Bernoulli **24**(4B), 2018, 3683–3710
https://doi.org/10.3150/17-BEJ975

Adaptive estimation of high-dimensional signal-to-noise ratios

NICOLAS VERZELEN[1] and ELISABETH GASSIAT[2]

[1]*INRA, UMR 729 MISTEA, F-34060 Montpellier, France. E-mail: nicolas.verzelen@inra.fr*
[2]*Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, CNRS, Université Paris-Saclay, 91405 Orsay, France. E-mail: elisabeth.gassiat@math.u-psud.fr*

We consider the equivalent problems of estimating the residual variance, the proportion of explained variance η and the signal strength in a high-dimensional linear regression model with Gaussian random design. Our aim is to understand the impact of not knowing the sparsity of the vector of regression coefficients and not knowing the distribution of the design on minimax estimation rates of η. Depending on the sparsity k of the vector regression coefficients, optimal estimators of η either rely on estimating the vector of regression coefficients or are based on U-type statistics. In the important situation where k is unknown, we build an adaptive procedure whose convergence rate simultaneously achieves the minimax risk over all k up to a logarithmic loss which we prove to be non avoidable. Finally, the knowledge of the design distribution is shown to play a critical role. When the distribution of the design is unknown, consistent estimation of explained variance is indeed possible in much narrower regimes than for known design distribution.

Keywords: heritability; minimax analysis; quadratic functional; signal to noise ratio

1. Introduction

1.1. Motivations

In this paper, we investigate the estimation of the proportion of explained variation in high-dimensional linear models with random design, that is the ratio of the variance of the signal to the total amount of variance of the observation. Although this question is of great importance in many applications where the aim is to quantify to what extent covariates explain the variation of the response variable, our analysis is mainly motivated by problems of heritability estimation. In such studies, the response variable is a phenotype measured on n individuals and the predictors are genetic markers on each of these individuals. Then, heritability corresponds to the proportion of phenotypic variance which can be explained by genetic factors. Usually, the number of predictors p greatly exceeds the number n of individuals. When the phenotype under investigation can be explained by a small number of genetic factors, the corresponding regression coefficient vector is sparse, and methods exploiting sparsity are of utmost interest. It appeared recently in biological studies that, for some complex human traits, there was a huge gap (which has been called the "dark matter" of the genome) between the genetic variance explained by populations studies and the one obtained by genome wide associations studies (GWAS), see [25,28] or [17]. To explain this gap, it has been hypothesized that some traits might be "highly polygenic", meaning that genetic factors explaining the phenotype could be so numerous that the corresponding

1350-7265 © 2018 ISI/BS

regression coefficient vector may not be considered as sparse. This may be the case for instance, when psychiatric disorders are associated to neuroanatomical changes as in [2] or [27], see also [30]. As a consequence, sparsity-based methods would be questionable in this situation. When the researcher faces the data, she does not know in general the proportion of relevant predictors, that is the level of sparsity of the parameter. In this work, our first aim is to understand the impact of the ignorance of the sparsity level on heritability estimation. Another important feature of the model when estimating proportion of explained variation is the covariance matrix of the predictors. There is a long standing gap between estimation procedures that assume the knowledge of this covariance (e.g., [7,21]) (which mathematically is the same as assuming that the covariance is the identity matrix) and practical situations where it is generally unknown. Our second aim is to evaluate the impact of the ignorance of the covariance matrix on heritability estimation.

To be more specific, consider the random design high-dimensional linear model

$$y_i = \mathbf{x}_i \beta^* + \varepsilon_i, \qquad i = 1, \dots, n, \tag{1}$$

where $y_i, \varepsilon_i \in \mathbb{R}$, $i = 1, \dots, n$, $\beta^* \in \mathbb{R}^p$, and

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n \end{pmatrix} \in \mathbb{R}^{n \times p}.$$

We assume that the noise $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$ and the rows $\mathbf{x}_i$, $i = 1, \dots, n$, of $\mathbf{X}$ are independent random variables. We also assume that the ε_i, $i = 1, \dots, n$, are independent and identically distributed (i.i.d.) with distribution $\mathcal{N}(0, \sigma^2)$, and that the rows $\mathbf{x}_i$, $i = 1, \dots, n$, of $\mathbf{X}$ are also i.i.d. with distribution $\mathcal{N}(0, \mathbf{\Sigma})$. Throughout the paper, the covariance matrix $\mathbf{\Sigma}$ is assumed to be invertible and the noise level σ is unknown (the case of known noise level is evoked in the discussion section). Our general objective is the optimal estimation of the signal-to-noise ratio

$$\theta := \frac{\mathbb{E}[\|\mathbf{x}_1^T \beta^*\|_2^2]}{\sigma^2} = \frac{\|\mathbf{\Sigma}^{1/2} \beta^*\|_2^2}{\sigma^2}, \tag{2}$$

or equivalently the proportion of explained variation

$$\eta = \eta(\beta^*, \sigma) := \frac{\mathbb{E}[\|\mathbf{x}_1^T \beta^*\|_2^2]}{\operatorname{Var}(y_1)} = \frac{\theta}{1 + \theta} \tag{3}$$

when the vector β^* is unknown and possibly sparse. In the sequel, β^* is said to be k-sparse, when at most k coordinates of β^* are non-zero.

Note that estimating η amounts to deciphering the signal strength from the noise level in $\operatorname{Var}(y_1) = \sigma^2 + \|\mathbf{\Sigma}^{1/2} \beta^*\|_2^2$. Since $\|Y\|_2^2 / \operatorname{Var}(y_1)$ follows a χ^2 distribution with n degrees of freedom, it follows that $\|Y\|_2^2 / n = \operatorname{Var}(y_1)[1 + O_P(n^{-1/2})]$ and it is therefore almost equivalent (up to a parametric loss) to estimate the proportion of explained variation η, the quadratic function $\beta^{*T} \mathbf{\Sigma} \beta^*$ or the noise level σ^2. For the sake of presentation, we mostly express our results in terms of the estimation of η, but they can be easily extended to the signal strength or to the noise estimation problems.

1.2. Main results

There are two main lines of research for estimating σ or η in a high-dimensional setting. Under the assumption that β^* is k-sparse with some small k, it has been established that β^* can be estimated at a fast rate (roughly $k \log p/n$) using for instance Lasso-type procedures, so that using an adequate plug-in method one could hope to estimate η well. Note that in this paper the rate is understood as that of the quadratic risk. Following this general approach, some authors have obtained $(k\log(p)/n)^2$-consistent [29] and $1/n$-consistent [5,16] estimators of σ in some specific regimes. When β^* is dense (that is when many coordinates of β^* are nonzero), such approaches fail. In this regime, a U-type estimator [13] has been proved to achieve consistency at the rate p/n^2. However, its optimality has never been assessed.

Our first main contribution is the proof that the adaptation to unknown sparsity is indeed possible when $\mathbf{\Sigma}$ is known, but at the price of a $\log(p)$ loss factor in the convergence rate when β^* is dense. The idea is the following. Let $\widehat{\eta}^D(\mathbf{\Sigma}^{-1})$ be a U-type estimator which is p/n^2-consistent, the true parameter β^* being sparse or not. We shall denote it the dense estimator. Let also $\widehat{\eta}^{\mathrm{SL}}$ be a $(k\log(p)/n)^2$-consistent estimator when β^* is k-sparse for some small k. Then, if the real β^* is sparse, both estimators should be fairly accurate and should give similar answers, and if the real β^* is dense, or not sparse enough, then $\widehat{\eta}^{\mathrm{SL}}$ will be quite wrong and will give an answer slightly different from the dense estimator. Therefore, the idea is to choose the sparse estimator $\widehat{\eta}^{\mathrm{SL}}$ when both estimators are close enough, so that the quick convergence rate is obtained when the unknown sparsity k is small, and to choose the dense estimator when both estimators are not close, in which case the slower rate is attained which is appropriate in the dense regime. Such a procedure should adapt well to unknown sparsity. Now, to be able to give a precise definition of the estimator, that is to set what "close enough" quantitatively means, one needs a precise understanding of the behavior of the dense and of the sparse estimators. Thus as a first and preliminary step, we obtain a deviation inequality for the dense estimator, see Theorem 2.1. We also establish the minimax estimation risk of η as a function of (k, n, p) when the parameter β^* is k-sparse (see Table 1) and when $\mathbf{\Sigma}$ is known, thereby assessing that Dicker's procedure [13] is optimal in the dense regime ($k \geq \sqrt{p}$) and an estimator based on the square-root Lasso [29] is near optimal in the sparse regime ($k \leq \sqrt{p}$). Again for known $\mathbf{\Sigma}$, we finally construct a data-driven combination of $\widehat{\eta}^D(\mathbf{\Sigma}^{-1})$ (the dense estimator) and $\widehat{\eta}^{\mathrm{SL}}$ (the sparse estimator) following the idea explained before. We prove that such a procedure is indeed

Table 1. Optimal estimation risk $\mathbb{E}[(\widehat{\eta}-\eta)^2]$ when β^* is k-sparse and $\mathbf{\Sigma}$ is known. Here, $a \in (0, 1/2)$ is any arbitrarily small constant and it is assumed below that $n \leq p \leq n^2$. The results remain valid for $p \geq n^2$ if we replace the quantities $\frac{k^2\log^2(p)}{n^2}$ and $\frac{p}{n^2}$ by $\frac{k^2\log^2(p)}{n^2} \wedge 1$ and $\frac{p}{n^2} \wedge 1$, respectively

Sparsity regimes	Minimax risk	Near-optimal procedure
$k \leq \frac{\sqrt{n}}{\log(p)}$	$\frac{1}{n}$	square-root Lasso estimator $\widehat{\eta}^{\mathrm{SL}}$ (12)
$\frac{\sqrt{n}}{\log(p)} \leq k \leq p^{1/2-a}$	$\frac{k^2\log^2(p)}{n^2}$	square-root Lasso estimator $\widehat{\eta}^{\mathrm{SL}}$ (12)
$k \geq \sqrt{p}$	$\frac{p}{n^2}$	Dense estimator $\widehat{\eta}^D(\mathbf{\Sigma}^{-1})$ (8) (see also [13])

adaptive to unknown sparsity, see Theorem 3.2, and that it achieves the minimax adaptive rate with a $\log(p)$ loss factor compared to the non adaptive minimax rate. This logarithmic term is proved to be unavoidable, see Proposition 3.1.

Our second main contribution is an analysis of the proportion of explained variance estimation problem under unknown $\mathbf{\Sigma}$. The construction of dense estimators such as $\widehat{\eta}^D(\mathbf{\Sigma}^{-1})$ requires the knowledge of the covariance matrix $\mathbf{\Sigma}$. But in many practical situations, the covariance structure of the covariates is unknown. For unknown $\mathbf{\Sigma}$, there are basically two main situations:

- Under sufficiently strong structural assumptions on $\mathbf{\Sigma}$ so that $\mathbf{\Sigma}^{-1}$ can be estimated at the rate p/n^2 in operator norm, a simple plug-in method allows to build a minimax and an adaptive minimax procedure with the same rates as when $\mathbf{\Sigma}$ is known, see Corollary 4.4.
- Our main result is that, for a general covariance matrix $\mathbf{\Sigma}$, it is basically impossible to build a consistent estimator of η when k is much larger than n; see Theorem 4.5 and its comments for a precise statement. This is in sharp contrast with the situation where $\mathbf{\Sigma}$ is known, for which the problem of estimating η can be handled in regimes where β^* is impossible to estimate (e.g. $k = p$ and $p = n^{1+\kappa}$ with $\kappa \in (0, 1)$ as depicted in Table 1). For unknown and arbitrary $\mathbf{\Sigma}$, the range of (k, n, p) for which η can be consistently estimated seems to be roughly the same as for estimating β^*, suggesting that signal estimation (β^*) is nearly as difficult as signal strength estimation ($\beta^{*T}\mathbf{\Sigma}\beta^*$). This impossibility result unveils that, in the high-dimensional dense case, the knowledge of the covariance matrix is fundamental and one cannot extend known procedures such as [13,14] or $\widehat{\eta}^D(\mathbf{\Sigma}^{-1})$ to this unknown variance setting.

1.3. Related work

The literature on minimax estimation of quadratic functionals initiated in [15] is rather extensive (see, e.g., [10,24]). In the Gaussian sequence model, that is $n = p$ and $\mathbf{X} = \mathbf{I}_p$, Collier et al. [12] have derived the minimax estimation rate of the functional $\|\beta^*\|_2^2$ for k-sparse vector β^* when the noise level σ is known. However, we are not aware of any minimax result in the high-dimensional linear model even under known noise level.

