in the proof of Theorem 4.1. The likelihood ratio $L_{\mu_\rho}(\mathbf{X}, \mathbf{Y}) = P_{\mu_\rho}/P_0$ writes:

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

As in the random design case (proof of Theorem 4.1), we compute the expectation of $L^2_{\mu_\rho}(\mathbf{X}, \mathbf{Y})$:

$$\mathbb{E}_0\left[L^2_{\mu_\rho}(\mathbf{X}, \mathbf{Y})\right] = \mathbb{E}_{m_1,m_2,\xi_1,\xi_2}\left[\exp\left(\langle \mathbf{X}\theta_{m_1,\xi_1}, \mathbf{X}\theta_{m_2,\xi_2}\rangle_n\right)\right] . \tag{9.15}$$

We want to prove that for some design $\mathbf{X}$ the quantity $\mathbb{E}_0[L^2_{\mu_\rho}(\mathbf{X}, \mathbf{Y})]$ is smaller than $1 + \eta^2$. Suppose that the design $\mathbf{X}$ is the observation of a standard normal design: for any $1 \le i \le p$ and $1 \le j \le n$, $\mathbf{X}_{i,j} \sim \mathcal{N}(0, 1)$. By Lemma 9.2, we have $\mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_0\{L^2_{\mu_\rho}(\mathbf{Y}, \mathbf{X})\}\right] \le 1 + \eta^2/2$ if we take

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

For such a ρ^2, we apply Markov's inequality and get

$$\mathbb{P}_{\mathbf{X}}\left[\mathbb{E}_0\left\{L^2_{\mu_\rho}(\mathbf{Y}, \mathbf{X})\right\} \le 1 + \eta^2\right] \ge \frac{\eta^2}{2(1 + \eta^2)} . \tag{9.16}$$

With positive probability, the design $\mathbf{X}$ satisfies $\mathbb{E}_0\{L^2_{\mu_\rho}(\mathbf{Y}, \mathbf{X})\} \le 1 + \eta^2$. To conclude, we need to study different cases depending on the values of k, n, and p.

CASE 1. $k \log\left(1 + \frac{p}{k^2} \vee \sqrt{\frac{p}{k^2}}\right) \le \sqrt{n}/2$. It follows that $k \le \sqrt{p}$ since $p \ge n$. Hence, we derive that

$$k(1 + \log(p/k)) \le C\sqrt{n}\log(n).$$

Applying Lemma A.2, we control the largest restricted eigenvalue of order k of the random matrix $\mathbf{X}^*\mathbf{X}$. With probability larger $1 - 2\exp(-\sqrt{n}/2)$,

$$\begin{aligned}
\Phi_{k,+}(\mathbf{X}^*\mathbf{X}/n) &\le \left(1 + 3\sqrt{\frac{k(1 + \log(p/k))}{n}} + n^{-1/4}\right)^2 , \\
\Phi_{k,-}(\mathbf{X}^*\mathbf{X}/n) &\ge \left(1 - 3\sqrt{\frac{k(1 + \log(p/k))}{n}} - n^{-1/4}\right)^2 .
\end{aligned}$$

If n is larger than some constant C, then $\Phi_{k,+}(\mathbf{X}^*\mathbf{X}/n) \le 3/2$ and $\Phi_{k,-}(\mathbf{X}^*\mathbf{X}/n) \le 1/2$ with probability larger than $1 - 2\exp(-\sqrt{n}/2)$.

Gathering this result with (9.16), we conclude that for n larger than some quantity $C(\eta)$ with probability larger than $\eta^2/(4(1 + \eta^2)$ the design $\mathbf{X}/\sqrt{n}$ satisfies a $1/\sqrt{2}$-restricted isometry of order

k, and $\mathbb{E}_{0,\mathbf{Y}}\{L^2_{\mu_\rho}(\mathbf{Y},\mathbf{X})\} \leq 1+\eta^2$. Let us consider such a design $\mathbf{X}$. For any subset $m \in \mathcal{M}(k,p)$ and any $\xi \in \{-1,1\}^k$, we have

$$\|\mathbf{X}\theta_{m,\xi}\|_n^2/n \geq 1/2\|\theta_{m,\xi}\|_p^2 = \rho^2/2.$$

As a consequence μ_ρ is supported by $\Theta'[k,p,\rho\sqrt{n}/\sqrt{2}]$ defined by

$$\Theta'[k,p,\rho/\sqrt{2}] := \left\{\theta \in \Theta[k,p]\ ,\ \|\mathbf{X}\theta\|_n^2/n \geq \rho^2/2\right\}\ .$$

Applying Lemma 9.1, we conclude that

$$\beta_{\mathbf{X},1,\alpha}\left(\Theta'\left[k,p,\rho/2\right]\right) \geq \delta\ .$$

CASE 2. $k\log\left(1+\frac{p}{k^2}\vee\sqrt{\frac{p}{k^2}}\right) \geq \sqrt{n}/2$.

CASE 2.a. $\log(1+p) \leq \sqrt{n}/2$. In such a case, there exists some k' between 1 and k such that

$$\sqrt{n}/4 \leq k'\log\left(1+\frac{p}{k'^2}\vee\sqrt{\frac{p}{k^2}}\right) \leq \sqrt{n}/2\ .$$

As proved in CASE 1, it is possible to build designs $\mathbf{X}$ of size $n\times p$ such that the $\rho^*_F[k',\mathbf{X}] \geq C(\eta)/\sqrt{n}$. Since $\rho^*_F[k,\mathbf{X}] \geq \rho^*_F[k',\mathbf{X}]$, we can conclude.

CASE 2.b. $\log(1+p) > \sqrt{n}/2$. For n large enough, there exists some p' between n and p, such that

$$\sqrt{n}/4 \leq \log\left(1+p'\right) \leq \sqrt{n}/2\ .$$

By CASE 2.a, we can build design $\mathbf{X}'$ of size $n\times p'$ such that the (α,δ) minimax separation distance over $\Theta[k,p']$ is larger than $C(\eta)\sqrt{n}$. To conclude, we only have to take any completion of the design $\mathbf{X}'$ to get a design $\mathbf{X}$ of size $n\times p$.

9.4. Proof of Proposition 4.7

Take some $\rho < \rho^*_R[k,I_p]/2$ as defined in Theorem 4.5. Assume that for any design $\mathbf{X}$ of size $n\times p$, there exists a test $\phi_\alpha[\mathbf{X}]$ of level α that satisfies the following property. For any $\sigma^2>0$ and any $\theta\in\Theta[k,p]$ such that $\|\mathbf{X}\theta\|_n^2/n \geq \rho^2\sigma^2$, we have

$$\mathbb{P}_{\theta,\sigma}\left[\phi_\alpha[\mathbf{X}]=0\right] \leq \delta/2\ .$$

Let us consider a Gaussian random design with the identity covariance as in the proof of Theorem 4.5. Then, the test T defined by $T(\mathbf{Y},\mathbf{X}) = \phi_\alpha[\mathbf{X}](\mathbf{Y})$ has a level α and does not require the knowledge of σ^2.

Take any $\theta\in\Theta[k,p]$ and and $\sigma>0$ such that $\|\theta\|_p^2/\sigma^2 = 2\rho^2$. The random variable $\|\mathbf{X}\theta\|_n^2/\|\theta\|_p^2$ follows a χ^2 distribution with n degrees of freedom. Applying Lemma A.1, we derive that with probability larger than $1-e^{-n/16}$, we have $\|\mathbf{X}\theta\|_n^2/n \geq \rho^2\sigma^2$. If follows that

$$\mathbb{P}_{\theta,\sigma}\left[T(\mathbf{X},\mathbf{Y})=0\right] \leq \delta/2 + e^{-n/16}\ ,$$

which contradicts Theorem 4.5.

9.5. Proof of Proposition 5.1

We derive this minimax lower bound from the hypothesis testing problem "$\theta = 0$" studied in Section 4. Since the covariance $\Sigma = I_p$, the loss $\mathbb{E}\left[\{X(\theta_1 - \theta_2)\}^2\right]$ is simply $\|\theta_1 - \theta_2\|_p^2$. 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 $\sigma^2 = 1$ and we consider the two sets

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

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 $\|\widetilde{\theta} - \theta\|_p \leq 2\|\widehat{\theta} - \theta\|_p$, for any $\theta \in \Theta_1[\rho] \cup \Theta_2[\rho]$.

$$\sup_{i=1,2} \sup_{\theta \in \Theta_i[\rho]} \mathbb{E}\left[\|\widehat{\theta} - \theta\|_p^2\right] \geq \frac{\rho^2}{4} \sup_{i=1,2} \sup_{\theta \in \Theta_i[\rho]} \mathbb{P}_{\theta,1}[\mathrm{supp}(\widetilde{\theta}) \not\subseteq M_i] \ . \tag{9.17}$$

It is enough to prove that for $\rho^2 = C\frac{k}{n}\log\left(\frac{p}{k}\right)\exp\{C\frac{k}{n}\log\left(\frac{p}{k}\right)\}$, the supremum of the probabilities $\mathbb{P}_\theta[\mathrm{supp}(\widetilde{\theta}) \not\subseteq M_i]$ is lower bounded by a positive constant. This is equivalent to lower bounding the minimax separation distance for $H_0 : \theta \in \Theta_1[\rho]$ against H_1: $\theta \in \Theta_2[\rho]$.