Another problem related to the estimation of the quadratic functional $\beta^{*T}\mathbf{\Sigma}\beta^*$ is signal detection, which aims at testing the null hypothesis H_0: "$\beta^* = 0$" versus $H_{1,k}[r]$: "$\|\mathbf{\Sigma}^{1/2}\beta^*\|_2^2 \geq r$ and $|\beta^*|_0 \leq k$" (where $|\beta^*|_0$ denotes the number of non null coordinates of β^*). The minimax separation distance is then the smallest r such that a test of H_0 vs $H_{1,r}$ is able to achieve small type I and type II error probabilities. This minimax separation distance is somewhat analogous to a local minimax estimation risk of $\|\mathbf{\Sigma}^{1/2}\beta^*\|_2^2$ around $\beta^* = 0$. In the Gaussian sequence model, minimax separation distances haven been studied in [4,19]. These results have been extended to the high-dimensional linear model under both known [3,20] and unknown [20,34] noise level. Our first minimax lower bound (Proposition 2.4) is largely inspired from these earlier contributions, but the minimax lower bounds for adaptation problems require more elaborate arguments. In particular, the proof of Theorem 4.5 is largely based on new ideas.

Recent works have been devoted to the adaptive estimation of sparse parameters β^* in (1) under unknown variance. As a byproduct, one can then obtain estimators of the variance [5,29]. See also [16] for more direct approaches to variance estimation. In Section 2, we rely on the square-root Lasso estimator to construct the estimator $\widehat{\eta}^{\mathrm{SL}}$ which turns out to be minimax in the sparse regime.

In the dense regime, we already mentioned the contribution of Dicker [13] that proposes method of moments and maximum likelihood based procedures to estimate η when $\mathbf{\Sigma}$ is known. It is shown that the square risk of these estimators goes to 0 at rate p/n^2. When p/n converges to a finite non-negative constant, these estimators are asymptotically normally distributed. Dicker also considers the case of unknown $\mathbf{\Sigma}$ when $\mathbf{\Sigma}$ is highly structured (allowing $\mathbf{\Sigma}$ to be estimable in operator norm at the parametric rate n^{-1}). Janson et al. [21] introduce the procedure EigenPrism for computing confidence intervals of η and study its asymptotic behavior when $\mathbf{\Sigma}$ is known and p/n converges to a constant $c \in (0, \infty)$. Under similar assumptions, Dicker et al. [14] have considered a maximum likelihood based estimator. Bonnet et al. [7] consider a mixed effect model, which is equivalent to assuming that the parameter β^* follows a prior distribution. In the asymptotic regime where $p/n \to c$, they also propose a n^{-1}-rate consistent estimator of η. To summarize, none of the aforementioned contributions has studied minimax convergence rates, the problem of adaptation to sparsity or the estimation problem for unknown $\mathbf{\Sigma}$ (to the exception of [13]).

Finally, there has been a recent interest in the adaptive estimation of other functionals in the linear model (1), such as the coordinates β_i^* of β^* or the sum of coordinates $\sum_{i=1}^{n} \beta_i^*$ [9,22,23,31,35]. However, both the statistical methods and the regimes are qualitatively different for these functionals. After our work was made publicly available, Guo et al. [18] have introduced a procedure based on square-root Lasso to estimate the functional $\|\beta^*\|_2^2$ which, for $\mathbf{\Sigma} = \mathbf{I}$, is equivalent to estimating $\beta^{*T}\mathbf{\Sigma}\beta$. When the parameter β^* is k-sparse with $k \ll \sqrt{p}$ and for $\mathbf{\Sigma} = \mathbf{I}$, the convergence rate of their estimator corresponds to the one of our sparse estimator $\widehat{\eta}^{\mathrm{SL}}$. However, Guo et al. did not study denser settings. Also, for general $\mathbf{\Sigma}$, it is not clear whether optimal rates are the same for estimating $\|\beta^*\|_2^2$ and $\beta^{*T}\mathbf{\Sigma}\beta^*$ in the dense case ($k \gg \sqrt{p}$).

1.4. Notations and organization

The set of integers $\{1, \dots, p\}$ is denoted $[p]$. For any subset J of $[p]$, $\mathbf{X}_J$ is the $n \times |J|$ corresponding submatrix of $\mathbf{X}$. Given a symmetric matrix $\mathbf{A}$, $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$ respectively stand for the largest and the smallest eigenvalue of $\mathbf{A}$, $|\mathbf{A}|$ denotes the determinant of $\mathbf{A}$. For a vector u, $\|u\|_p$ denotes its l_p norm and $|u|_0$ stands for its l_0 norm (ie its number of non-zero components). For any matrix $\mathbf{A}$, $\|\mathbf{A}\|_p$ denotes the l_p norm of the vectorized version of $\mathbf{A}$, that is $(\sum |\mathbf{A}_{i,j}|^p)^{1/p}$. The Frobenius norm is also denoted $\|\mathbf{A}\|_F$. Finally, the l_2 operator norm of a matrix $\mathbf{A}$ writes $\|\mathbf{A}\|_{\mathrm{op}}$. In what follows, $C, C', \dots$ denote universal constants whose value may vary from line to line whereas C_1, C_2 and C_3 denote numerical constants that will be used in several places of our work.

In Section 2, we introduce the two main procedures and characterize the minimax estimation risk of η when both the covariance matrix $\mathbf{\Sigma}$ and the sparsity are known. Section 3 is devoted to the problem of adaptation to the unknown sparsity, whereas the case of unknown covariance $\mathbf{\Sigma}$ is studied in Section 4. Extensions to fixed design regression and other related problems are discussed in Section 5. All the proofs are postponed to the end of the paper and to the supplement (see [33]).

2. Minimax rates for known sparsity

In this section, we consider two estimators. In the spirit of [13], the first estimator $\widehat{\eta}^D(\mathbf{\Sigma}^{-1})$ is designed for the dense regime ($|\beta^*|_0 \geq p^{1/2}$) and it is proved to be consistent with rate p/n^2 irrespectively of the parameter sparsity. When β^* is in fact highly sparse, the estimator $\widehat{\eta}^{\mathrm{SL}}$ based on the square-root Lasso better exploits the structure of β^* and achieves the estimation rate $(\frac{|\beta^*|_0 \log(p)}{n})^2 + n^{-1}$. It turns out that these two procedures (almost) achieve the minimax estimation rate when it is known whether β^* is sparse or not.

2.1. Dense regime

In this subsection, we introduce an estimator of η which will turn out to be mostly interesting for dense parameters β^*. Its definition is close to that in [13]. We provide a detailed analysis of this estimator, and our concentration inequality in Theorem 2.1 below will turn out to be useful both for the adaptation problem and for the case of unknown $\mathbf{\Sigma}$.

Since $\mathrm{Var}(y_1)$ is easily estimated by $\|Y\|_2^2/n$, the main challenge is to estimate $\|\mathbf{\Sigma}^{1/2}\beta^*\|^2$. Thus, the question is how to separate in Y the randomness coming from $\mathbf{X}\beta^*$ from that coming from the ε_i's, $i = 1, \dots, n$. The idea is to use the fact that the noise ε is isotropic whereas, conditionally on $\mathbf{X}$, $\mathbf{X}\beta^*$ is not isotropic. Respectively, denote (λ_i, u_i), $i = 1, \dots, n$ the eigenvalues and eigenvectors of $(\mathbf{X}\mathbf{X}^T)/p$. We will prove, that in a high-dimensional setting where $p > n$, $\mathbf{X}\beta^*$ is slightly more aligned with left eigenvectors of $\mathbf{X}$ associated to large eigenvalues than with those associated to small eigenvalues. This subtle phenomenon suggests that the distribution of the random variable T

$$T := \frac{p}{n^2} \sum_{i=1}^n (\lambda_i - \bar{\lambda})\big(Y^T u_i\big)^2, \qquad \text{where } \bar{\lambda} := \sum_{i=1}^n \lambda_i / n,$$

(almost) does not depend on the noise level σ and, at the same time, captures some functional of the signal β^*. This functional turns out to be $\beta^{*T}\mathbf{\Sigma}^2\beta^*$. One can rewrite the random variable as a quadratic form of Y

$$T = \frac{Y^T(\mathbf{X}\mathbf{X}^T - \mathrm{tr}(\mathbf{X}\mathbf{X}^T)\mathbf{I}_n/n)Y}{n^2}. \tag{4}$$

Working with a normalized estimator $\widehat{V} := \frac{Tn^2}{\|Y\|_2^2(n+1)}$, we state in the following theorem that $\widehat{V}$ concentrates exponentially fast around $\beta^{*T}\mathbf{\Sigma}^2\beta^*/\mathrm{Var}(y_1)$. Note that, due to the fact that $\mathbf{\Sigma}$ is squared in the numerator, $\widehat{V}$ does not concentrate around η (except in case $\mathbf{\Sigma} = \mathbf{I}_p$). However, an appropriate modification explained below will allow to use the following theorem to estimate η for known general $\mathbf{\Sigma}$.

Theorem 2.1. *Assume that $p \geq n$.*

There exist numerical constants C_1 and C_2 such that for all $t \leq n^{1/3}$,

$$\mathbb{P}\left[\left|\widehat{V} - \frac{\beta^{*T}\mathbf{\Sigma}^2\beta^*}{\mathrm{Var}(y_1)}\right| \leq C_1 \|\mathbf{\Sigma}\|_{\mathrm{op}} \frac{\sqrt{pt}}{n}\right] \geq 1 - C_2 e^{-t}. \tag{5}$$

There exists a numerical constant C such that

$$\mathbb{E}\left[\left(\widehat{V}-\frac{\beta^{*T}\mathbf{\Sigma}^2\beta^*}{\mathrm{Var}(y_1)}\right)^2\right]\leq C\|\mathbf{\Sigma}\|_{\mathrm{op}}^2\frac{p}{n^2}. \tag{6}$$

Remark 2.1. The proof relies on recent exponential concentration inequalities for Gaussian chaos [1] and a new concentration inequality of the spectrum of $\mathbf{X}\mathbf{X}^T/n$ around $\mathrm{tr}(\mathbf{\Sigma})/n$ (Lemma C.2). The concentration inequality (5) will be the key tool in the construction of adaptive estimators in the next section.

Remark 2.2. When $\mathbf{\Sigma}=\mathbf{I}_p$, the above theorem enforces that $\widehat{V}$ estimates the proportion of explained variation η at the rate p/n^2, uniformly over all β^* and $\sigma>0$. Note that $\widehat{V}$ is only consistent in the regime where n^2 is large compared to p.

For arbitrary $\mathbf{\Sigma}$ (with bounded eigenvalues), the above theorem only implies that $\widehat{V}$ is *of the same order as* η, that is, there exists positive constant c and C such that $c\lambda_{\min}(\mathbf{\Sigma})\leq\widehat{V}/\eta\leq C\lambda_{\max}(\mathbf{\Sigma})$.

Nevertheless, when the covariance $\mathbf{\Sigma}$ is known, it is possible to get a consistent estimator of η. Replace the design matrix $\mathbf{X}$ in the linear regression model by $\tilde{\mathbf{X}}:=\mathbf{X}\mathbf{\Sigma}^{-1/2}$ in such a way that its rows $\tilde{\mathbf{x}}_i$ follow i.i.d. standard normal distributions and

$$Y=\tilde{\mathbf{X}}\mathbf{\Sigma}^{1/2}\beta^*+\varepsilon. \tag{7}$$

Then, we define the estimator $\widehat{\eta}^D$ as $\widehat{V}$ where $\mathbf{X}$ is replaced by $\tilde{\mathbf{X}}$, so that $\widehat{\eta}^D$ is a quadratic form of $\mathbf{Y}$ with a matrix involving the precision matrix, that is the inverse covariance matrix $\mathbf{\Sigma}^{-1}$. Let us denote $\mathbf{\Omega}:=\mathbf{\Sigma}^{-1}$, and define

$$\widehat{\eta}^D(\mathbf{\Omega}):=\frac{Y^T(\mathbf{X}\mathbf{\Omega}\mathbf{X}^T-\mathrm{tr}(\mathbf{X}\mathbf{\Omega}\mathbf{X}^T)\mathbf{I}_n/n)Y}{(n+1)\|Y\|^2} \tag{8}$$

(we could replace $\mathrm{tr}(\mathbf{X}\mathbf{\Omega}\mathbf{X}^T)$ by p in the above definition without changing the rate in the corollary below). We straightforwardly derive from Theorem 2.1 that $\widehat{\eta}^D(\mathbf{\Omega})$ estimates η at the rate p/n^2.

Corollary 2.2. *Assume that $p\geq n$. There exists a numerical constant C such that the estimator $\widehat{\eta}^D(\mathbf{\Omega})$ satisfies*

$$\mathbb{E}\left[\left(\widehat{\eta}^D(\mathbf{\Omega})-\eta\right)^2\right]\leq C\frac{p}{n^2}. \tag{9}$$

Remark 2.3. It turns out that $\widehat{\eta}^D(\mathbf{\Omega})$ is consistent for p small compared to n^2 even though consistent estimation of β^* is impossible in this regime. Although developed independently, the estimator $\widehat{\eta}^D(\mathbf{\Omega})$ shares some similarities with the method of moment based estimator of Dicker [13], which also achieves the p/n^2 convergence rate.

Remark 2.4. In the low-dimensional case $p\leq n$, one may easily adapt the proof of Theorem 2.1 to get that $\mathbb{E}[(\widehat{\eta}^D(\mathbf{\Omega})-\eta)^2]\leq C/n$ for some constant $C>0$.

2.2. Sparse regime: Square-root lasso estimator

When β^* is highly sparse, the signal to noise ratio estimator is based on a Lasso-type estimator of β^* proposed in [6,29]. As customary for Lasso-type methods, we shall work with a standardized version $\mathbf{W}$ of the matrix $\mathbf{X}$, whose columns $\mathbf{W}_{\bullet j}$ satisfy $\|\mathbf{W}_{\bullet j}\|_2 = 1$. Since the noise-level σ is unknown, we cannot readily use the classical Lasso estimator whose optimal value of the tuning parameter depends on σ. Instead, we rely on the square-root Lasso [6] defined by

$$\widetilde{\beta}_{\mathrm{SL}} := \arg\min_{\beta \in \mathbb{R}^p} \sqrt{\|Y - \mathbf{W}\beta\|_2^2} + \frac{\lambda_0}{\sqrt{n}}\|\beta\|_1, \qquad (\widehat{\beta}_{\mathrm{SL}})_j := (\widetilde{\beta}_{\mathrm{SL}})_j / \|\mathbf{x}_j\|_2. \tag{10}$$

In the sequel, the tuning parameter λ_0 is set to $\lambda_0 := 13\sqrt{\log(p)}$ (there is nothing specific with this particular choice). In the proof, we will also use an equivalent definition of the square-root estimator introduced in [29]

$$(\widetilde{\beta}_{\mathrm{SL}}, \widetilde{\sigma}_{\mathrm{SL}}) = \arg\min_{\beta \in \mathbb{R}^p, \sigma' > 0} \left[\frac{n\sigma'}{2} + \frac{\|Y - \mathbf{W}\beta\|_2^2}{2\sigma'}\right] + \lambda_0\|\beta\|_1. \tag{11}$$

(To prove the equivalence between the two definitions, minimize (11) with respect to σ'.) Notice that $\widetilde{\sigma}_{\mathrm{SL}} = \|Y - \mathbf{W}\widetilde{\beta}_{\mathrm{SL}}\|_2/\sqrt{n} = \|Y - \mathbf{X}\widehat{\beta}_{\mathrm{SL}}\|_2/\sqrt{n}$. Then, we define the estimator

$$\widehat{\eta}^{\mathrm{SL}} := 1 - \frac{n\widetilde{\sigma}_{\mathrm{SL}}^2}{\|Y\|_2^2} = 1 - \frac{\|Y - \mathbf{X}\widehat{\beta}_{\mathrm{SL}}\|_2^2}{\|Y\|_2^2}. \tag{12}$$

The following proposition is a consequence of Theorem 2 in [29].

Proposition 2.3. *There exist two numerical constants C and C' such that the following holds. Assume that β^* is k-sparse, that $p \geq n$ and*

$$k\log(p)\frac{\lambda_{\max}(\mathbf{\Sigma})}{\lambda_{\min}(\mathbf{\Sigma})} \leq Cn. \tag{13}$$

Then the square-root Lasso based estimator $\widehat{\eta}^{\mathrm{SL}}$ satisfies

$$\mathbb{E}\big[(\widehat{\eta}^{\mathrm{SL}} - \eta)^2\big] \leq C'\left[\frac{1}{n} + \frac{k^2\log^2(p)}{n^2}\frac{\lambda_{\max}^2(\mathbf{\Sigma})}{\lambda_{\min}^2(\mathbf{\Sigma})}\right]. \tag{14}$$

Remark 2.5. Condition (13) is unavoidable, as the minimax risk of proportion of explained variation estimation is bounded away from zero when $k\log(p)$ is large compared to n (see Proposition 2.4 later). To ease the presentation, we have expressed Condition (13) in terms of largest and smallest eigenvalues of $\mathbf{\Sigma}$. One could also replace these quantities by local ones such as compatibility constants (see the proof for more details).

2.3. Minimax lower bound

We shall prove in the sequel that a combination of the estimators $\widehat{\eta}^D(\mathbf{\Omega})$ and $\widehat{\eta}^{\mathrm{SL}}$ essentially achieves the minimax estimation risk. In the following minimax lower bound we assume that the covariance $\mathbf{\Sigma}$ is the identity matrix $\mathbf{I}_p$.

Define $\mathbb{B}_0[k]$ the collection of k-sparse vectors of size p. Given any estimator $\widehat{\eta}$, define the maximal risk $R(\widehat{\eta}, k)$ over k-sparse parameters by

$$R(\widehat{\eta}, k) := \sup_{\beta \in \mathbb{B}_0[k], \sigma > 0} \mathbb{E}_{\beta,\sigma}\big[\{\hat{\eta} - \eta(\beta, \sigma)\}^2\big], \tag{15}$$

where $\mathbb{E}_{\beta,\sigma}[\cdot]$ is the expectation with respect to $(Y, \mathbf{X})$ where $Y = \mathbf{X}\beta + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$ and the covariance matrix of the rows of $\mathbf{X}$ is $\mathbf{I}_p$. Then, the minimax risk is denoted $R^*(k) := \inf_{\widehat{\eta}} R(\widehat{\eta}, k)$, where the infimum is taken over all estimators $\widehat{\eta}$ measurable with respect to $(Y, \mathbf{X})$. Notice that the minimax risk is therefore a function of (k, n, p). For ease of notations, we use $R^*(k)$ and not $R^*(k, n, p)$ since we are interested in the adaptivity to unknown k.

Proposition 2.4 (Minimax lower bound). *There exists a numerical constant* $C > 0$ *such that for any* $1 \leq k \leq p$,

$$R^*(k) \geq C\left(\left\{\left[\frac{k}{n}\log\left(1 + \frac{p}{k^2} \vee \sqrt{\frac{p}{k^2}}\right)\right]^2 \wedge 1\right\} + \frac{1}{n}\right). \tag{16}$$

The proof of this proposition follows the lines developed to derive minimax lower bounds for the signal detection problem (see, e.g., Theorem 4.3 in [34]). Nevertheless, as this proposition is a first step towards more complex settings, we provide a self-contained proof in Section 6.1.

In (16), we recognize three regimes:

- If $k \geq p^{1/2}$, the minimax rate is larger than $(p/n^2) \wedge 1$. This optimal risk is achieved by the dense estimator $\widehat{\eta}^D(\mathbf{\Omega})$ up to a constant number.
- If $k \leq p^{1/2-\gamma}$ for some arbitrary small $\gamma > 0$, the minimax rate is of order

  $$\frac{1}{n} + \left(\frac{k\log(p)}{n}\right)^2 \wedge 1.$$

  More precisely for $k \leq [\sqrt{n}/\log(p)]$, it is of order n^{-1}, whereas for larger k it is of order $(k\log(p)/n)^2 \wedge 1$. This bound is achieved by the square-root Lasso estimator $\widehat{\eta}^{\mathrm{SL}}$, which does not require the knowledge of $\mathbf{\Sigma}$ and k.
- For k close to $p^{1/2}$ (e.g., $k = (p/\log(p))^{1/2}$), the minimax lower bound (16) and the upper bound (14) only match up to some $\log(p)$ factors. Such a logarithmic mismatch has also been obtained in the related work [4] on minimax detection rates for testing the null hypothesis $\beta^* = 0$ when the design matrix is fixed and orthonormal, that is $p = n$ and $\mathbf{X} = \mathbf{I}_p$. In this orthonormal setting, Collier et al. [12] have very recently closed this gap. Transposed in our setting, their results would suggest that the optimal risk is of order $(k\log(p/k^2)/n)^2$,

suggesting that Proposition 2.4 is sharp. In the specific case where $\mathbf{\Sigma} = \mathbf{I}_p$, it seems possible to extend the estimator of $\|\beta^*\|_2^2$ introduced by [12] to our setting by considering the pairwise correlations $Y^T \mathbf{W}_{\bullet j}$ for $j = 1, \dots, p$. Such estimator would then presumably be $(k \log(p/k^2)/n)^2$ consistent. As this approach does not seem to extend easily to arbitrary $\mathbf{\Sigma}$, we did not go further in this direction.

Remark 2.6. In the definition (15) of $R[\widehat{\eta}, k]$ and in the definition of the minimax risk $R^*[k]$, we restricted ourselves to the case where the covariance matrix $\mathbf{\Sigma}$ of the covariates is the identity. For known but general $\mathbf{\Sigma}$, we do not have a matching lower bound. Nevertheless, one deduces from the above results that, when either $k \ll p$ and (13) is satisfied or when $k \geq \sqrt{p}$ and $\mathbf{\Sigma}$ is invertible, the optimal risk is (up to numerical constants) not larger than than $R^*[k]$.

3. Adaptation to unknown sparsity

In practice, the number $|\beta^*|_0$ of non-zero components of β^* is unknown. In this section, our purpose is to build an estimator $\widehat{\eta}$ that adapts to the unknown sparsity $|\beta^*|_0$. Although the computation of the estimators $\widehat{\eta}^D(\mathbf{\Omega})$ and $\widehat{\eta}^{\mathrm{SL}}$ does not require the knowledge of $|\beta^*|_0$, the choice of one estimator over the other depends on this quantity. Observe that, when $p \geq n^2$, the dense estimator $\widehat{\eta}^D(\mathbf{\Omega})$ is not consistent. Therefore, only the estimator $\widehat{\eta}^{\mathrm{SL}}$ is useful and $\widehat{\eta}^{\mathrm{SL}}$ alone is minimax adaptive to the sparsity k (up to a possible log factor when k is of the order of $p^{1/2}$). This is why we focus on the regime where p is large compared to n and where $p \log p \leq n^2$.

It turns out that no estimator $\widehat{\eta}$ can simultaneously achieve the minimax risk $R^*(k)$ over all $k = 1, \dots, p$, and that there is an unavoidable loss for adaptation. This may be seen in the following proposition.

Proposition 3.1. *Assume that $p \log p \leq n^2$, and that for some $a \in]0, 1/2[$, $p^{1-a}(\log p)^2 \geq 16n$. Then for any estimator $\widehat{\eta}$, for all k such that $\sqrt{p \log p} \leq k \leq p$, one has*

$$\frac{R(\widehat{\eta}, 1)}{\frac{1}{n}\sqrt{\frac{p}{n}}} + \frac{R(\widehat{\eta}, k)}{\frac{p \log p}{n^2}} \geq \frac{a^2}{4^5}.$$

Recall that $R^*(1)$ is of order $1/n$ and $R^*(k)$ is of order p/n^2. Proposition 3.1 implies that any estimator $\widehat{\eta}$ whose maximal risk over $\mathbf{B}_0[k]$ is smaller than $p \log(p)/n^2$ exhibits a huge maximal risk over $\mathbf{B}_0[1]$. As a consequence, any estimator admitting a reasonable risk bound over $\mathbf{B}_0[1]$ should have a maximal risk at least of order $p \log(p)/n^2$ for all $k \in [\sqrt{p \log(p)}, p]$. Next, we define an estimator $\widehat{\eta}^A$ simultaneously achieving the risk $R^*(k)$ for k small compared to $\sqrt{p}$ and achieving the risk $R^*(k) \log p$ in the dense regime where $k \geq \sqrt{p \log p}$.

Define the numerical constant c_0 as two times the constant C_1 arising in the deviation bound (5) of Theorem 2.1. We build an adaptive estimator by combining the estimator $\widehat{\eta}^{\mathrm{SL}}$ and $\widehat{\eta}^D$ as follows

$$\widehat{\eta}^A := \begin{cases} \widehat{\eta}^{\mathrm{SL}} & \text{if } \left|\widehat{\eta}_T^D(\mathbf{\Omega}) - \widehat{\eta}^{\mathrm{SL}}\right| \leq c_0\sqrt{p \log(p)}/n, \\ \widehat{\eta}_T^D(\mathbf{\Omega}) & \text{else,} \end{cases} \tag{17}$$

where, for technical reasons, we consider $\widehat{\eta}_T^D(\mathbf{\Omega}) := \min(1, \max(0, \widehat{\eta}^D(\mathbf{\Omega})))$ a truncated version of $\widehat{\eta}^D(\mathbf{\Omega})$ which lies in $[0, 1]$.

The rationale behind $\widehat{\eta}^A$ is the following. Suppose that β^* is k-sparse, with $k \leq \sqrt{p}$, in which case, $\widehat{\eta}^{\mathrm{SL}}$ achieves the optimal rate. With large probability, $|\widehat{\eta}_T^D(\mathbf{\Omega}) - \eta|$ is smaller than $c_0\sqrt{p\log(p)}/(2n)$ (this is true for arbitrary β^*) and $|\widehat{\eta}^{\mathrm{SL}} - \eta|$ is smaller than $(1/\sqrt{n} + k\log(p)/n)$ which is smaller than $c_0\sqrt{p\log(p)}/(2n)$. Hence, $\widehat{\eta}^A$ equals $\widehat{\eta}^{\mathrm{SL}}$ with large probability. Now assume that $k \geq \sqrt{p}$, in which case the optimal rate is of order p/n^2 and is achieved by $\widehat{\eta}_T^D(\mathbf{\Omega})$. Observe that $\widehat{\eta}^A = \widehat{\eta}_T^D(\mathbf{\Omega})$ except if $\widehat{\eta}^{\mathrm{SL}}$ is at distance less than $c_0\sqrt{p\log(p)}/n$ from $\widehat{\eta}_T^D(\mathbf{\Omega})$. Consequently, $|\widehat{\eta}^A - \eta| \leq c_0\sqrt{p\log(p)}/n + |\widehat{\eta}_T^D(\mathbf{\Omega}) - \eta|$. Formalizing the above argument, we arrive at the following.