As in the proof of Theorem 4.5, we build a prior distribution $\mu_{1,\rho}$ on $\Theta_1[\rho]$. Consider the collection $\mathcal{M}_1(k)$ of subsets of M_1 of size k. Let $\widehat{m}$ be be some random variable uniformly distributed over $\mathcal{M}_1(k)$. For each $m \in \mathcal{M}_1(k)$, let $\xi^m = (\epsilon_j^m)_{j \in m}$ be a sequence of independent Rademacher random variables. We assume that for all $m \in \mathcal{M}_1(k)$, ξ^m and $\widehat{m}$ are independent. Then, $\mu_{1,\rho}$ is the distribution $\widehat{\theta} = \sum_{j \in \widehat{m}} \rho/\sqrt{k}\xi_j^{\widehat{m}} e_j$. Similarly, we define the prior distribution $\mu_{2,\rho}$ on $\Theta_2[\rho]$. We note $\mathbb{P}_{\mu_i} = \int \mathbb{P}_{\theta,1} d\mu_{i,\rho}$. We have

$$\begin{aligned}\sup_{i=1,2} \sup_{\theta \in \Theta_i[r]} \mathbb{P}_\theta[\mathrm{supp}(\widetilde{\theta}) \not\subseteq M_i] &\geq 1 - \frac{1}{2}\|P_{\mu_1} - P_{\mu_2}\|_{TV} \ . \\ &\geq 1 - \|P_{\mu_1} - P_{0,1+\rho^2}\|_{TV} \ , \end{aligned} \tag{9.18}$$

by the triangle inequality. Lemma 9.1 states that

$$\|P_{\mu_1} - P_0\|_{TV} \leq \mathbb{E}_0\left[L_{\mu_{1,\rho}}^2 - 1\right] \ ,$$

where $L_{\mu_{1,\rho}} = \mathrm{d}\mathbb{P}_{\mu_{1,\rho}}/\mathrm{d}\mathbb{P}_{0,1+\rho^2}$. In fact, the second moment of $L_{\mu_{1,\rho}}$ has been studied in the proof of Theorem 4.5 and in the proof of Theorem 4.3 in Verzelen and Villers [38]. If we take $\alpha + \delta = 53\%$ in these two proofs, we derive from (9.13) and from the proof of Theorem 4.3 [38] that

$$\mathbb{E}_0\left[L_{\mu_{1,\rho}}^2\right] \leq e^{1/2} \ , \qquad \text{if } \rho^2 \leq Ck/n\log(p/k)\exp(Ck/n\log(p/k)) \text{ and if } p \geq k^3 \vee C \ .$$

Gathering this result with Equations (9.17) and (9.18) allows to conclude.

9.6. Proof of Proposition 5.3

This lower bound is based on Birgé version of Fano's Lemma [8]. Consider θ_1 and θ_2 in $\mathcal{C}_k^p(\sqrt{k}r)$, where the set $\mathcal{C}_k^p(.)$ has been introduced in Definition 6.1. We upper bound the Kullback distance between $\mathcal{K}(\theta_1,\theta_2)$ between the probability distribution $\mathbb{P}_{\theta_1,\sigma}$ and $\mathbb{P}_{\theta_2,\sigma}$.

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

Let us also lower bound the loss $\|\mathbf{X}(\theta_1-\theta_2)\|_n^2$.

$$\|\mathbf{X}(\theta_1-\theta_2)\|_n^2\geq r^2\Phi_{2k,-}(\mathbf{X})d_H(\theta_1,\theta_2)\ ,$$

where d_H is the Hamming distance. The rest of the proof is analogous to the case of Gaussian sequence model (e.g. Proposition 4.11 in [32]). Thanks to combinatorial results such Varshamov lemma or Lemma 4.10 in [32], we build a subset $\mathcal{C}_k^{'p}(\sqrt{k}r)$ of $\mathcal{C}_k^p(\sqrt{k}r)$ whose points are at least $k/2$-separated with respect to the Hamming distance. The cardinal of $\mathcal{C}_k^{'p}(\sqrt{k}r)$ is larger than $Ck\log(ep/k)$. Then, we apply Birgé's version of Fano's lemma [8] to conclude that:

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

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

Let us turn to the proof of (5.7). Assume that $k\log(ep/k)\leq n/16$ and assume that all the entries of $\mathbf{X}$ follow independent centered normal distribution with variance $1/\sqrt{n}$. Applying Lemma A.2, we derive that:

$$\mathbb{P}\left[\Phi_{2k,-}(\mathbf{X})\leq 1/16\right]\leq e^{-Cn}\ \text{and}\ \mathbb{P}\left[\Phi_{2k,+}(\mathbf{X})\geq 4\right]\leq e^{-Cn}\ .$$

Then, we deduce from (5.6) that for such $\mathbf{X}$,

$$\mathcal{R}_F[k,\mathbf{X}]\geq C\frac{k}{n}\log\left(\frac{ep}{k}\right)\ .$$

Let us now assume that $k\log(ep/k)>n/16$. For n larger than a numerical constant, we can find $k'<k$ and $p'<p$ such that

$$n/32\leq k'\log(ep'/k')\leq n/16\ .$$

Hence, there exists a design $\mathbf{X}'$ of size $n\times p'$ such that

$$\inf_{\widehat{\theta}}\sup_{\theta\in\theta[k',p']}\mathbb{E}\left[\|\mathbf{X}'(\widehat{\theta}-\theta)\|_n^2/n\right]\geq C\sigma^2\ .$$

Then, we only have to take any completion of $\mathbf{X}'$ since

$$\inf_{\widehat{\theta}}\sup_{\theta\in\theta[k',p']}\mathbb{E}\left[\|\mathbf{X}'(\widehat{\theta}-\theta)\|_n^2\right]\leq\inf_{\widehat{\theta}}\sup_{\theta\in\theta[k,p]}\mathbb{E}\left[\|\mathbf{X}(\widehat{\theta}-\theta)\|_n^2\right]\ .$$

9.7. Proof of Proposition 5.6

Let us set $\alpha = \delta = 0.01$. Consider a design $\mathbf{X}$ that achieves the bound (4.13) and take $\rho = \rho^*_{F,U}[k, \mathbf{X}]/2$. If $k\log(p/k)/n$ 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 $\mathbf{H_0}$: "$\theta = 0$" against $\mathbf{H_1}$: "$\theta \in \Theta[k,p]$ and $\|\mathbf{X}\theta\|_n^2/(n\sigma^2) \geq \rho$

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

By Proposition 4.7, we have at least one of the two following properties:

$$\sup_{\sigma>0} \mathbb{P}_{0,\sigma}(T=1) \quad \geq \quad \alpha \tag{9.19}$$

$$\sup_{\sigma>0,\ \theta\in\Theta[k,p],\ \|\mathbf{X}\theta\|_n^2/(n\sigma^2)\geq\rho^2} \mathbb{P}_{\theta,\sigma}(T=0) \quad \geq \quad \delta \tag{9.20}$$

CASE 1: (9.19) holds. With probability larger $1 - e^{-n/16}$, we have $\|\mathbf{Y}\|_n^2 \geq n\sigma^2/2$. 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

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

CASE 2: (9.20) 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 $\|\mathbf{X}\theta\|_n^2/\sigma^2$. By Lemma 1 in Birgé [7], we have $\|\mathbf{Y}\|_n^2 \leq 3/2\left[n\sigma^2 + \|\mathbf{X}\theta\|_n^2\right]$, with probability larger than $1 - e^{-Cn}$. Consequently, there exists $\sigma > 0$ and $\theta \in \Theta[k,p]$ such that $\|\mathbf{X}\theta\|_n^2/(n\sigma^2) \geq \rho^2$ and

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

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

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

9.8. Proof of Proposition 6.1

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 $\theta_1 - \theta_2 = 2\kappa\log(2)\theta/\Phi_{2k,-}(\mathbf{X})$ and $\mathrm{supp}(\theta_1) \cap \mathrm{supp}(\theta_2) = \emptyset$. 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 $\kappa\log(2)$. Applying Corollary 2.19 in [32], we derive the first part of the lower bound:

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

The proof of the second part follows closely the proof of the minimax lower bound for prediction (Proposition 5.3). Given some $r > 0$, we consider the set $\mathcal{C}_k^p(\sqrt{k}r)$ (Definition 6.1). The Kullback discrepansy $\mathcal{K}(\theta_1, \theta_2)$ between any two elements of this set is smaller than $4\Phi_{2k\wedge p,+}(\mathbf{X})r^2$, while the loss $\|\theta_1 - \theta_2\|_p^2$ is lower bounded by $r^2 d_H(\theta_1, \theta_2)$. We recall that $d_H(.)$ is the Hamming distance. As in the proof of Proposition 5.3, we find a subset $\mathcal{C}_k^{'p}(\sqrt{k}r) \subset \mathcal{C}_k^p(\sqrt{k}r)$ whose points are well separated with respect to the Hamming distance. Applying Fano's lemma, we conclude that

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

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

9.9. Proof of Corollary 6.3

Consider a standard Gaussian design $\mathbf{W}$ of size $n \times p$. Rescaling to one each column of $\mathbf{W}$, we get a new design $\mathbf{X}$. If the constant C in the statement of Corollary 6.3 is large enough, we can Apply Lemma A.2 and 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 $\mathbf{X}\theta = \mathbf{W}\theta'/\sqrt{n}$. Moreover we have

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

Hence, we derive that

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

With high probability the $2k$-restricted eigenvalues of $\mathbf{X}$ therefore lie between $1/49$ and 49. Gathering the minimax bounds of Propositions 6.1 and 6.2 allows to conclude.

9.10. Proof of Proposition 6.4

We fix some $1 \geq \delta > 0$. We consider $\mathcal{M}(k, p)$ the collections of subsets of $\{1, \ldots, p\}$ of size k. For any $m \in \mathcal{M}(k, p)$, we define the vector θ_m by $(\theta_m)_i = 1/\sqrt{k}$ if $i \in m$ and 0 else. For any $m \neq m'$, we have $\|\theta_m - \theta_{m'}\|_p^2 \geq 2/k$. If there exists two sets $(m, m') \in \mathcal{M}(k, p)$ such that $\|\mathbf{X}(\theta_m - \theta_{m'})\|_n^2 \leq \delta^2$, then the design $\mathbf{X}$ satisfies $\Phi_{2k,-}(\mathbf{X}) \leq \delta^2 k/2$. A necessary condition for $\mathbf{X}$ to satisfy $\Phi_{2k,-}(\mathbf{X}) \geq \delta^2 k/2$ is therefore that the vectors $\mathbf{X}\theta_m$ are δ-separated. For any $m \in \mathcal{M}(k, p)$, we have $\|\mathbf{X}\theta_m\|_n \leq \sqrt{k}$.

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

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

which allows to prove the first result.

Let us turn to the second result (6.4). By the first result, we have

$$\left(\frac{p_n}{k_n}\right)^{2k_n/n} \overset{\log}{\sim} Ck_n^{-2}\left(\frac{p_n}{k_n}\right)^{2k_n/n} = O\left[\sup_{\mathbf{X}\in\mathcal{D}_{n,p_n}} \Phi^{-1}_{2k_n,-}(\mathbf{X})\right].$$

Comparing this results with the minimax lower bounds and upper bounds (Propositions 6.1 and 6.2), we realize $\mathcal{RI}[k]$ is log-equivalent to $\sup_{\mathbf{X}\in\mathcal{D}_{n,p_n}} \Phi^{-1}_{2k_n,-}(\mathbf{X})$.

In order to finish the proof, it remains to asymptotically upper bound $\sup_{\mathbf{X}\in\mathcal{D}_{n,p_n}} \Phi^{-1}_{2k_n,-}(\mathbf{X})$. Consider a standard Gaussian design $\mathbf{X}_n$ with size $n\times p_n$. Applying the deviation inequality (A.3) of Lemma A.2, we derive that with probability going to one, we have

$$\Phi^{-1}_{2k_n,-}(\mathbf{X}_n/\sqrt{n}) \le C\left(\frac{p_n}{k_n}\right)^{4k_n/n(1+o(1))}\frac{k_n}{n}\log\left(\frac{p_n}{k_n}\right).$$

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

$$\Phi_{2k_n,-}(\mathbf{X}'_n) = \inf_{\theta\in\Theta[k_n,p_n]}\frac{\|\mathbf{X'}_n\theta\|_n^2}{\|\theta\|_{p_n}^2} = \inf_{\theta\in\Theta[k_n,p_n]}\frac{\|\mathbf{X}_n/\sqrt{n}\theta\|_n^2}{\|D_n\theta\|_{p_n}^2} \ge \frac{\Phi_{2x_n,-}(\mathbf{X}_n/\sqrt{n})}{\varphi_{\max}(D_n^2)},$$

Each diagonal element of nD_n^2 follows of χ^2 distribution with n degrees of freedom. Apply Lemma A.1, we derive that $\varphi_{\max}(D_n)\le C\sqrt{1\vee\log(p_n)/n}$ with probability going to one. We conclude that

$$\Phi^{-1}_{2k_n,-}(\mathbf{X}'_n) \le C\left(\frac{p_n}{k_n}\right)^{4k_n/n(1+o(1))}\left[\frac{k_n}{n}\log\left(\frac{p_n}{k_n}\right)\right]^2,$$

with probability going to one. This allows to conclude.

9.11. Proof of Proposition 6.5

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 6.4, 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 $1/\sqrt{k}$.
2. $\mathrm{supp}(\theta_1)\neq\mathrm{supp}(\theta_2)$.
3. $\|\mathbf{X}(\theta_1-\theta_2)\|_n^2 \le (\sqrt{k}+1)^2\binom{p}{k}^{-2/n} := \rho^{*-2}$.

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 discrepansy between $P_{\theta_1^*}$ and $P_{\theta_2^*}$ is smaller than $\log(2)2e/(2e+1)$. Consider an estimator $\widehat{\theta}$ taking its values in $\{\theta_1^*,\theta_2^*\}$. Applying Birgé's lemma or more precisely Corollary 2.18 in [32], we derive that $\inf_{\theta_1^*,\theta_2^*}\mathbb{P}_\theta(\widehat{\theta}=\theta)\le 2e/(2e+1)$. This allows to conclude.

9.12. Proof of Proposition 6.6

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

$$\rho^2 = Ck/(2n)\log(p/k)\exp[Ck/(2n)\log(p/k)]$$

where the constants C 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

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

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 play a symmetric role in the distribution of $(\mathbf{Y},\mathbf{X})$. This estimator $\tilde{\theta}_k$ has the same form as the estimator $\widehat{\theta}$ introduced in Definition 5.1. Arguing as in the proof of Theorem 5.2, we derive that

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

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

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

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 (9.17) and (9.18), we derive that for any estimator $\widehat{\theta}$, there exists $\theta\in\mathcal{C}_{k,1}^p(\rho)\cup\mathcal{C}_{k,2}^p(\rho)$ such that

$$\|\widehat{\theta} - \theta\|_p^2 \geq \frac{\rho^2}{4}\ ,$$

with probability larger than 1/4.

The last lower bound contradicts (9.22) is $\log(p_0)/\log(p)\leq\delta$, where $\delta>0$ is a positive constant.

10. Proofs of the upper bounds

10.1. Analysis of the testing procedures

Proof of Proposition 4.2. By homogeneity, we can assume that $\sigma^2 = 1$. Let us consider a subset $m\subset\{1,\dots,p\}$. We note d_m the rank of the covariance matrix Σ_m of X_m. Under $\mathbf{H_0}$, $\|\Pi_m\mathbf{Y}\|_n^2$

follows a χ^2 distribution with $d_m \leq |m|$ degrees of freedom

$$\mathbb{P}_0\left[\|\Pi_m \mathbf{Y}\|_n^2 > \bar{\chi}_{|m|}^{-1}\left\{\alpha/[2k^*\binom{|m|}{p}]\right\}\right] \leq \alpha/[2k^*\binom{|m|}{p}] \ .$$

Under $\mathbf{H_0}$, $\|\mathbf{Y}\|_n^2$ is χ^2 distribution of size n. If follows that T_α^* is of level α.

Let us turn to the type II error probability. First, take some k between 1 and k^*-1. Consider some $\theta \in \Theta[k,p] \setminus \{0\}$. We call m the support of θ. We can assume that the size of this support is k. If the size of the support is smaller than k, then we only have to complete m to obtain a set of size k.

$$\|\Pi_m \mathbf{Y}\|_n^2 = \|\mathbf{X}\theta + \Pi_m \boldsymbol{\epsilon}\|_n^2$$

The random vector $\boldsymbol{\epsilon}$ is independent of $\mathbf{X}$. Conditionally to $\mathbf{X}$, $\|\Pi_m \mathbf{Y}\|_n^2$ follows a non-centered χ^2 distribution with $d_m \leq k$ degrees of freedom and a non centrality parameter $\|\mathbf{X}\theta\|_n^2$.