Theorem 3.2. *There exists a numerical constant C such that the following holds. Assume that $p \geq n$. For any integer $k \in [p]$, any k-sparse vector β^* and any $\sigma > 0$, the estimator $\widehat{\eta}^A$ satisfies*

$$\mathbb{E}\big[(\widehat{\eta}^A - \eta)^2\big] \leq C\left[\frac{1}{n} + \left(\frac{k^2\log^2(p)}{n^2}\frac{\lambda_{\max}^2(\mathbf{\Sigma})}{\lambda_{\min}^2(\mathbf{\Sigma})}\right) \wedge \left(\frac{p\log(p)}{n^2}\right)\right].$$

For all $k \geq \sqrt{p\log(p)}$, the risk $R(\widehat{\eta}^A, k)$ is of order $p\log(p)/n^2$ whereas $R(\widehat{\eta}^A, 1)$ is of order $1/n$. In view of Proposition 3.1, it is therefore impossible to improve the rates $R(\widehat{\eta}^A, k)$ for $k \geq \sqrt{p\log(p)}$ without drastically deteriorating $R(\widehat{\eta}^A, 1)$. As a consequence of Propositions 2.4, 3.1 and Theorem 3.2, and, in the asymptotic regime where $p\log p \leq n^2$ and p^{1-a} is large compared to n for some positive a, $\widehat{\eta}^A$ achieves the optimal adaptive risk for all $k \in \{1, \dots, p^{1/2-\gamma}\} \cup \{(p\log(p))^{1/2}, \dots, p\}$ where $\gamma > 0$ is arbitrary small. For k close to $\sqrt{p}$, there is still a logarithmic gap between the upper and lower bounds as in the non-adaptive section.

Remark 3.1. Theorem 2.1 is the basic stone for the construction of $\widehat{\eta}^A$ by the use of the deviation inequality.

Remark 3.2. In the low-dimensional case $p \leq n$, the estimator $\widehat{\eta}^D$ achieves the parametric rate $\mathbb{E}[(\widehat{\eta}^D(\mathbf{\Omega}) - \eta)^2] \leq C/n$ regardless of the sparsity of the parameter β^*. Thus, $\widehat{\eta}^D$ alone is adaptive to the unknown sparsity.

4. Minimax estimation when $\mathbf{\Sigma}$ is unknown

In this section, we investigate the case where the covariance matrix $\mathbf{\Sigma}$ is unknown. As the computation of the sparse estimator $\widehat{\eta}_{\mathrm{SL}}$ does not require the knowledge of $\mathbf{\Sigma}$, the optimal estimation rate is unchanged when $|\beta^*|_0$ is much smaller than $\sqrt{p}$. In what follows we therefore focus on the regime where $|\beta^*|_0 \geq \sqrt{p}$.

4.1. Positive results under restrictions on $\mathbf{\Sigma}$

Here, we prove that a simple plug-in method allows to achieve the minimax rate as long as one can estimate the inverse covariance matrix $\mathbf{\Omega}$ sufficiently well. Without loss of generality, we may assume that we have at our disposal an independent copy of $\mathbf{X}$, denoted $\mathbf{X}^{(2)}$ (if it is not the case, simply divide the data set into two subsamples of the same size).

Given an estimator $\widehat{\mathbf{\Omega}}$ of $\mathbf{\Omega} := \mathbf{\Sigma}^{-1}$ based on the matrix $\mathbf{X}^{(2)}$, the proportion of explained variation η is estimated as in Section 2.1, using (8), except that the true inverse covariance matrix is replaced by its estimator:

$$\widehat{\eta}^D(\widehat{\mathbf{\Omega}}) := \frac{Y^T(\mathbf{X}\widehat{\mathbf{\Omega}}\mathbf{X}^T - \operatorname{tr}(\mathbf{X}\widehat{\mathbf{\Omega}}\mathbf{X}^T)\mathbf{I}_n/n)Y}{(n+1)\|Y\|^2}. \tag{18}$$

Notice that Dicker [13] has already proposed a similar procedure and has derived asymptotic rates similar to the ones we may deduce from Proposition 4.1. However, deviation inequalities (and not only asymptotic rates) are required to construct an estimator that is adaptive to the unknown sparsity.

Proposition 4.1. *Assume that $p \geq n$. For any non-singular estimator $\widehat{\mathbf{\Omega}}$ based on the sample $\mathbf{X}^{(2)}$,*

$$\mathbb{P}\left[\left|\widehat{\eta}^D(\widehat{\mathbf{\Omega}}) - \eta\right| \geq C_1\|\mathbf{\Sigma}\|_{\mathrm{op}}\|\widehat{\mathbf{\Omega}}\|_{\mathrm{op}}\frac{\sqrt{pt}}{n} + \|\mathbf{\Sigma}\|_{\mathrm{op}}\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}\|_{\mathrm{op}}|\mathbf{X}^{(2)}\right] \leq C_2 e^{-t}, \tag{19}$$

for all $t < n^{1/3}$. Here, C_1 and C_2 are the numerical constants that appear in Theorem 2.1.

Thus, if one is able to estimate $\mathbf{\Omega}$ at the rate p/n^2, then $\widehat{\eta}^D(\widehat{\mathbf{\Omega}})$ achieves the same estimation rate as if $\mathbf{\Sigma}$ was known. To illustrate this qualitative situation, we describe an example of a class $\mathcal{U}$ of precision matrices and an estimator $\widehat{\mathbf{\Omega}}$ satisfying this property.

For any square matrix $\mathbf{A}$, define its matrix l_1 operator norm by

$$\|\mathbf{A}\|_{1\to 1} = \max_{1\leq j\leq p}\sum_{1\leq i\leq p}|\mathbf{A}_{i,j}|.$$

Given any $M > 0$ and $M_1 > 0$, consider the following collection $\mathcal{U}$ of sparse inverse covariance matrices

$$\begin{aligned}\mathcal{U} :=& \mathcal{U}(M, M_1)\\ :=& \left\{\mathbf{\Omega} : \mathbf{\Omega} \succ 0 : \frac{1}{M_1} \leq \lambda_{\min}(\mathbf{\Omega}) \leq \lambda_{\max}(\mathbf{\Omega}) \leq M_1, \|\mathbf{\Omega}\|_{1\to 1} \leq M,\right.\\ &\left.\max_{1\leq j\leq p}\sum_{i=1}^{p}\mathbf{1}_{\mathbf{\Omega}_{i,j}\neq 0} \leq \sqrt{\frac{p}{n\log(p)}}\right\}.\end{aligned} \tag{20}$$

Cai et al. [8] introduced the CLIME estimator to estimate sparse precision matrices. Let $\lambda_n > 0$ and $\rho > 0$ be two tuning parameters, whose value will be fixed in Lemma 4.2 below. Denote $\widehat{\mathbf{\Sigma}}^{(2)} := \mathbf{X}^{(2)T}\mathbf{X}^{(2)}/n$ the empirical covariance matrix based on the observations $\mathbf{X}^{(2)}$.

Let $\widehat{\mathbf{\Omega}}_1$ be the solution of the following optimization problem

$$\min \|\mathbf{\Omega}'\|_1, \qquad \text{subject to } \|\widehat{\mathbf{\Sigma}}^{(2)}\mathbf{\Omega}' - \mathbf{I}_p\|_\infty \le \lambda_n, \mathbf{\Omega}' \in \mathbb{R}^{p\times p}. \tag{21}$$

Then, the CLIME estimator $\widehat{\mathbf{\Omega}}_{\mathrm{CL}}$ is obtained by symmetrizing $\widehat{\mathbf{\Omega}}_1$: for all i, j, we take $(\widehat{\mathbf{\Omega}}_{\mathrm{CL}})_{i,j} = (\widehat{\mathbf{\Omega}}_1)_{i,j}$ if $|(\widehat{\mathbf{\Omega}}_1)_{i,j}| \le |(\widehat{\mathbf{\Omega}}_1)_{j,i}|$ and $(\widehat{\mathbf{\Omega}}_{\mathrm{CL}})_{i,j} = (\widehat{\mathbf{\Omega}}_1)_{j,i}$ in the opposite case. We may now apply Theorem 1.a in [8] to our setting with $\eta = 1/5 \wedge 1/\sqrt{M_1}$, $K = e^{1/2}$ and $\tau = 1$. This way we obtain the following.

Lemma 4.2. *There exists a numerical constant $C_3 > 0$ such that the following holds. Fix $\lambda_n = 2[25 \vee M_1](3 + e^3(5 \vee \sqrt{M_1})^2 M\sqrt{\log(p)/n}$. Assume that $\log(p) \le n/8$ and that $\mathbf{\Omega}$ belongs to $\mathcal{U}$. Then, the CLIME estimator satisfies*

$$\|\widehat{\mathbf{\Omega}}_{\mathrm{CL}} - \mathbf{\Omega}\|_{\mathrm{op}} \le C_3 M^2 M_1^2 \frac{\sqrt{p}}{n}, \tag{22}$$

with probability larger than $1 - 4/p$.

Let us modify the estimator of η so that it effectively lies in $[0, 1]$. Let $\widehat{\eta}^D_T(\widehat{\mathbf{\Omega}}) := \min(1, \max(0, \widehat{\eta}^D(\widehat{\mathbf{\Omega}})))$.

Corollary 4.3. *Assume that $p \ge n$ and that $\mathbf{\Omega}$ belongs to the collection $\mathcal{U}$ defined above. Then, there exists a universal constant $C > 0$ such that the following holds. For any β^* and $\sigma > 0$,*

$$\mathbb{E}\big[\{\widehat{\eta}^D_T(\widehat{\mathbf{\Omega}}_{\mathrm{CL}}) - \eta\}^2\big] \le \left[C M^4 M_1^6 \frac{p}{n^2}\right] \wedge 1.$$

We shall now define an adaptive estimator $\widehat{\eta}^A_{\mathrm{CL}}$ in the same spirit as $\widehat{\eta}^A$ in the previous subsection. Define $c_0(M, M_1)$ by

$$c_0(M, M_1) := 4C_1 M_1^2 + 2C_3 M^2 M_1^3.$$

Here, C_1 is the numerical constant that appears in Theorem 2.1 and C_3 the numerical constant that appears in Lemma 4.2. Define the estimator as:

$$\widehat{\eta}^A_{\mathrm{CL}} := \begin{cases} \widehat{\eta}^{\mathrm{SL}} & \text{if } |\widehat{\eta}^D_T(\widehat{\mathbf{\Omega}}_{\mathrm{CL}}) - \widehat{\eta}^{\mathrm{SL}}| \le c_0(M, M_1)\sqrt{p\log(p)}/n, \\ \widehat{\eta}^D_T(\widehat{\mathbf{\Omega}}_{\mathrm{CL}}) & \text{else.} \end{cases} \tag{23}$$

We then obtain that $\widehat{\eta}^A_{\mathrm{CL}}$ is asymptotically minimax adaptive to $\mathbf{\Omega}$ (if it is known that $\mathbf{\Omega} \in \mathcal{U}$) and to sparsity, in the same regimes as those in which $\widehat{\eta}^A$ is asymptotically minimax adaptive to sparsity.

Corollary 4.4. *Assume that $\mathbf{\Omega}$ belongs to the collection $\mathcal{U}$ defined above. Then, there exists a constant $C(M, M_1) > 0$ only depending on M and M_1 such that the following holds. For any* *integer $k \in [p]$, any k-sparse vector β^* and any $\sigma > 0$,*

$$\mathbb{E}\big[\big(\widehat{\eta}^A_{\mathrm{CL}} - \eta\big)^2\big] \le C(M, M_1)\left[\frac{1}{n} + \left(\frac{k^2 \log^2(p)}{n^2}\right) \wedge \left(\frac{p \log(p)}{n^2}\right)\right]. \tag{24}$$

Remark 4.1. When $\mathbf{\Omega}$ belongs to $\mathcal{U}$, the estimator $\widehat{\eta}^D_T(\widehat{\mathbf{\Omega}}_{\mathrm{CL}})$ achieves a similar risk bound to that of $\widehat{\eta}^D_T(\mathbf{\Omega})$. Also, $\widehat{\eta}^A_{\mathrm{CL}}$ performs as well as estimator $\widehat{\eta}^A$ which requires the knowledge of $\mathbf{\Omega}$. As a consequence, there does not seem to be a price to pay for the adaptation to $\mathbf{\Omega}$ under the restriction $\mathbf{\Omega} \in \mathcal{U}$.

Remark 4.2. If the quantity $\sqrt{p/(n\log(p))}$ in the sparsity condition $\max_{1\le j\le p} \sum_{i=1}^p \mathbf{1}_{\mathbf{\Omega}_{i,j}\neq 0} \le \sqrt{p/(n\log(p))}$ in the definition (20) of $\mathcal{U}$ is replaced by some $s \ge \sqrt{p/(n\log(p))}$, the CLIME-based estimator $\widetilde{\eta}^D_T(\widehat{\mathbf{\Omega}}_{\mathrm{CL}})$ will only be consistent at the rate $s^2\log(p)/n$ which is slower than the desired p/n^2. This is not completely unexpected as we prove in the next subsection that a reliable estimation of η becomes almost impossible when the collection of precision matrices is too large.

4.2. Impossibility results

We now turn to the general problem where $\mathbf{\Sigma}$ is only assumed to have bounded eigenvalues. As explained in the beginning of Section 3, the estimator $\widehat{\eta}^{\mathrm{SL}}$, which does not require the knowledge of $\mathbf{\Sigma}$, is minimax adaptive to $\mathbb{B}_0[k]$ when $p \ge n^2$. Hence, we focus in the remainder of this section on the regime $n \le p \le n^2$.

In this subsection and the corresponding proofs, we denote $\mathbb{P}_{\beta,\sigma,\mathbf{\Sigma}}$ the distribution of $(Y, \mathbf{X})$, in order to emphasize the dependency of the data distributions with respect to the covariance matrix of $\mathbf{X}$. For any $M > 1$, let us introduce $\Xi[M]$ the set of positive symmetric matrices of size p whose eigenvalues lie in the compact $[1/M, M]$. The purpose of these bounded eigenvalues in $(1/M, M)$ is to prove that the difficulty in the estimation problem does not simply arise because of poorly invertible covariance matrices.

Denote $\overline{R}^*[p, M]$ the minimax estimation risk of the proportion of explained variation η when the covariance matrix is unknown

$$\overline{R}^*[p, M] := \inf_{\widehat{\eta}} \sup_{\beta\in\mathbb{B}_0[p],\sigma>0} \sup_{\mathbf{\Sigma}\in\Xi[M]} \mathbb{E}_{\beta,\sigma,\mathbf{\Sigma}}\big[\big(\widehat{\eta} - \eta(\beta,\sigma)\big)^2\big], \tag{25}$$

where the infimum is taken over all estimator $\widehat{\eta}$ measurable with respect to $(Y, \mathbf{X})$.

When the covariance matrix $\mathbf{\Sigma}$ is known, the minimax rate has been shown to be of order at most p/n^2 and therefore goes to 0 as soon as p is small compared to n^2. The following proposition shows that, for unknown $\mathbf{\Sigma}$, there is no consistent estimators of η when p is large compared to n.

Theorem 4.5. *Consider an asymptotic setting where both n and p go to infinity. Then, there exists a positive function $C : (0, \infty) \times (1, \infty) \mapsto (0, 1)$ such that the following holds. If for some*

$\varsigma > 0$,

$$\frac{n^{1+\varsigma}}{p} \to 0, \tag{26}$$

then, for any $M > 1$, the minimax risk $\overline{R}^*[p, M]$ is bounded away from zero, that is $\underline{\lim}_{n,p} \overline{R}^*[p, M] \geq C(\varsigma, M)$.

Remark 4.3. Theorem 4.5 tells us that it is impossible to consistently estimate the proportion of explained variation in a high-dimensional setting where p is much larger than n. This lower bound straightforwardly extends to $\overline{R}^*[k, M]$ when k is much larger than n in the sense $n^{1+\varsigma}/k \to 0$ for some $\varsigma > 0$.

Let us get a glimpse of the proof by trying to build an estimator of $\eta(\beta^*, \sigma)$ in the high-dimensional regime $p \geq n$. As $\mathbf{\Omega}$ is unknown and cannot be consistently estimated in this regime, a natural candidate would be to consider $\widehat{\eta}^D(\mathbf{I}_p) = \widehat{V}$ as defined below (4). By Theorem 2.1, one has

$$\widehat{\eta}^D(\mathbf{I}_p) = \frac{\beta^{*T}\mathbf{\Sigma}^2\beta^*}{\operatorname{Var}(y_i)} + O_P\left(\frac{\sqrt{p}}{n}\right).$$

Although the signal strength $\beta^{*T}\mathbf{\Sigma}\beta^*$ cannot be consistently estimated for unknown $\mathbf{\Sigma}$ (Theorem 4.5), it is interesting to note that some regularized version of the signal strength $\beta^{*T}\mathbf{\Sigma}^2\beta^*$ is estimable at the rate p/n^2 (this phenomenon was already observed in [13]).

Going one step further, one can consistently estimate $\beta^{*T}\mathbf{\Sigma}^3\beta^*$ for $p \leq n^{3/2}$ by considering a quadratic form of Y as in T (4) but with higher-order polynomials of $\mathbf{X}$. For p of order $n^{1+\varsigma}$ for some small $\varsigma > 0$, it will be possible to consistently estimate all $a_q := \beta^{*T}\mathbf{\Sigma}^q\beta^*$ for $q = 2, 3, \dots, r(\varsigma)$ where $r(\varsigma)$ is a positive integer only depending on ς.

Then, one may wonder whether it is possible to reconstruct $a_1 = \beta^{*T}\mathbf{\Sigma}\beta^*$ from (a_q), $q = 2, \dots, r(\varsigma)$. Observe that a_q is the qth moment of a positive discrete measure μ supported by the spectrum of $\mathbf{\Sigma}$ and whose corresponding weights are the square norms of the projections of β^* on the eigenvectors of $\mathbf{\Sigma}$. As a consequence, estimating $\beta^{*T}\mathbf{\Sigma}\beta^*$ from (a_q), $q = 2, \dots, r(\varsigma)$ is a partial moment problem where one aims at recovering the first moment of the measure μ given its higher order moments up to $r(\varsigma)$. Following these informal arguments, we build, in the proof of Theorem 4.5, two discrete measures μ_1 and μ_2 supported on $(1/M, M)$ whose qth moments coincide for $q = 2, \dots, r(\varsigma)$ and whose first moments are far from each other. Define $\mathcal{B}_1$ (resp. $\mathcal{B}_2$) the collection of parameter $(\beta^*, \mathbf{\Sigma})$ whose corresponding measure is μ_1 (resp. μ_2). Then, we show that no test can consistently distinguish the hypothesis $H_0 : (\beta^*, \mathbf{\Sigma}) \in \mathcal{B}_1$ from $H_1 : (\beta^*, \mathbf{\Sigma}) \in \mathcal{B}_2$. As the signal strengths $\beta^{*T}\mathbf{\Sigma}\beta^*$ of parameters in $\mathcal{B}_1$ are far from those in $\mathcal{B}_2$, this implies that consistent estimation is impossible in this setting.

Remark 4.4. Let us summarize our findings on the minimax estimation risk when $\mathbf{\Sigma}$ is unknown and $n \leq p \leq n^2$:

- if k is small compared to $\sqrt{p}$, the minimax risk is of order $[k\log(p)/n \wedge 1]^2 + n^{-1}$ and is achieved by the square-root Lasso estimator $\widehat{\eta}^{\mathrm{SL}}$.
- if k is large compared to n (in the sense $n^{1+\varsigma}/k \to 0$ for some $\varsigma > 0$), then consistent estimation is impossible.
- if k lies between $\sqrt{p}$ and $n/\log(p)$, the square-root Lasso estimator $\widehat{\eta}^{\mathrm{SL}}$ is consistent at the rate $(k\log(p)/n)^2$. We conjecture that this rate is optimal.
- if k lies between $n/\log(p)$ and n, we are not aware of any consistent estimator η and we conjecture that consistent estimation is impossible.

5. Discussion and extensions

We focused in this work on the estimation risk of η in high-dimensional linear models under two major assumptions: the design is random (with possibly unknown covariance matrix) and the level of noise σ is unknown. We first discuss how the difficulty of the problem is modified when the two assumptions are not satisfied: when the design is not random, then consistent estimation of η is impossible in the dense regime, and when the level of noise is known, then the estimation of η becomes much easier in the dense regime. Finally, we mention the problem of constructing optimal confidence intervals.

5.1. Fixed design

If the regression design $\mathbf{X}$ is considered as fixed, then the counterpart of the proportion of explained variation would be

$$\eta[\beta^*, \sigma, \mathbf{X}] := \frac{\|\mathbf{X}\beta^*\|_2^2/n}{\|\mathbf{X}\beta^*\|_2^2/n + \sigma^2}.$$

In this new setting, the square-root Lasso estimator still estimates $\eta[\beta^*, \sigma, \mathbf{X}]$ at the rate $n^{-1} + (k\log(p)/n)^2$ up to multiplicative constants only depending on the sparse eigenvalues and compatibility constants of $\mathbf{X}$. In contrast, the construction of $\widehat{V}$ relies on the fact that $\mathbf{X}$ is random and is independent of the isotropic noise ε. When $\mathbf{X}$ is considered as fixed, $\widehat{V}$ does not consistently estimate $\eta[\beta^*, \sigma, \mathbf{X}]$ for p small compared to n^2. As a simple example, take $\sigma = 1$ and define β^* by $\beta^{*T} v_i = \lambda_i^{-1/2}$ for $i = 1, \dots, n$ where $(v_i)_i$ denote the right eigenvectors of $\mathbf{X}$ and $(\lambda_i^{1/2})_i$ its singular values. Then, the random variables T and $\hat{V}$ (defined in Section 2.1) are concentrated around 0, whereas $\eta[\beta^*, \sigma, \mathbf{X}]$ equals 1/2.

More generally, the next proposition states that it is impossible to consistently estimate $\eta[\beta^*, \sigma, \mathbf{X}]$ in a high-dimensional setting $p \geq n + 1$. The randomness of $\mathbf{X}$ therefore plays a fundamental role in the problem.

Proposition 5.1. *Assume that $p > n$ and consider any fixed design* $\mathbf{X}$ *such that* $\mathrm{Rank}(\mathbf{X}) = n$. *Given β^* and σ, denote $\underline{\mathbb{P}}_{\beta^*,\sigma}$ and $\underline{\mathbb{E}}_{\beta^*,\sigma}$ the probability and expectation with respect to the* *distribution* $Y = \mathbf{X}\beta^* + \varepsilon$ *with* $\varepsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I}_n)$. *Then, the minimax estimation risk satisfies*

$$\inf_{\widehat{\eta}} \sup_{\beta^* \in \mathbb{R}^p, \sigma \geq 0} \mathbb{E}_{\beta^*,\sigma}\big[\big(\widehat{\eta} - \eta[\beta^*, \sigma, \mathbf{X}]\big)^2\big] \geq \frac{1}{4}, \tag{27}$$

where the infimum is taken over all estimator $\widehat{\eta}$ *measurable with respect to* Y.

5.2. Knowledge of the noise level

Throughout this manuscript, we assumed that the noise level σ was unknown. As explained in the introduction, the situation is qualitatively different when σ is known. Let us briefly sketch the optimal convergence rates in this setting, still restricting ourselves to $p \geq n$. For any $k = 1, \dots, p$ define the maximal risk of an estimator $\widehat{\eta}$ and the minimax risk as

$$R(\widehat{\eta}, k, \sigma) := \sup_{\beta \in \mathbb{B}_0[k]} \mathbb{E}_{\beta,\sigma}\big[\big\{\hat{\eta} - \eta(\beta, \sigma)\big\}^2\big], \qquad R^*(k, \sigma) := \inf_{\hat{\eta}} R(\widehat{\eta}, k, \sigma).$$

It follows from the minimax lower bounds for signal detection [3,20], that for some $C > 0$ (lower bounds in [3,20] are asymptotic but it is not difficult to adapt the arguments to obtain non-asymptotic bounds to the price of worse multiplicative constants),

$$R^*(k, \sigma) \geq C\left(\frac{k}{n}\log\left(1 + \frac{p}{k^2}\right)\right)^2 \wedge \frac{1}{n}, \tag{28}$$

which is of order $[k\log(p)/n]^2 \wedge n^{-1}$ except in the regime where n is of order p and where k is of order $p^{1/2}$ in which case the logarithmic factors do not match. As for the upper bounds, since $\|Y\|_2^2/[\sigma^2 + \beta^{*T}\mathbf{\Sigma}\beta^*]$ follows a χ^2 distribution with n degrees of freedom, the estimator $\widehat{\eta}^{D,\sigma} := 1 - \frac{n\sigma^2}{\|Y\|_2^2}$ admits a quadratic risk (up to constants) smaller than $1/n$. This implies that the proportion of explained variation η can be efficiently estimated for arbitrarily large p. For small k, one can use the Gauss-Lasso estimator based on $\tilde{\beta}^{\mathrm{SL}}$. Let $\hat{J}$ be the set of integers j such that $\tilde{\beta}^{\mathrm{SL}} \neq 0$ and define:

$$\widehat{\eta}^{\mathrm{GL},\sigma} := \frac{\|\mathbf{\Pi}_{\hat{J}}Y\|_2^2/n}{\sigma^2 + \|\mathbf{\Pi}_{\hat{J}}Y\|_2^2/n}$$

where $\mathbf{\Pi}_{\hat{J}} = \boldsymbol{X}_{\hat{J}}(\boldsymbol{X}_{\hat{J}}^T\boldsymbol{X}_{\hat{J}})^{-1}\boldsymbol{X}_{\hat{J}}^T$ is the orthogonal projector of $\mathbb{R}^n$ onto the space spanned by the columns of $\boldsymbol{X}_{\hat{J}}$. The Gauss-Lasso estimator was introduced to get an estimator of heritability in the sparse situation in a first version of this work [32]. Following the proof of Theorem 2.3 in [32] we may obtain that, under Assumption (13) and when $|\beta^*|_0 = k$,

$$\mathbb{E}\big[\big(\widehat{\eta}^{\mathrm{GL},\sigma} - \eta\big)^2\big] \leq C'\frac{k^2\log^2(p)}{n^2}\frac{\lambda_{\max}^2(\mathbf{\Sigma})}{\lambda_{\min}^2(\mathbf{\Sigma})}.$$

In conclusion, the rate $[k\log(p)/n]^2 \wedge n^{-1}$ is (up to a possible logarithmic multiplicative term) optimal. These results contrast with the case of unknown σ in two ways: (i) The optimal rate is order-wise faster when σ is known especially when k is small (n^{-1} versus $(k\log(p)/n)^2$) and when k, p are larger (p/n^2 versus n^{-1}). (ii) Since $\widehat{\eta}^{D,\sigma}$ and $\widehat{\eta}^{\mathrm{GL},\sigma}$ do not use the knowledge of $\mathbf{\Sigma}$, adaptation to unknown covariance of the covariates is possible.