Let us denote $Q(a,D,u)$ the $1-u$ quantile of a noncentral χ^2 distribution with D degrees of freedom and noncentrality parameter a. Thanks to Lemma 1 in Birgé [7], we know that the following inequalities hold for all $u \in (0,1)$:

$$Q(a,D,1-u) \geq D + a - 2\sqrt{(D+2a)\log(1/u)} \ .$$

We derive that

$$\begin{aligned} Q\left(\|\mathbf{X}\theta\|_n^2, d_m, 1-\delta/2\right) &\geq d_m + \frac{4}{5}\|\mathbf{X}\theta\|_n^2 - 2\sqrt{d_m \log(2/\beta)} \\ &- 10\log(2/\delta) \ . \end{aligned} \tag{10.1}$$

The variable $\|\mathbf{X}\theta\|_n^2/\|\sqrt{\Sigma}\theta\|_p^2$ follows a χ^2 distribution with n degrees of freedom. With probability $1-\delta/2$, we have $\|\mathbf{X}\theta\|_n^2 \geq n\|\sqrt{\Sigma}\theta\|_p^2[1-2\sqrt{\log(2/\delta)/n}]$ by Lemma A.1. Applying again Lemma A.1, we derive that

$$\bar{\chi}_k^{-1}\left[\alpha/(2k^*\tbinom{k}{p})\right] \leq k + 5k\log\left(\frac{ep}{k}\right) + 3\log\left(\frac{2k^*}{\alpha}\right) \leq C(\alpha)k\log\left(\frac{ep}{k}\right) \ .$$

With probability larger than $1-\delta$, we have

$$\|\Pi_m \mathbf{Y}\|_n^2 - \bar{\chi}_k^{-1}\left[\alpha/(2k^*\tbinom{k}{p})\right] \geq Cn\|\sqrt{\Sigma}\theta\|_p^2 - C(\alpha,\delta)\frac{k}{n}\log\left(\frac{ep}{k}\right) \ ,$$

since we assume that $\log(2/\delta) \leq n/8$. Hence, $T_\alpha^* > 0$ with probability larger than $1-\delta$ as soon as

$$\|\sqrt{\Sigma}\theta\|_p^2 > C(\alpha,\delta)\frac{k}{n}\log\left(\frac{ep}{k}\right) \ .$$

We now consider the case $k \geq k^*$. Consider some $\theta \in \Theta[k,p]$. Arguing as previously, we derive that

$$\|\mathbf{Y}\|_n^2 - \bar{\chi}_n^{-1}(\alpha/2) \geq C\|\mathbf{X}\theta\|_n^2 - C(\alpha,\delta)\frac{1}{\sqrt{n}} \ ,$$

with probability larger than $1-\delta$. □

Proof of Proposition 4.4. We argue as in the proof of the previous proposition, the only difference being that the design $\mathbf{X}$ is now fixed.

Let us consider a subset $m \subset \{1,\dots,p\}$. We note d_m, the rank of the $n \times |m|$ submatrix $\mathbf{X}_m$ of $\mathbf{X}$. Under $\mathbf{H_0}$, $\|\Pi_m \mathbf{Y}\|_n^2$ follows a χ^2 distribution with $d_m \leq |m|$ degrees of freedom. Then, arguing as in the proof of Proposition 4.2, we derive that the level of T_α^* is smaller than α.

Let us turn to the type II error probability. Take some k between 1 and k^*-1 and consider some $\theta \in \Theta[k,p] \setminus \{0\}$. We call m the support of θ. We can assume that the size of this support is k. Then, $\|\Pi_m \mathbf{Y}\|_n^2$ follows a χ^2 distribution with d_m degrees of freedom and a non centrality parameter $\|\mathbf{X}\theta\|_n^2$. Arguing as in the proof of Proposition 4.2, we derive that

$$\|\Pi_m \mathbf{Y}\|_n^2 - \bar{\chi}_k^{-1}\left[\alpha/(2k^*\tbinom{k}{p})\right] \geq C\|\mathbf{X}\theta\|_p^2 - C(\alpha,\delta)\frac{k}{n}\log\left(\frac{ep}{k}\right) \ ,$$

with probability larger than $1-\delta$. Then, we conclude as in the proof of Proposition 4.2.

□

Proof of Proposition 4.6. A similar procedure has been studied in Theorem 3.3 in [38], the main difference being that Σ was assumed to be non singular and that the ultra-high dimension setting was not taken into account.

Consider a subset $m \subset \{1,\dots,p\}$ whose size is smaller than n. We note d_m the rank of the covariance matrix Σ_m of X_m. Almost surely, we have $d_m(\mathbf{X}) = d_m$. Under $\mathbf{H_0}$, conditionally to $\mathbf{X}$, $\phi_m(\mathbf{X},\mathbf{Y})$ follows a Fisher distribution with d_m and $n-d_m$ degrees of freedom. Integrating with respect to $\mathbf{X}$, we derive that $\phi_m(\mathbf{X},\mathbf{Y})$ still follows a Fisher distribution with d_m and $n-d_m$ degrees of freedom. As a consequence, T_α is a Bonferroni multiple testing procedure based on Fisher statistics. It follows that T_α has level α.

Let us turn to the type II error probability. We mainly follow the steps the proof of Theorem 3.3 of [38], since their specific assumptions of Theorem 3.3 are only used in the last three lines of their proof.

Fix some $1 \leq k \leq n/2$. Observe that $\log[\tbinom{k}{p}] \leq k\log(ep/k)$. Consider a set $m \in \mathcal{M}(k,p)$. Adapting Eq.(7.5), (7.6), and (7.9) in [38] and using the inequality $d_m \leq k$, we find that the test T_α is rejected with probability larger than $1-\delta$ as soon as

$$\frac{\mathrm{Var}(Y) - \mathrm{Var}(Y|X_m)}{\mathrm{Var}(Y|X_m)} \geq \frac{\bar{\Delta}_m(\delta)}{n\left(1-2\sqrt{\log(2/\delta)/n}\right)} \ , \tag{10.2}$$

where $\bar{\Delta}_m(\delta)$ is defined as follows

$$\bar{\Delta}_m(\delta) := 2.5\sqrt{1+K_m^2(U)}\sqrt{k\log\left(\frac{4}{\alpha_m\delta}\right)}\left(1+\sqrt{\frac{k}{n-k}}\right) + 2.5\left[k_m K_m(U) \vee 5\right]\log\left(\frac{4}{\alpha_m\delta}\right)\left(1+\frac{2k}{n-k}\right),$$

where $\alpha_m = \alpha/(\lfloor n/2 \rfloor |\mathcal{M}(k,p)|)$, $U := \log(2/\delta)$,

$$\begin{aligned} k_m &:= 2\exp\left[4\frac{k}{n-k}\log(ep/k) + 4\frac{\log(1/\alpha)+\log(n)}{n-k}\right] \ , \\ K_m(U) &:= 1 + 2\sqrt{\frac{U}{n-k}} + 2k_m\frac{U}{n-k} \ . \end{aligned}$$

Since $k \leq n/2$, $\log(2/\delta) \leq n/8$, $p \geq n$, we get

$$\bar{\Delta}_m(\delta) \quad \leq \quad C(\alpha,\delta) k \log\left(\frac{ep}{k}\right) \exp\left(C_2(\alpha,\delta)\frac{k\log(ep/k)}{n}\right) .$$

Now, we take m to be the support of θ. Gathering (10.2) with this last equation, we derive that T_α is rejected with probability larger than $1-\delta$ if

$$\frac{\|\sqrt{\Sigma}\theta\|_p^2}{\mathrm{Var}(Y|X)} \geq C(\alpha,\delta)\frac{k}{n}\log\left(\frac{ep}{k}\right)\exp\left(C_2(\alpha,\delta)\frac{k\log(ep/k)}{n}\right) .$$

□

Proof of Proposition 4.8. The testing procedure has already been studied in Baraud et al. [5]. They proved that the size T_α is less than α.