5.3. Minimax confidence intervals

In practice, one may not only be interested in the estimation of $\eta(\beta^*,\sigma)$, but also on building confidence intervals [21]. In the proof of Theorem 2.1 and in Proposition 2.3, we obtain exponential concentration inequalities of $\widehat{\eta}^D(\mathbf{\Omega})$ and $\widehat{\eta}^{\mathrm{SL}}$ around β^*. This allows to get, for any $\alpha>0$ and any $k=1,\dots,p$, confidence intervals

$$\mathrm{CI}_\alpha^D := \left[\widehat{\eta}^D(\mathbf{\Omega}) \pm C(\alpha)\frac{\sqrt{p}}{n}\right],$$

$$\mathrm{CI}_{\alpha,k}^{\mathrm{SL}} := \left[\widehat{\eta}^{\mathrm{SL}} \pm C'(\alpha)\left(\frac{1}{n^{1/2}} + \frac{k\log(p)}{n}\frac{\lambda_{\max}^2(\mathbf{\Sigma})}{\lambda_{\min}^2(\mathbf{\Sigma})}\right)\right],$$

where $C(\alpha)$ and $C'(\alpha)$ are universal constants only depending on α. When $p\geq n$, CI_α^D is honest over $\mathbb{R}^p$ in the sense that

$$\inf_{\beta\in\mathbb{B}_0[p],\sigma>0}\mathbb{P}_{\beta,\sigma}\left[\eta\in\mathrm{CI}_\alpha^D\right]\geq 1-\alpha.$$

For $p\geq n$ and if Assumption (13) is satisfied, then the confidence interval $\mathrm{CI}_{\alpha,k}^{\mathrm{SL}}$ is honest over $\mathbb{B}_0[k]$ in the sense that

$$\inf_{\beta\in\mathbb{B}_0[k],\sigma>0}\mathbb{P}_{\beta,\sigma}\left[\eta\in\mathrm{CI}_{\alpha,k}^{\mathrm{SL}}\right]\geq 1-\alpha.$$

In high-dimensional linear regressions, there have been recent advances towards the construction of optimal confidence regions both for the unknown vector β^* [26] or low-dimensional functional of the parameters such as components β_i^* [9,22,31,35] or $\sum_i \beta_i^*$ [9]. Building on this line of work, it seems at hand to prove the minimax optimality of CI_α^D and $\mathrm{CI}_{\alpha,k}^{\mathrm{SL}}$, proving the existence of such honest confidence intervals. Of course, as already noticed when constructing our adaptive estimator, the choice of the constants $C(\alpha)$ and $C'(\alpha)$ are probably far to be optimal in applications.

A further step would be to study the problem of the construction (if possible) of adaptive confidence intervals. We leave those important questions for future research.

6. Proofs of the minimax lower bounds

6.1. Proof of Proposition 2.4

6.1.1. *Proof of the parametric rate $R^*(k)\geq R^*(1)\geq Cn^{-1}$*

First, we prove that η cannot be estimated faster than the parametric rate n^{-1}. Fix $\sigma=1$, $\beta_1^*=(1,0,\dots,0)^T$ and $\beta_2^*=(1+n^{-1/2},0,\dots,0)^T$. Then $\eta_1=\eta(\beta_1^*,\sigma)=1/2$ and $\eta_2=\eta(\beta_2^*,\sigma)\geq$

$1/2+n^{-1/2}/4$. Denoting $\mathbb{K}(\mathbb{P}_{\beta_1^*,\sigma};\mathbb{P}_{\beta_2^*,\sigma})$ the Kullback-Leibler divergence between $\mathbb{P}_{\beta_1^*,\sigma}$ and $\mathbb{P}_{\beta_2^*,\sigma}$, we have

$$\mathbb{K}(\mathbb{P}_{\beta_1^*,\sigma};\mathbb{P}_{\beta_2^*,\sigma})=\mathbb{E}\left[\frac{\|\mathbf{X}(\beta_1^*-\beta_2^*)\|_2^2}{2}\right]=\frac{1}{2}.$$

Using Pinsker's inequality, we provide a lower bound of $R^*(1)$ in terms of $\mathbb{K}(\mathbb{P}_{\beta_1^*,\sigma};\mathbb{P}_{\beta_2^*,\sigma})$ and $(\eta_1-\eta_2)^2$ as follows:

$$\begin{aligned}
R^*(1) &\geq \inf_{\widehat{\eta}}\mathbb{E}_{\beta_1^*,\sigma}\big[(\widehat{\eta}-\eta_1)^2\big]\vee\mathbb{E}_{\beta_2^*,\sigma}\big[(\widehat{\eta}-\eta_2)^2\big]\\
&\geq \frac{(\eta_2-\eta_1)^2}{4}\inf_{\widehat{\eta}}\mathbb{P}_{\beta_1^*,\sigma}\big[\widehat{\eta}\geq(\eta_1+\eta_2)/2\big]\vee\mathbb{P}_{\beta_2^*,\sigma}\big[\widehat{\eta}\leq(\eta_1+\eta_2)/2\big]\\
&\geq \frac{(\eta_2-\eta_1)^2}{8}\inf_{\mathcal{A}}\mathbb{P}_{\beta_1^*,\sigma}(\mathcal{A})+\mathbb{P}_{\beta_2^*,\sigma}\big(\mathcal{A}^c\big), \qquad \text{where } \mathcal{A} \text{ is any measurable event}\\
&\geq \frac{(\eta_2-\eta_1)^2}{8}\big[1-\|\mathbb{P}_{\beta_1^*,\sigma}-\mathbb{P}_{\beta_2^*,\sigma}\|_{\mathrm{TV}}\big]\\
&\geq \frac{(\eta_2-\eta_1)^2}{8}\big[1-2^{-1/2}\mathbb{K}^{1/2}(\mathbb{P}_{\beta_1^*,\sigma};\mathbb{P}_{\beta_2^*,\sigma})\big], \qquad \text{by Pinsker's inequality}\\
&\geq \frac{(\eta_2-\eta_1)^2}{16}\geq\frac{1}{16^2n},
\end{aligned}$$

which concludes the proof.

6.1.2. *Proof of* $R^*(k)\geq C\{[\frac{k}{n}\log(1+\frac{p}{k^2}\vee\sqrt{\frac{p}{k^2}})]^2\wedge 1\}$

In this proof, we follow the standard strategy of reducing the heritability estimation problem to a detection problem, thereby taking advantage on available bounds of [34]. We could simply derive Proposition 2.4 from Theorem 4.3 in [34], but we prefer to detail the arguments as a first step towards the minimax lower bounds for adaptation problems.

Denote $\mathbb{P}_0$ the distribution of $(Y,\mathbf{X})$ when $\beta^*=0$ and $\sigma=1$. Let $\rho>0$ be a positive quantity that will be fixed later. Also, denote $\mathcal{B}$ the collection of all vectors $\beta\in\mathbb{R}^p$ with exactly k nonzero components that are either equal to $\frac{\rho}{[(1+\rho^2)k]^{1/2}}$ or $-\frac{\rho}{[(1+\rho^2)k]^{1/2}}$. Defining $\sigma_\rho^2:=(1+\rho^2)^{-1}$, we obtain, for all $\beta\in\mathcal{B}$, $\eta(\beta,\sigma_\rho)=\rho^2/(1+\rho^2)$. Following the beaten path of Le Cam's approach, we consider μ the uniform measure on $\mathcal{B}$ and denote $\mathbf{P}_\mu$ the mixture probability measure

$$\mathbf{P}_\mu=\int_{\mathcal{B}}\mathbb{P}_{\beta,\sigma_\rho}\mu(d\beta). \tag{29}$$

The minimax risk $R^*(k)$ is obviously lower bounded as follows:

$$\begin{aligned}R^*(k) &\geq \inf_{\widehat{\eta}}\left\{\mathbb{E}_0\big[\widehat{\eta}^2\big] \vee \bigvee_{\beta\in\mathcal{B}} \mathbb{E}_{\beta,\sigma_\rho}\left[\left(\widehat{\eta}-\frac{\rho^2}{1+\rho^2}\right)^2\right]\right\}\\
&\geq \frac{1}{2}\inf_{\widehat{\eta}}\left[\mathbb{E}_0\big[\widehat{\eta}^2\big]+\mathbf{E}_\mu\left[\left(\widehat{\eta}-\frac{\rho^2}{1+\rho^2}\right)^2\right]\right]\\
&\geq \frac{\rho^4}{8(1+\rho^2)^2}\inf_{\widehat{\eta}}\left[\mathbb{P}_0\left[\widehat{\eta}>\frac{\rho^2}{2(1+\rho^2)}\right]+\mathbf{P}_\mu\left[\widehat{\eta}\leq\frac{\rho^2}{2(1+\rho^2)}\right]\right].\end{aligned}$$

Defining the test statistic $\widehat{T} := \mathbf{1}\{\widehat{\eta} > \rho^2/[2(1+\rho^2)]\}$, one recognizes in the bound above the sum of type I and type II errors of a test of $\mathbb{P}_0$ versus $\mathbf{P}_\mu$. Taking the infimum over all tests $\widehat{T}$, that is all measurables function of $(Y, \mathbf{X})$ to $\{0, 1\}$, we arrive at

$$\begin{aligned}R^*(k) &\geq \frac{\rho^4}{8(1+\rho^2)^2}\inf_{\widehat{T}}\big[\mathbb{P}_0[\widehat{T}=1]+\mathbf{P}_\mu[\widehat{T}=0]\big]\\
&\geq \frac{\rho^4}{8(1+\rho^2)^2}\inf_{\widehat{T}}\big[1-\big|\mathbb{P}_0(\widehat{T}=0)-\mathbf{P}_\mu(\widehat{T}=0)\big|\big]\\
&\geq \frac{\rho^4}{8(1+\rho^2)^2}\big[1-\mathbb{E}_0|L_\mu-1|\big] \qquad \text{where } \mathbb{L}_\mu=\frac{d\mathbf{P}_\mu}{d\mathbb{P}_0}\\
&\geq \frac{\rho^4}{8(1+\rho^2)^2}\big[1-\big(\chi^2(\mathbf{P}_\mu,\mathbb{P}_0)\big)^{1/2}\big] \qquad \text{(by Cauchy–Schwarz inequality)}\end{aligned} \tag{30}$$

where $\chi^2(\mathbf{P}_\mu, \mathbb{P}_0) = \mathbb{E}_0[(L_\mu - 1)^2]$ stands for the χ^2 distance between probability distributions. As a consequence, we only need to bound the χ^2 distance between $\mathbf{P}_\mu$ and $\mathbb{P}_0$. Fortunately, this distance has been controlled in [34] (take $v = 1$, $\mathrm{Var}(y) = 1$ in [34], p. 741, line 14, and note that $k\lambda^2 = \rho^2/(1 + \rho^2)$).

Lemma 6.1 ([34]). *We have*

$$\chi^2(\mathbf{P}_\mu,\mathbb{P}_0)\leq \exp\left[k\log\left(1+\frac{k}{p}\left(\cosh\left(\frac{n\rho^2}{k}\right)-1\right)\right)\right]-\frac{1}{2}. \tag{31}$$

Let us fix ρ^2 in such a way that

$$\frac{n\rho^2}{k}=\log\left[1+\frac{p}{k^2}\log(5/4)+\sqrt{\left(1+\frac{p}{k^2}\log(5/4)\right)^2-1}\right]. \tag{32}$$

Using the classical equality $\cosh(\log(1 + x + \sqrt{x^2 + 2x})) = 1 + x$ for $x \geq 0$, we arrive at

$$\chi^2(\mathbf{P}_\mu,\mathbb{P}_0)\leq \exp\big[k\log\big(1+\log(5/4)/k\big)\big]-1/2\leq 3/4,$$

which, together with (30), implies

$$R^*(k) \geq \frac{\rho^4}{8(1+\rho^2)^2}\big(1-(3/4)^{1/2}\big).$$

Since $\log(1+ux) \geq u\log(1+x)$ for any $u \in (0,1)$ and $x > 0$, we derive from (32) that

$$\rho^2 \geq \log\left(\frac{5}{4}\right)\left[\frac{k}{n}\log\left(1+\frac{p}{k^2}\vee\sqrt{\frac{p}{k^2}}\right)\right]^2.$$

But

$$\frac{\rho^2}{1+\rho^2} \geq \frac{\rho^2}{2}\wedge 1,$$

which concludes the proof.

6.2. Proof of Proposition 3.1

Define the quantity $\rho > 0$ by

$$\rho^2 := \frac{a\sqrt{p\log p}}{4n}. \tag{33}$$

We consider μ, $\mathbf{P}_\mu$, $\mathbf{E}_\mu$ as introduced in the proof of Proposition 2.4.

Let $\widehat{\eta}$ be any given estimator. Define

$$R := n\sqrt{\frac{n}{p}}\mathbf{E}_0\big[\widehat{\eta}^2\big] + \frac{n^2}{p\log p}\mathbf{E}_\mu\left[\left(\widehat{\eta}-\frac{\rho^2}{1+\rho^2}\right)^2\right].$$

Then,

$$\frac{R(\widehat{\eta},1)}{\frac{1}{n}\sqrt{\frac{p}{n}}} + \frac{R(\widehat{\eta},k)}{\frac{p\log p}{n^2}} \geq R.$$

Now define the event $\mathcal{A}(\widehat{\eta}) := \{\widehat{\eta} \geq \rho^2/[2(1+\rho^2)]\}$. Then, one has

$$n\sqrt{\frac{n}{p}}\mathbf{E}_0\big[\widehat{\eta}^2\big] \geq n\sqrt{\frac{n}{p}}\mathbb{P}_0\big[\mathcal{A}(\widehat{\eta})\big]\frac{\rho^4}{4(1+\rho^2)^2} \geq \frac{a^2}{4^4}\sqrt{\frac{p}{n}}\log p\mathbb{P}_0\big[\mathcal{A}(\widehat{\eta})\big].$$

Similarly, $\mathbf{E}_\mu(\widehat{\eta}-\rho^2/(1+\rho^2))^2 \geq \mathbf{P}_\mu(\mathcal{A}^c(\widehat{\eta}))\rho^4/4(1+\rho^2)^2 \geq \frac{a^2}{4^4}\frac{p\log p}{n^2}\mathbf{P}_\mu(\mathcal{A}^c(\widehat{\eta}))$ so that

$$R \geq \frac{a^2}{4^4}\inf_{\mathcal{A}}\left\{\mathbb{P}_0[\mathcal{A}]\sqrt{\frac{p}{n}}\log p + \mathbf{P}_\mu\big[\mathcal{A}^c\big]\right\},$$

where the infimum is taken over all measurable events $\mathcal{A}$. Restricting the events $\mathcal{A}$ to have small probability, we arrive at

$$R \geq \frac{a^2}{4^4}\Big\{1 \wedge \inf_{\mathcal{A}, \mathbb{P}_0[\mathcal{A}] \leq \sqrt{n}/(\sqrt{p}\log p)} \mathbf{P}_\mu\big[\mathcal{A}^c\big]\Big\}, \tag{34}$$

so that it suffices to obtain a uniform lower bound for $\mathbf{P}_\mu[\mathcal{A}^c]$ over events $\mathcal{A}$ of small $\mathbb{P}_0$-probability.

$$\begin{aligned}
\mathbf{P}_\mu\big(A^c\big) &\geq 1 - \mathbb{P}_0(\mathcal{A}) - \big|\mathbf{P}_\mu(\mathcal{A}) - \mathbb{P}_0(\mathcal{A})\big| \\
&\geq 1 - \mathbb{P}_0(\mathcal{A}) - \big|\mathbf{E}_0\big[(\mathbb{L}_\mu - 1)\mathbf{1}_{\mathcal{A}}\big]\big| \qquad \text{where } \mathbb{L}_\mu = \frac{d\mathbf{P}_\mu}{d\mathbb{P}_0} \\
&\geq 1 - \mathbb{P}_0(\mathcal{A}) - \big(\mathbb{P}_0[\mathcal{A}]\chi^2(\mathbf{P}_\mu, \mathbb{P}_0)\big)^{1/2} \qquad \text{(by Cauchy–Schwarz inequality)}
\end{aligned} \tag{35}$$

Define $x = \frac{ap}{2k^2}\log(p)$. Since $\sqrt{p\log p} \leq k$, and since $\log(1+ux) \geq u\log(1+x)$ for any $u \in (0,1)$ and $x > 0$, we have

$$\frac{n\rho^2}{k} \leq \log[1 + x \vee \sqrt{x}] \leq \log\big[1 + x + \sqrt{2x + x^2}\big].$$

Together with Lemma 6.1 and the classical identity $\cosh[\log(1 + u + \sqrt{2u + u^2})] = 1 + u$ for all $u > 0$, we arrive at

$$\frac{\sqrt{n}}{\sqrt{p}\log p}\chi^2(\mathbf{P}_\mu, \mathbb{P}_0) \leq \frac{\sqrt{n}}{\sqrt{p}\log p}\exp\left[\frac{k^2}{p}x\right] \leq \frac{\sqrt{n}}{\sqrt{p}\log p}p^{a/2} = \sqrt{\frac{n}{p^{1-a}}\frac{1}{\log p}}. \tag{36}$$

Coming back to the lower bound (35), we conclude that, for any event $\mathcal{A}$ satisfying $\mathbb{P}_0(\mathcal{A}) \leq \sqrt{n}/(\sqrt{p}\log p)$, we have

$$\mathbf{P}_\mu\big(A^c\big) \geq 1 - \sqrt{\frac{n}{p}\frac{1}{\log p}} - \left(\sqrt{\frac{n}{p^{1-a}}\frac{1}{\log p}}\right)^{1/2}.$$

Plugging this result in (34) and using the fact that $p^{1-a}(\log p)^2 \geq 16n$ leads to the desired result.

6.3. Proof of Theorem 4.5

6.3.1. *General arguments*

Fix $M > 1$ and suppose that Condition (26) is satisfied for some $\varsigma > 0$. Define r to be the smallest integer such that $\varsigma \geq 1/(2r)$ so that we can assume henceforth that $n^{1+1/(2r)}/p \to 0$.

In this proof, we follow the same general approach as in the other minimax lower bounds, that is we define two mixture distributions $\mathbf{P}_0$ and $\mathbf{P}_1$

$$\mathbf{P}_0 := \int \mathbb{P}_{\beta,\sigma_0,\mathbf{\Sigma}}\mu_0(d\beta, d\mathbf{\Sigma}), \qquad \mathbf{P}_1 := \int \mathbb{P}_{\beta,\sigma_1,\mathbf{\Sigma}}\mu_1(d\beta, d\mathbf{\Sigma}),$$

in such a way that $\mathbf{P}_0$ and $\mathbf{P}_1$ are almost indistinguishable and at the same time the function $\eta(\beta, \sigma)$ takes different values for parameters in the support of the prior distribution μ_0 and parameters in the support of the prior distribution μ_1. The main difference with previous proofs lies in the fact that μ_0 and μ_1 are now prior probabilities on both the regression coefficient vector β and the covariance matrix $\mathbf{\Sigma}$.

Let $\alpha_0 = (\alpha_{i,0})$, $\gamma_0 = (\gamma_{i,0})$, $i = 1, \dots r$ and $\alpha_1 = (\alpha_{i,1})$, $\gamma_1 = (\gamma_{i,1})$, $i = 0, \dots r$ be positive parameters whose exact values will be fixed later. We emphasize that the values of these parameters will only depend on r and not on n and p. Given a positive integer q and $\alpha = (\alpha_1, \dots, \alpha_q)$ whose coordinates α_j are positive, define the probability distribution π_α on vectors of $\mathbb{R}^{q \times p}$ whose density is proportional to $(|\mathbf{I}_p + \sum_{i=1}^q \alpha_i x_i x_i^T|)^{-n/2} e^{-\sum_{i=1}^q p\|x_i\|_2^2/2}$ for $x_i \in \mathbb{R}^p$, $i = 1, \dots, q$.

The distribution μ_0 is defined as follows. Let $(v_{i,0})$, $i = 1, \dots, r$ be independently sampled according to the distribution π_{α_0}. Then, conditionally to $(v_{1,0}, \dots, v_{r,0})$, β and $\mathbf{\Sigma}$ are fixed to the following values

$$\beta = \sum_{i=1}^{r} \gamma_{i,0} v_{i,0}; \qquad \mathbf{\Sigma}^{-1} = \mathbf{I}_p + \sum_{i=1}^{r} \alpha_{i,0} v_{i,0} v_{i,0}^T. \tag{37}$$

Similarly, under μ_1,

$$\beta = \sum_{i=0}^{r} \gamma_{i,1} v_{i,1}; \qquad \mathbf{\Sigma}^{-1} = \mathbf{I}_p + \sum_{i=0}^{r} \alpha_{i,1} v_{i,1} v_{i,1}^T, \tag{38}$$

where the vectors $(v_{i,1})$, $i = 0, \dots, r$ are independently sampled according to the distribution π_{α_1}. Finally, the noise variances are fixed to the following values.

$$\sigma_0^2 = 1, \qquad \sigma_1^2 = 1 - \zeta(r, M), \tag{39}$$

where $\zeta(r, M) > 0$ is introduced in Lemma 6.2.

To prove that $\mathbf{P}_0$ and $\mathbf{P}_1$ are almost indistinguishable we will consider separately the marginal distribution of $\mathbf{X}$ and the conditional distribution of Y given $\mathbf{X}$. We will see that the centered Gaussian distribution of $\mathbf{X}$ under both $\mathbf{P}_0$ and $\mathbf{P}_1$ are indistinguishable from the standard normal distribution when $n = o(p)$, see Lemma 6.4 below.

Let us now choose the parameters $\gamma_{i,j}$ and $\alpha_{i,j}$ in such a way that the conditional distribution of Y given $\mathbf{X}$ under $\mathbf{P}_0$ is indistinguishable from that under $\mathbf{P}_1$ when $n^{1+1/(2r)} = o(p)$. We first consider a truncated moment problem.

Lemma 6.2. *There exist two discrete positive measures $\rho_0 = \sum_{i=1}^r \xi_{i,0} \delta_{\tau_{i,0}}$ on $\rho_1 = \sum_{i=0}^r \xi_{i,1} \delta_{\tau_{i,1}}$ supported on $(0, 1)$ such that*

1. *The atoms $\tau_{i,j}$ for $j = 0, 1$ lie in $[(5 + M)/(6M), 1)$.*
2. *The total mass of ρ_0 equals 1, whereas the total mass of ρ_1 is $1 + \zeta(r, M)$, where $\zeta(r, M) > 0$ is introduced in the proof.*
3. *For all $q = 1, \dots, 2r - 1$, the qth moment of ρ_0 and ρ_1 coincide*

$$\int x^q \, d\rho_0 = \int x^q \, d\rho_1 = \frac{2M}{M-1} \int_{(3+M)/(4M)}^{(1+3M)/(4M)} x^q \, dx := m_q.$$

For $j = 0, 1$, we set the values $\gamma_{i,j} := [\xi_{i,j}/\tau_{i,j}]^{1/2}$ and $\alpha_{i,j} := \tau_{i,j}^{-1} - 1$.

Let us give a hint why such a choice leads to what we need. As a consequence of our parameter choices, the following identities are satisfied

$$\sum_{i=1}^{r} \frac{\gamma_{i,0}^2}{1+\alpha_{i,0}} + \sigma_0^2 = \sum_{i=0}^{r} \frac{\gamma_{i,1}^2}{1+\alpha_{i,1}} + \sigma_1^2 = 2, \tag{40}$$

$$\sum_{i=1}^{r} \frac{\gamma_{i,0}^2}{(1+\alpha_{i,0})^q} = \sum_{i=0}^{r} \frac{\gamma_{i,1}^2}{(1+\alpha_{i,1})^q} = m_{q-1}, \qquad \forall q = 2, \dots, 2r. \tag{41}$$

Had the random vectors $(v_{i,j})$ introduced in μ_j formed an orthonormal family, then we would have had $\beta^T \mathbf{\Sigma}^q \beta = \sum_i \frac{\gamma_{i,j}^2}{(1+\alpha_{i,j})^q}$ for any positive integer q. We shall prove later that, under the distribution μ_j, the vectors $v_{i,j}$ have a norm close to one and are almost orthogonal with large probability. Hence, identities (41) imply that the moments $\beta^T \mathbf{\Sigma}^q \beta$ concentrate around the same value under μ_0 and μ_1, this for all $q = 2, \dots, 2r$. This will lead to the fact that the conditional distribution of Y given $\mathbf{X}$ under $\mathbf{P}_0$ is indistinguishable from that under $\mathbf{P}_1$ when $n^{1+1/(2r)} = o(p)$ as proved in Lemma 6.5 below. In the same way, (40) will imply that $\beta^T \mathbf{\Sigma} \beta + \sigma_j^2$ concentrate around 2 under μ_j for $j = 0, 1$ so that η will concentrate around different values under $\mathbf{P}_0$ and $\mathbf{P}_1$ since $\sigma_0^2 \neq \sigma_1^2$. This is stated in Lemma 6.3 below.