Consider some $k \leq n/2$ and $\theta \in \Theta[k,p]$. Applying Theorem 1 in [5], we derive that the test rejects H_0 with probability larger than $1-\delta$ if

$$\begin{aligned}\frac{\|\mathbf{X}\theta\|_n^2}{\sigma^2} &\geq C(\alpha,\delta)k\log\left(\frac{2ep}{k\alpha\delta}\right)\exp\left[C\frac{k}{n}\log\left(\frac{ep}{k\alpha}\right)\right]\left[1\vee\exp\left[4\frac{k}{n}\log\left(\frac{ep}{k\alpha}\right)\right]\frac{\log(2/\delta)}{n}\right]\\ &\geq C(\alpha,\delta)k\log\left(\frac{ep}{k}\right)\exp\left[C(\alpha,\delta)\frac{k}{n}\log\left(\frac{ep}{k}\right)\right] .\end{aligned}$$

□

10.2. Proof of Theorem 5.2

Proof of Theorem 5.2. We note $\mathcal{M}$ the collection of subsets of $\{1,\dots,p\}$ of size smaller than $(n-1)/4$. For any $m \in \mathcal{M}$, $\widehat{\theta}_m$ refers to the least-squares estimator of θ whose support is included in m. We have

$$\widehat{m}^V \in \arg\min_{m\in\mathcal{M}} \|\mathbf{Y}-\mathbf{X}\widehat{\theta}_m\|_n^2\left[1+\mathrm{pen}'(|m|)\right] ,$$

where $\mathrm{pen}'(|m|) = -1+\exp[Kk/n\log(ep/k)]$. In the following, θ_m refers to the best approximation of θ whose support is included in m with respect to the distance $\|\sqrt{\Sigma}(.-\theta)\|_p$. We shall prove a stronger result than (5.5):

Proposition 10.1. *Assume that $n \geq C$. There exists a universal choice of K in the penalty (5.4) such that the following holds. For any covariance Σ and any $\theta \in \mathbb{R}^p$, we have*

$$\mathbb{E}\left[\|\sqrt{\Sigma}(\widetilde{\theta}^V-\theta)\|_p^2\right] \quad \leq \quad C(K)\inf_{m\in\mathcal{M}}\left[\|\sqrt{\Sigma}(\theta_m-\theta)\|_p^2\{1+\mathrm{pen}'(|m|)\}+\sigma^2\left\{\mathrm{pen}'(|m|)\vee\frac{1}{n}\right\}\right] \quad (10.3)$$

Take any $\theta \in \Theta[k,p]$. Consider the risk bound (10.3) with the set $m = \mathrm{supp}(\theta)$. We obtain

$$\mathbb{E}\left[\|\sqrt{\Sigma}(\widetilde{\theta}^V-\theta)\|_p^2\right] \leq C(K)\frac{k}{n}\log\left(\frac{ep}{k}\right)\exp\left\{C\frac{k}{n}\log\left(\frac{ep}{k}\right)\right\}\sigma^2 .$$

□

Proof of Proposition 10.1. Given a subset $m \in \mathcal{M}$, Σ_m denotes the covariance matrix of X_m. We can assume that for all $m \in \mathcal{M}$, the covariance matrices Σ_m are non-singular.

If this is not the case, we can define the subcollection $\mathcal{M}' \subset \mathcal{M}$, which contains all subsets m such that Σ_m is non-singular. We have

$$\widehat{m}^V \in \arg\min_{m\in\mathcal{M}'} \|\mathbf{Y} - \mathbf{X}\widehat{\theta}_m\|_n^2 \left[1 + \mathrm{pen}'(|m|)\right] \ a.s. \ .$$

Thus, $\tilde{\theta}^V$ can be analyzed using the collection $\mathcal{M}'$ instead of $\mathcal{M}$.

Let us fix a set $m \in \mathcal{M}$. By definition of $\widehat{m}$, we have

$$\|\mathbf{Y} - \mathbf{X}\widehat{\theta}_{\widehat{m}}\|_n^2 \left[1 + \mathrm{pen}'(|\widehat{m}|)\right] \leq \|\mathbf{Y} - \mathbf{X}\theta_m\|_n^2 [1 + \mathrm{pen}'(|m|)] \tag{10.4}$$

As in [37], we define the random variable ϵ_m by

$$Y = X\theta_m + \epsilon_m + \epsilon \ \text{ a.s. } .$$

By definition of θ_m, ϵ_m is independent of X_m and follows a centered normal distribution with variance $\|\sqrt{\Sigma}(\theta_m - \theta)\|_p^2$. Consider some $0 < u \leq 1$, which will be fixed later. For any subset m and any vector Z of size n, $\Pi_m^\perp Z$ stands for $Z - \Pi_m Z$. Since $\|\mathbf{Y} - \mathbf{X}\widehat{\theta}_{\widehat{m}}\|_n^2 = \|\Pi_{\widehat{m}}^\perp \boldsymbol{\epsilon} + \boldsymbol{\epsilon}_{\widehat{m}}\|_n^2$, we derive from (10.4) that

$$\begin{aligned} u\|\sqrt{\Sigma}(\theta - \widehat{\theta}_{\widehat{m}})\|_p^2 &\leq \|\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_m\|_n^2/n[1 + \mathrm{pen}'(|m|)] + u\left[\|\sqrt{\Sigma}(\theta_{\widehat{m}} - \theta)\|_p^2 + \|\sqrt{\Sigma}(\theta_{\widehat{m}} - \widehat{\theta}_{\widehat{m}})\|_p^2\right] \\ &- \|\Pi_{\widehat{m}}^\perp \boldsymbol{\epsilon} + \boldsymbol{\epsilon}_{\widehat{m}}\|_n^2/n[1 + \mathrm{pen}'(|\widehat{m}|)] \ . \end{aligned}$$

Then, we get

$$\begin{aligned} u\|\sqrt{\Sigma}(\theta - \widehat{\theta}_{\widehat{m}})\|_p^2 &\leq \left[2\langle \boldsymbol{\epsilon}, \boldsymbol{\epsilon}_m\rangle_n + \|\boldsymbol{\epsilon}_m\|_n^2\right]/n(1 + \mathrm{pen}'(|m|)) + \|\boldsymbol{\epsilon}\|_n^2 \mathrm{pen}'(|m|)/n \\ &+ \frac{\|\boldsymbol{\epsilon}\|_n^2}{n} + u\left[\|\sqrt{\Sigma}(\theta_{\widehat{m}} - \theta)\|_p^2 + \|\sqrt{\Sigma}(\theta_{\widehat{m}} - \widehat{\theta}_{\widehat{m}})\|_p^2\right] - \frac{\|\Pi_{\widehat{m}}^\perp \boldsymbol{\epsilon} + \boldsymbol{\epsilon}_{\widehat{m}}\|_n^2}{n}[1 + \mathrm{pen}'(|\widehat{m}|)] \ . \end{aligned}$$

We call the first line A_m and the second line $B_{\widehat{m}}$. We shall prove that for a good choice of u, $B_{\widehat{m}}$ is non-positive with large probability. First, we provide a control of A_m with large probability.

Lemma 10.2. *With probability larger than $1 - 3e^{-x}$, we have*

$$A_m \leq C\|\sqrt{\Sigma}(\theta_m - \theta)\|_p^2 [1 + \mathrm{pen}'(|m|)] \left(1 + \frac{x}{n}\right) + C\sigma^2 \frac{1 \vee x}{n} + C\sigma^2 \mathrm{pen}'(|m|) \left(1 + \frac{x}{n}\right) \ . \tag{10.5}$$

Let us consider a partition $(\mathcal{M}_1, \mathcal{M}_2)$ of $\mathcal{M}$. such that the collection $\mathcal{M}_1$ contains all the sets m such that $|m| \log(ep/|m|) \leq n/16$. One of the collections $\mathcal{M}_1$ or $\mathcal{M}_2$ is possibly empty. We shall first upper bound the loss $\|\sqrt{\Sigma}(\widetilde{\theta} - \theta)\|_p^2 \mathbf{1}_{\widehat{m}\in\mathcal{M}_2}$ and then $\|\sqrt{\Sigma}(\widetilde{\theta} - \theta)\|_p^2 \mathbf{1}_{\widehat{m}\in\mathcal{M}_1}$.

Lemma 10.3. *Let us define $V_{\widehat{m}} = |\widehat{m}|/n \log(ep/|\widehat{m}|)$. For any $x > 0$ and any $0 < u \leq 1$, we have*

$$\begin{aligned} B_{\widehat{m}} &\leq \sigma^2\left(3/2 + 4\frac{x}{n}\right) + u\|\sqrt{\Sigma}(\theta_{\widehat{m}} - \theta)\|_p^2 \\ &+ \left[\sigma^2 + \|\sqrt{\Sigma}(\theta_{\widehat{m}} - \theta)\|_p^2\right]\left[uC_1 \exp\left[C_2 V_{\widehat{m}}\right] e^{C_3 x/n} - C_4 \exp\left[(K - C_5)V_{\widehat{m}}\right] e^{-C_6 x/n}\right] \ , \end{aligned}$$

with probability larger than $1 - Ce^{-x}$.

CASE 1. **Large sets:** $\widehat{m} \in \mathcal{M}_2$. First, we apply Lemma 10.3 with $x = n$ and $u = C_4/(2C_1)e^{-C_6-C_3} \wedge 1$. If we constrain K to satisfy $K \geq C_2 + C_5$, we have

$$\frac{B_{\widehat{m}}}{\sigma^2 + \|\sqrt{\Sigma}(\theta_{\widehat{m}} - \theta)\|_p^2} \mathbf{1}_{\widehat{m} \in \mathcal{M}_2} \leq \left[\frac{11}{2} - C_4 \exp\left[(K - C_5)V_{\widehat{m}}\right] e^{C_6} \right]_+ ,$$

which is zero for K large enough since $V_{\widehat{m}} \geq 1/16$ for $\widehat{m} \in \mathcal{M}_2$. Combining this result with (10.5), we obtain

$$\|\sqrt{\Sigma}(\widetilde{\theta} - \theta)\|_p^2 \mathbf{1}_{\widehat{m} \in \mathcal{M}_2} \leq C \left[\|\sqrt{\Sigma}(\theta_m - \theta)\|_p^2 (1 + \mathrm{pen}'(|m|)) + \sigma^2 \mathrm{pen}'(|m|) \right] + \sigma^2 \frac{1 \vee x}{n} , \quad (10.6)$$

with probability larger than $1 - Ce^{-n} - Ce^{-x}$.

CASE 2. Small sets: $\widehat{m} \in \mathcal{M}_1$. In this case, we use a slightly different decomposition for $B_{\widehat{m}}$:

$$\begin{aligned} B_{\widehat{m}} &\leq u\|\sqrt{\Sigma}(\theta_{\widehat{m}} - \theta)\|_p^2 - \frac{\|\Pi_{\widehat{m}}^{\perp} \boldsymbol{\epsilon}_{\widehat{m}}\|_n^2}{2n} \\ &+ u\|\sqrt{\Sigma}(\widetilde{\theta} - \theta_{\widehat{m}})\|_p^2 + \frac{\|\Pi_{\widehat{m}} \boldsymbol{\epsilon}\|_n^2}{n} + \frac{2}{n} \langle \Pi_{\widehat{m}}^{\perp} \boldsymbol{\epsilon}, \frac{\Pi_{\widehat{m}}^{\perp} \boldsymbol{\epsilon}_{\widehat{m}}}{\|\Pi_{\widehat{m}}^{\perp} \boldsymbol{\epsilon}_{\widehat{m}}\|_n} \rangle_n^2 - \frac{\|\Pi_{\widehat{m}}^{\perp} \boldsymbol{\epsilon}_{\widehat{m}} + \boldsymbol{\epsilon}\|_n^2}{n} \mathrm{pen}'(|\widehat{m}|) . \end{aligned}$$

Lemma 10.4. *Let us constrain $K > C$ and $0 < u < C(K)$ (explained in the proof). For any $x \leq n$, we have*

$$B_{\widehat{m}} \mathbf{1}_{\widehat{m} \in \mathcal{M}_1} \leq C \frac{x}{n} \sigma^2 ,$$

with probability larger than $1 - Ce^{-x} - Ce^{-Cn}$.

We also bound the term A_m in probability by Lemma 10.2. All in all, we get that for any $x < n$

$$\|\sqrt{\Sigma}(\widetilde{\theta} - \theta)\|_p^2 \mathbf{1}_{\widehat{m} \in \mathcal{M}_1} \leq C(K) \left[\|\sqrt{\Sigma}(\theta_m - \theta)\|_p^2 (1 + \mathrm{pen}'(|m|)) + \sigma^2 \mathrm{pen}'(|m|) + \sigma^2 \frac{1 \vee x}{n} \right] , \quad (10.7)$$

with probability larger than $1 - Ce^{-Cn} - Ce^{-x}$. Gathering (10.6) and (10.7), we derive that for any $x < n$

$$\|\sqrt{\Sigma}(\widetilde{\theta} - \theta)\|_p^2 \leq C(K) \left[\|\sqrt{\Sigma}(\theta_m - \theta)\|_p^2 (1 + \mathrm{pen}'(|m|)) + \sigma^2 \mathrm{pen}'(|m|) + \sigma^2 \frac{1 \vee x}{n} \right] , \quad (10.8)$$

with probability larger than $1 - Ce^{-Cn} - Ce^{-x}$.

Control of the tail. Consider Lemma 10.3 with any $x > n$ and take u defined by

$$u = C_4/[2(C_1 \vee 1)] \exp\left[-(C_3 + C_6)x/n\right] \wedge 1 .$$

Constraining $K \geq C_2 + C_5$ yields

$$B_{\widehat{m}} \leq C\sigma^2 \left(1 + \frac{x}{n}\right) ,$$

with probability larger than $1-Ce^{-x}$. Gathering this bound with Lemma 10.2, we derive that for any $x > n$

$$\|\sqrt{\Sigma}(\widetilde{\theta}-\theta)\|_p^2 \leq C\exp\left[Cx/n\right]\left[\|\sqrt{\Sigma}(\theta_m-\theta)\|_p^2+\sigma^2\right](1+\mathrm{pen}'(|m|)) \;, \tag{10.9}$$

with probability larger than $1-Ce^{-x}$. Integrating the upper bounds (10.8) and (10.9), we conclude

$$\mathbb{E}\left[\|\sqrt{\Sigma}(\widetilde{\theta}-\theta)\|_p^2\right] \leq C(K)\left[\|\sqrt{\Sigma}(\theta_m-\theta)\|_p^2(1+\mathrm{pen}'(|m|))+\sigma^2\mathrm{pen}'(|m|)\right]+C(K)\frac{\sigma^2}{n} \;.$$

Taking the minimum over all $m \in \mathcal{M}$ concludes the proof. □

Proof of Lemma 10.2.

$$A_m \leq \left[2\|\boldsymbol{\epsilon}_m\|_n^2\right]/n[1+\mathrm{pen}'(|m|)]+\|\boldsymbol{\epsilon}\|_n^2\mathrm{pen}'(|m|)/n+2\langle\boldsymbol{\epsilon},\frac{\boldsymbol{\epsilon}_m}{\|\boldsymbol{\epsilon}_m\|_n}\rangle_n^2/n[1+\mathrm{pen}'(|m|)] \;.$$

The three random variables $\|\boldsymbol{\epsilon}_m\|_n^2/\|\sqrt{\Sigma}(\theta_m-\theta)\|_p^2$, $\|\boldsymbol{\epsilon}\|_n^2/\sigma^2$ and $\langle\boldsymbol{\epsilon},\frac{\boldsymbol{\epsilon}_m}{\|\boldsymbol{\epsilon}_m\|_n}\rangle_n^2/\sigma^2$ respectively follow χ^2 distributions with n, n and 1 degrees of freedom. Applying the deviation inequalities of Lemma A.1 allows to conclude.

□

Proof of Lemma 10.3. For any subset m of $\{1,\dots,p\}$, we recall that Σ_m denotes the covariance matrix of the vector X_m^*. Moreover, we define the row vector $Z_m := X_m\sqrt{\Sigma_m^{-1}}$ in order to deal with standard Gaussian vectors. Similarly to the matrix $\mathbf{X}_m$, the $n \times d_m$ matrix $\mathbf{Z}_m$ stands for the n observations of Z_m. By Lemma 7.1 in Verzelen [37], $\|\sqrt{\Sigma}(\widehat{\theta}_{\widehat{m}}-\theta_{\widehat{m}})\|_p^2$ decomposes as

$$\begin{aligned}\|\sqrt{\Sigma}(\widehat{\theta}_{\widehat{m}}-\theta_{\widehat{m}})\|_p^2 &= (\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}})^*\mathbf{Z}_{\widehat{m}}(\mathbf{Z}_{\widehat{m}}^*\mathbf{Z}_{\widehat{m}})^{-2}\mathbf{Z}_{\widehat{m}}^*(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}}) \\ &\leq \varphi_{\min}^{-1}\left[\mathbf{Z}_{\widehat{m}}^*\mathbf{Z}_{\widehat{m}}\right]\|\Pi_{\widehat{m}}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}})\|_n^2 \;. \end{aligned}\tag{10.10}$$

Moreover, the random variables $\|\Pi_{m'}\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'}\|_n^2/(\sigma^2+\|\sqrt{\Sigma}(\theta_{m'}-\theta)\|_p^2)$ and $\|\Pi_{m'}^{\perp}\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'}\|_n^2/(\sigma^2+\|\sqrt{\Sigma}(\theta_{m'}-\theta)\|_p^2)$ respectively follow χ^2 distributions with $|m'|$ and $n-|m'|$ degrees of freedom.

By Lemma A.1, we have

$$\|\boldsymbol{\epsilon}_n\|_n^2/\sigma^2 \leq 3/2+4x/n \;, \tag{10.11}$$

with probability larger than $1-e^{-x}$. We recall that the maximal size of any subset $m \in \mathcal{M}$ is smaller than $n/2$. Applying the deviation inequality (A.1) and Lemma A.2, we get for any set $m' \in \mathcal{M}$ and for any $x > 0$,

$$\frac{\|\Pi_{m'}^{\perp}\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'}\|_n^2}{n\left(\sigma^2+\|\sqrt{\Sigma}(\theta_{m'}-\theta)\|_p^2\right)} \geq C_1\exp\left[-C_2\frac{|m'|\log(ep/|m'|)}{n}\right]\exp(-C_3x/n) \;, \tag{10.12}$$

$$\varphi_{\min}^{-1}\left[\mathbf{Z}_{m'}^*\mathbf{Z}_{m'}/n\right] \leq C_1\exp\left[C_2\frac{|m'|\log(ep/|m'|)}{n}\right]\exp(C_3x/n) \;, \tag{10.13}$$

$$\frac{\|\Pi_{m'}\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'}\|_n^2}{n\left(\sigma^2+\|\sqrt{\Sigma}(\theta_{m'}-\theta)\|_p^2\right)} \leq C_1\frac{|m'|}{n}\log\left(\frac{ep}{|m'|}\right)+\frac{C_2x}{n} \;, \tag{10.14}$$

with probability larger than $1-3\binom{|m'|}{p}^{-1}e^{-|m'|}e^{-x}$. We derive from (10.10), (10.13) and (10.14), that

$$\|\sqrt{\Sigma}(\widehat{\theta}_{\widehat{m}}-\theta_{\widehat{m}})\|_p^2 \leq \left[\sigma^2+\|\sqrt{\Sigma}(\theta_{\widehat{m}}-\theta)\|_p^2\right] C_1 \exp\left[C_2\frac{|\widehat{m}|}{n}\log\left(\frac{ep}{|\widehat{m}|}\right)\right]\exp\left(C_3x/n\right) , \quad (10.15)$$

with probability larger than $1-Ce^{-x}$. We derive from (10.12) and the definition (5.4) that

$$\frac{\|\Pi^{\perp}_{\widehat{m}}\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}}\|_n^2}{n}\left[1+\mathrm{pen}'(|\widehat{m}|)\right] \geq \left[\sigma^2+\|\sqrt{\Sigma}(\theta_{\widehat{m}}-\theta)\|_p^2\right]\exp(-C_6x/n) \times C_4\exp\left[(K-C_5)\frac{|\widehat{m}|}{n}\log\left(\frac{ep}{|\widehat{m}|}\right)\right] , \quad (10.16)$$

with probability larger than $1-Ce^{-x}$. Gathering the bound (10.11), (10.15) and (10.16) allows to conclude.

□

Proof of Lemma 10.4. Applying deviations inequalities for χ^2 distributions (Lemma A.1) to all $m\in\mathcal{M}_1$ we prove

$$\frac{\|\Pi^{\perp}_{\widehat{m}}\boldsymbol{\epsilon}_{\widehat{m}}+\boldsymbol{\epsilon}\|_n^2}{n} \geq C(\sigma^2+\|\sqrt{\Sigma}(\theta_{\widehat{m}}-\theta)\|_p^2) \quad (10.17)$$

$$\frac{\|\Pi^{\perp}_{\widehat{m}}\boldsymbol{\epsilon}_{\widehat{m}}\|_n^2}{n} \geq C\|\sqrt{\Sigma}(\theta_{\widehat{m}}-\theta)\|_p^2 , \quad (10.18)$$

with probability larger than $1-2e^{-Cn}$. We also apply the deviation inequality (A.2) for the largest eigenvalue of an inverse Wishart matrix and we get:

$$\varphi^{-1}_{\min}\left[\mathbf{Z}^*_{\widehat{m}}\mathbf{Z}_{\widehat{m}}/n\right] \leq C ,$$

with probability larger than $1-e^{-Cn}$. Applying again deviation inequalities for χ^2 random variables, we get

$$\frac{\|\Pi_{\widehat{m}}\boldsymbol{\epsilon}\|_n^2}{n}+\frac{2}{n}\langle\Pi^{\perp}_{\widehat{m}}\boldsymbol{\epsilon},\frac{\Pi^{\perp}_{\widehat{m}}\boldsymbol{\epsilon}_{\widehat{m}}}{\|\Pi^{\perp}_{\widehat{m}}\boldsymbol{\epsilon}_{\widehat{m}}\|_n}\rangle_n^2 \leq C\frac{|\widehat{m}|}{n}\log\left(\frac{ep}{|\widehat{m}|}\right)\sigma^2+C\frac{x}{n}\sigma^2 ,$$

with probability larger than $1-Ce^{-x}$. The two previous bounds allow to prove that

$$\|\sqrt{\Sigma}(\widetilde{\theta}-\theta_{\widehat{m}})\|_p^2 \leq C\left[\frac{|\widehat{m}|}{n}\log\left(\frac{ep}{|\widehat{m}|}\right)+\frac{x}{n}\right]\left[\sigma^2+\|\sqrt{\Sigma}(\theta_{\widehat{m}}-\theta)\|_p^2\right] ,$$

with probability larger than $1-Ce^{-x}-Ce^{-Cn}$.

Gathering (10.17) and (10.18) with the last bound, we derive that for any $x>0$,

$$\begin{aligned} B_{\widehat{m}} &\leq \sigma^2\left[C_1\frac{|\widehat{m}|}{n}\log\left(\frac{ep}{|\widehat{m}|}\right)+C_2\frac{x}{n}-C_3\mathrm{pen}'(|\widehat{m}|)\right] \\ &+ \|\sqrt{\Sigma}(\theta_{\widehat{m}}-\theta)\|_p^2\left[u\left(1+C_4\frac{|\widehat{m}|}{n}\log\left(\frac{ep}{|\widehat{m}|}\right)+C_5\frac{x}{n}\right)-C_6\{1+\mathrm{pen}'(|\widehat{m}|)\}\right] , \end{aligned}$$

with probability larger $1-Ce^{-x}-Ce^{-Cn}$. Since $e^x\geq 1+x$, we have $\mathrm{pen}'(|\widehat{m}|)\geq K|\widehat{m}|\log(ep/|\widehat{m}|)$. It follows that if we take K larger than some numerical constant, $x\leq n$ and u smaller than some constant $C(K)$, we have $B_{\widehat{m}}\leq Cx/n\sigma^2$, with probability larger than $1-Ce^{-x}-Ce^{-Cn}$. □

10.3. Proof of Proposition 5.5

Given $m \subset \{1,\dots,p\}$, we write $\widehat{\theta}_m$, the least-squares estimator of θ whose support is included in m. For any subset m, we note d_m the rank of the subdesign $\mathbf{X}_m$ of size $n \times |m|$. Consider the collection of $\mathcal{M}'$ of non-empty subsets m such that $|m| \leq (n-1)/4$. Upon defining the penalty $\mathrm{pen}'(m)$ by the identity

$$1 + \frac{\mathrm{pen}'(m)}{n - d_m} = \exp[\mathrm{pen}(|m|)] \ ,$$

we have almost surely

$$\widehat{m}^V \in \arg\min_{m \in \mathcal{M}'} \|\mathbf{Y} - \mathbf{X}\widehat{\theta}_m\|_n^2 \left[1 + \frac{\mathrm{pen}'(m)}{n - d_m}\right] \ .$$

Let us consider the penalty $\mathrm{pen}_{2,\mathcal{L}}\colon \mathcal{M}' \to \mathbb{R}^+$ introduced in Definition 3 in [4] with $L_m = |m| + \log[\binom{|m|}{p}]$. By Proposition 4 in [4], we have for any $|m| \in \mathcal{M}'$,

$$\begin{aligned}
\mathrm{pen}_{2,\mathcal{L}}(m) &\leq C|m| \left[1 + \sqrt{C \log\left(\frac{ep}{|m|}\right)} \exp\left\{C\frac{|m|}{n}\log(ep/|m|)\right\}\right]^2 \\
&\leq C|m| \log\left(\frac{ep}{|m|}\right) \exp\left\{C\frac{|m|}{n}\log(ep/|m|)\right\} \ .
\end{aligned}$$

If we choose K large enough in the penalty function (5.4), then we have for any $m \in \mathcal{M}'$, $\mathrm{pen}'(m) \geq \mathrm{pen}_{2,\mathcal{L}}(m)$. By Theorem 2 and Proposition 3 in Baraud et al. [4], we obtain that for any $1 \leq k \leq (n-1)/4$ and any $\theta \in \Theta[k,p]$,

$$\mathbb{E}\left[\|\mathbf{X}(\tilde{\theta}^V - \theta)\|_n^2/(n\sigma^2)\right] \leq C\frac{\mathrm{pen}'(k)}{n} + \frac{C}{n} \leq C(K)\frac{k}{n}\log\left(\frac{ep}{k}\right)\exp\left\{C\frac{k}{n}\log(ep/k)\right\} \ .$$

10.4. Proof of Proposition 6.2

This result is a again a consequence of Theorem 1 in Birgé and Massart [10]. We have

$$\mathbb{E}\left[\|\mathbf{X}(\widehat{\theta}_k - \theta)\|_n^2\right] \leq Ck\log\left(ep/k\right)\sigma^2 \ .$$

We conclude by using the fact that $\widehat{\theta}_k$ and θ are k-sparse.

Appendix A: Concentration inequalities

Lemma A.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 $0 < x < 1$

$$\mathbb{P}\left[\chi^2(d) \leq dCx^{2/d}\right] \leq x \ , \tag{A.1}$$

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

Proof of Lemma A.1. The two first bounds are classical and are shown by Laplace method. We refer to Lemma 1 in [29] for more details. We only provide a proof for the third bound (A.1). Consider some $0 \leq u < 1$ and some $\lambda > 0$. We write X_d for a random variable that follows a $\chi^2(d)$ distribution. Applying Laplace method, we get

$$\mathbb{P}\left[X_d \leq ud\right] \leq \mathbb{E}\left[e^{-\lambda X_d} e^{\lambda ud}\right] = \exp\left[\lambda ud\right] \exp\left[-\frac{d}{2}\log\left(1+2\lambda\right)\right] .$$

Taking $\lambda = (1-u)/(2u)$ leads to

$$\mathbb{P}\left[X_d \leq ud\right] \leq \exp\left[\frac{d}{2}\left(1-u+\log(u)\right)\right] \leq \exp\left[\frac{d}{2}\log(eu)\right] .$$

Consider some $0 < x < 1$. Then $x^{2/d}/e$ is smaller than one and we get

$$\mathbb{P}\left[X_d \leq \frac{d}{e}x^{2/d}\right] \leq x .$$

□

Lemma A.2 (Wishart distributions)**.** *Let Z^*Z be a standard Wishart matrix of parameters (n,d) with $n > d$. For any number $0 < x < 1$,*

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

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

$$\mathbb{P}\left[\varphi_{\min}\left(Z^*Z\right) \leq nCx^{\frac{2}{n-2d}}\left[1 \vee \frac{\log(2/x)}{n}\right]^{-1}\right] \leq x , \tag{A.3}$$

where C is a numerical constant.

The two first deviation inequalities are taken from Theorem 2.13 in [17]. The bound (A.3) allows to control the tail of the distribution of the largest eigenvalue of a Wishart distribution. Rudelson and Vershynin [36] have provided a control similar to (A.3) under subgaussian assumptions. However, their results only holds for events of probability smaller than $1 - e^{-n}$.

Proof of Lemma A.2. In this proof, we adopt the same approach as Litvak et al. [30]. Fix some $x > 0$ and some $\epsilon > 0$. We consider a minimal ϵ-net $\mathcal{N}(\epsilon)$ of the l_2 unit sphere in $\mathbb{R}^d$, whose elements are included in the unit sphere. Since the ϵ-covering number of the ball in $\mathbb{R}^d$ is smaller than $(5/\epsilon)^d$, we get

$$\left|\mathcal{N}(\epsilon)\right| \leq (10/\epsilon)^d . \tag{A.4}$$

Consider some $\theta \in \mathcal{N}(\epsilon)$. The random variable $\|Z\theta\|_n^2$ follows a χ^2 distribution with n degrees of freedom. Applying Lemma A.1, we derive that for any $0 < x < 1$

$$\mathbb{P}\left[\|Z\theta\|_n^2 \leq Cnx^{2/n}\right] \leq x .$$

Applying an union bound over all $\theta \in \mathcal{N}(\epsilon)$, we derive that

$$\mathbb{P}\left[\inf_{\theta\in\mathcal{N}(\epsilon)} \|Z\theta\|_n^2 \leq nCx^{2/n}\left(\frac{\epsilon}{10}\right)^{\frac{2d}{n}}\right] \leq x/2 \ .$$

Applying the first result of Lemma A.2, we control the largest singular value of Z.

$$\mathbb{P}\left[\sup_{\theta\in\mathbb{R}^d \ , \ \|\theta\|_d^2\leq 1} \|Z\theta\|_n^2 \leq n\left\{1+\sqrt{\frac{k}{n}}+\sqrt{\frac{2\log(2/x)}{n}}\right\}^2\right] \leq x/2 \ .$$

Consider any $\theta \in \mathbb{R}^d$ such that $\|\theta\|_d = 1$. There exists $\theta_0 \in \mathcal{N}(\epsilon)$ such that $\|\theta - \theta_0\|_d \leq \epsilon$.

$$\begin{aligned}\|Z\theta\|_n^2 &= \|Z\theta_0\|_n^2 + < Z(\theta-\theta_0), Z(\theta+\theta_0) >_n \\ &\geq \inf_{\theta_0\in\mathcal{N}(\epsilon)} \|Z\theta_0\|_n^2 - \sqrt{2}\epsilon \sup_{\theta, \ \|\theta\|_d^2=1} \|Z\theta\|_n^2 \ .\end{aligned}$$

Applying the two deviations inequalities, we derive that for any $\epsilon > 0$

$$\inf_{\theta, \ \|\theta\|_d^2=1} \|Z\theta\|_n^2/n \geq Cx^{2/n}\left(\frac{\epsilon}{10}\right)^{\frac{2d}{n}} - \sqrt{2}\epsilon\left\{1+\sqrt{\frac{d}{n}}+\sqrt{\frac{2\log(2/x)}{n}}\right\}^2 \ , \tag{A.5}$$

with probability larger than $1-x$. Since we assume that $2d/n < 1$, we can choose ϵ so that the first term of (A.5) is twice the second term. After some computations, this leads to the bound

$$\inf_{\theta, \ \|\theta\|_d^2=1} \|Z\theta\|_n^2/n \geq Cx^{\frac{2}{n-2d}}\left[1 \wedge \frac{n}{\log(2/x)}\right] \ ,$$

with probability larger than $1-x$. This allows to conclude.

□

Acknowledgements

I am grateful to Yannick Baraud and Christophe Giraud for interesting suggestions that lead to an improvement of the paper.

References

[1] 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.

[2] 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

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

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

[5] 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)

[6] 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

[7] 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)

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

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

[10] 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)

[11] 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

[12] 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)

[13] 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

[14] 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.*

[15] 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)

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

[17] 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)

[18] 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)

[19] 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)

[20] 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)

[21] 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)

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

[23] 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

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

542–563.
[25] Ingster, Y. (1993a). Asymptotically minimax hypothesis testing for nonparametric alternatives I. *Math. Methods Statist.* **2**, 85–114.
[26] Ingster, Y. (1993b). Asymptotically minimax hypothesis testing for nonparametric alternatives II. *Math. Methods Statist.* **3**, 171–189.
[27] Ingster, Y. (1993c). Asymptotically minimax hypothesis testing for nonparametric alternatives III. *Math. Methods Statist.* **4**, 249–268.
[28] 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)
[29] Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28**, 5, 1302–1338. MR1805785 (2002c:62052)
[30] Litvak, A. E., Pajor, A., Rudelson, M., and Tomczak-Jaegermann, N. (2005). Smallest singular value of random matrices and geometry of random polytopes. *Adv. Math.* **195**, 2, 491–523. http://dx.doi.org/10.1016/j.aim.2004.08.004. MR2146352 (2006g:52009)
[31] 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)
[32] 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)
[33] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34**, 3, 1436–1462. MR2278363 (2008b:62044)
[34] Raskutti, G., Wainwright, M., and Yu, B. (2009). Minimax rates of estimations for high-dimensional regression over l_q balls. Tech. rep., UC Berkeley.
[35] Rigollet, P. and Tsybakov, A. (2010). Exponential screening and optimal rates of sparse estimation. http://arxiv.org/pdf/1003.2654.
[36] 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
[37] Verzelen, N. (2010). High-dimensional gaussian model selection on a gaussian design. *Ann. Inst. H. Poincaré Probab. Statist.* **46**, 2, 480–524.
[38] Verzelen, N. and Villers, F. (2010). Goodness-of-fit tests for high-dimensional Gaussian linear models. *Ann. Statist.* **38**, 2, 704–752. MR2604699
[39] 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
[40] 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.
[41] Yu, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam.* Springer, New York, 423–435. MR1462963 (99c:62137)
[42] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res. 7*, 2541–2563. MR2274449
[43] 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