Let us now define the quantities

$$\eta_0 := 1 - \frac{\sigma_0^2}{\sum_{i=1}^r \frac{\gamma_{i,0}^2}{1+\alpha_{i,0}} + \sigma_0^2} = 1/2, \qquad \eta_1 := 1 - \frac{\sigma_1^2}{\sum_{i=0}^r \frac{\gamma_{i,1}^2}{1+\alpha_{i,1}} + \sigma_1^2} = 1/2 + \zeta(r, M)/2. \tag{42}$$

The next lemma states that, for $j = 0, 1$, $\eta(\beta, \sigma)$ is close to η_j under μ_j.

Lemma 6.3. *There exists three positive constants $C_1(r, M)$, $C_2(r, M)$ and $C_3(r)$ such that the following holds for $p \geq nC_1(r, M)$. First, one has*

$$\mu_j\left[\left|\eta(\beta, \sigma) - \eta_j\right| \geq C_2(r, M)\left(p^{-1/4} + \left(\frac{n}{p}\right)^{1/2}\right)\right] \leq e^{-C_3(r)p^{1/2}}, \tag{43}$$

for $j = 0, 1$. Also, the spectrum of $\mathbf{\Sigma}$ is bounded away from zero with large probability, that is for $j = 0, 1$,

$$\mu_j\left[\lambda_{\min}(\mathbf{\Sigma}) \leq 1/M\right] \leq e^{-C_3(r)p^{1/2}}. \tag{44}$$

By definition of μ_0 and μ_1, the largest eigenvalue of $\mathbf{\Sigma}$ is always equal to one. By Lemma 6.3, with μ_0 and μ_1 probability going to one, the spectrum of $\mathbf{\Sigma}$ lies in $[1/M, M]$.

Let us now bound the minimax risk $\overline{R}^*[p, M]$. Contrary to the prior distributions chosen in the proof of Proposition 3.1, the proportion of explained variation $\eta(\beta, \sigma)$ is not constant either on μ_0 or on μ_1, so that we cannot directly relate the minimax estimation rate to the total variation distance as done before. Nevertheless, these proportions of explained variation concentrate around

η_0 and η_1 so that it will be possible to work around this difficulty. This slight refinement of Le Cam's method has already been applied for other functional estimation problems (see, e.g., [11]). Also to circumvent the issue that some eigenvalues of $\mathbf{\Sigma}$ are smaller than $1/M$ with positive (but very small) probability, we consider a threshold version of the risk $\mathbf{E}_1^*[\cdot] := \mathbf{E}_1[\cdot \mathbf{1}_{\lambda_{\min}(\mathbf{\Sigma})\geq M^{-1}}]$ and $\mathbf{E}_0^*[\cdot] := \mathbf{E}_0[\cdot \mathbf{1}_{\lambda_{\min}(\mathbf{\Sigma})\geq M^{-1}}]$.

Without loss of generality, we may assume that all the estimators $\widehat{\eta}$ below only take values in [0, 1].

$$\begin{aligned}\overline{R}^*[p,M] &\geq \inf_{\widehat{\eta}} \mathbf{E}_0^*\big[\{\widehat{\eta}-\eta(\beta,\sigma)\}^2\big] \vee \mathbf{E}_1^*\big[\{\widehat{\eta}-\eta(\beta,\sigma)\}^2\big] \\ &\geq \inf_{\widehat{\eta}} \mathbf{E}_0\big[\{\widehat{\eta}-\eta(\beta,\sigma)\}^2\big] \vee \mathbf{E}_1\big[\{\widehat{\eta}-\eta(\beta,\sigma)\}^2\big] - \bigvee_{i=0,1} \mu_i\big[\lambda_{\min}(\mathbf{\Sigma}) \leq M^{-1}\big] \\ &\geq \inf_{\widehat{\eta}} \frac{1}{2} \bigvee_{i=1,2} \mathbf{E}_i\big[\{\widehat{\eta}-\eta_i\}^2\big] - \bigvee_{i=1,2} \mathbf{E}_i\big[\{\eta(\beta,\sigma)-\eta_i\}^2\big] - \bigvee_{i=0,1} \mu_i\big[\lambda_{\min}(\mathbf{\Sigma}) \leq M^{-1}\big],\end{aligned}$$

where we used $(x-y)^2 \geq (x-z)^2/2-(y-z)^2$. From (43) and the fact that $\eta(\beta,\sigma)$ belongs to [0, 1], we derive that for some positive constant $C(r, M)$,

$$\bigvee_{i=1,2} \mathbf{E}_i\big[\{\eta(\beta,\sigma)-\eta_i\}^2\big] \leq C(r,M)\big(p^{-1/2}+n/p\big),$$

when p is large enough. Besides, the probabilities $\mu_i[\lambda_{\min}(\mathbf{\Sigma}) \leq M^{-1}]$ are smaller than $e^{-C_3(r)p^{1/2}}$ by (44). Then, we control the maximum $\bigvee_{i=1,2} \mathbf{E}_i[\{(\widehat{\eta}-\eta_i\}^2]$ using the total variation distance between $\mathbf{P}_0$ and $\mathbf{P}_1$ as we did in the proof of Proposition 2.4. More precisely,

$$\begin{aligned}\overline{R}^*[p,M]+C(r,M)\big[p^{-1/2}+(n/p)\big] &\geq \frac{(\eta_1-\eta_0)^2}{8} \inf_{\widehat{\eta}} \mathbf{P}_0\bigg(\widehat{\eta} \geq \frac{\eta_1+\eta_0}{2}\bigg) \vee \mathbf{P}_1\bigg(\widehat{\eta} \leq \frac{\eta_1+\eta_0}{2}\bigg) \\ &\geq \frac{(\eta_1-\eta_0)^2}{16} \inf_{\mathcal{A}} \mathbf{P}_0(\mathcal{A})+\mathbf{P}_1\big(\mathcal{A}^c\big) \\ &\geq \frac{(\eta_1-\eta_0)^2}{16}\big[1-\|\mathbf{P}_1-\mathbf{P}_0\|_{\mathrm{TV}}\big],\end{aligned}$$

so that we only have to focus on $\|\mathbf{P}_1-\mathbf{P}_0\|_{\mathrm{TV}}$. Let us decompose the total variation distance between $\mathbf{P}_0$ and $\mathbf{P}_1$ in a way enabling to consider separately the marginal distribution of $\mathbf{X}$ and the conditional distributions of Y given $\mathbf{X}$. Since the total variation distance is, up to a multiplicative constant, the l_1 distance between the density functions, we obtain

$$\begin{aligned}2\|\mathbf{P}_1-\mathbf{P}_0\|_{\mathrm{TV}} &= \int \big|f_0(y,\mathbf{x})-f_1(y,\mathbf{x})\big|\,dy\,d\mathbf{x} \\ &= \int \big|f_0(y|\mathbf{x})f_0(\mathbf{x})-f_1(y|\mathbf{x})f_1(\mathbf{x})\big|\,dy\,d\mathbf{x}\end{aligned}$$

$$\begin{aligned}
&\leq \int f_1(y|\mathbf{x})\big|f_0(\mathbf{x}) - f_1(\mathbf{x})\big|\,dy\,d\mathbf{x} + \int f_0(x)\big|f_0(y|\mathbf{x}) - f_1(y|\mathbf{x})\big|\,dy\,d\mathbf{x} \\
&\leq \int \big|f_0(\mathbf{x}) - f_1(\mathbf{x})\big|\,d\mathbf{x} + \int f_0(x)\big|f_0(y|\mathbf{x}) - f_1(y|\mathbf{x})\big|\,dy\,d\mathbf{x} \\
&\leq 2\big\|\mathbf{P}_0^{\mathbf{X}} - \mathbf{P}_1^{\mathbf{X}}\big\|_{\mathrm{TV}} + 2\mathbf{E}_0^{\mathbf{X}}\big[\big\|\mathbf{P}_0^{Y|\mathbf{X}} - \mathbf{P}_1^{Y|\mathbf{X}}\big\|_{\mathrm{TV}}\big], \qquad (45)
\end{aligned}$$

where, for $i = 0, 1$, $\mathbf{P}_i^{\mathbf{X}}$ (resp. f_i) denotes the marginal probability distribution (resp. density) of $\mathbf{X}$ under $\mathbf{P}_i$, $\mathbf{P}_i^{Y|\mathbf{X}}$ (resp. $f_i(\cdot|\mathbf{x})$) is the conditional distribution (resp. density) of Y given $\mathbf{X}$ and $\mathbf{E}_0^{\mathbf{X}}$ stands for the expectation with respect to $\mathbf{P}_0^{\mathbf{X}}$. The main difficulty in the proof lies in controlling these two total deviation distances $\|\mathbf{P}_0^{\mathbf{X}} - \mathbf{P}_1^{\mathbf{X}}\|_{\mathrm{TV}}$ and $\mathbf{E}_0^{\mathbf{X}}[\|\mathbf{P}_0^{Y|\mathbf{X}} - \mathbf{P}_1^{Y|\mathbf{X}}\|_{\mathrm{TV}}]$.

The marginal distribution of $\mathbf{X}$ under $\mathbf{P}_0$ and $\mathbf{P}_1$ is that of a n sample of p-dimensional normal distribution whose precision matrix is a rank r perturbation of the identity matrix and whose r principal directions are sampled nearly uniformly. In a high-dimensional setting, such perturbations are indistinguishable from the standard normal distribution as shown in the next lemma.

Lemma 6.4. *There exist two positive constants $C(r, M)$ and $C'(r, M)$ only depending on r and M such that the following holds. If $p \geq C(r, M)n$, then*

$$\big\|\mathbf{P}_0^{\mathbf{X}} - \mathbf{P}_1^{\mathbf{X}}\big\|_{\mathrm{TV}} \leq C'(r, M)\sqrt{\frac{n}{p}}.$$

The intricate construction of μ_0 and μ_1 (and especially the choices of the parameters $\alpha_{i,j}$ and $\gamma_{i,j}$) has been made to force the conditional $\mathbf{P}_0^{Y|\mathbf{X}}$ and $\mathbf{P}_1^{Y|\mathbf{X}}$ to be close to each other. Informally, the fact that the quantities $\beta^T \mathbf{\Sigma}^q \beta$ almost coincide under μ_0 and μ_1, this for all $q = 2, \dots, 2r$, will translate into the total distance $\|\mathbf{P}_0^{Y|\mathbf{X}} - \mathbf{P}_1^{Y|\mathbf{X}}\|_{\mathrm{TV}}$ as illustrated by the next lemma.

Lemma 6.5. *There exist two positive constants $C(r, M)$ and $C'(r, M)$ only depending on r and M such that the following holds. If $p \geq C(r, M)n$, then*

$$\mathbf{E}_0^{\mathbf{X}}\big[\big\|\mathbf{P}_0^{Y|\mathbf{X}} - \mathbf{P}_1^{Y|\mathbf{X}}\big\|_{\mathrm{TV}}\big] \leq C'(r, M)\left(\frac{n^{1+1/(2r)}}{p}\right)^r.$$

Under assumption (26), the distance $\|\mathbf{P}_1 - \mathbf{P}_0\|_{\mathrm{TV}}$ goes to 0, and the minimax risk $\overline{R}^*[p, M]$ is therefore bounded away from zero:

$$\underline{\lim}\overline{R}^*[p, M] \geq \frac{(\eta_1 - \eta_0)^2}{32} \geq C\zeta^2(r, M).$$

Acknowledgements

We would like to thank the anonymous referees for their comments that helped us to improve the presentation of the manuscript.

Supplementary Material

Supplement to "Adaptive estimation of high-dimensional signal-to-noise ratios" (DOI: 10.3150/17-BEJ975SUPP; .pdf). This supplement contains the remaining proofs.

References

[1] Adamczak, R. and Wolff, P. (2015). Concentration inequalities for non-Lipschitz functions with bounded derivatives of higher order. *Probab. Theory Related Fields* **162** 531–586. MR3383337

[2] Amaral, D.G., Schumann, C.M. and Nordahl, C.W. (2008). Neuroanatomy of autism. *Trends Neurosci.* **31** 137–145.

[3] Arias-Castro, E., Candès, E.J. and Plan, Y. (2011). Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *Ann. Statist.* **39** 2533–2556. MR2906877

[4] Baraud, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli* **8** 577–606. MR1935648

[5] Bayati, M., Erdogdu, M.A. and Montanari, A. (2013). Estimating lasso risk and noise level. In *Advances in Neural Information Processing Systems* 944–952.

[6] Belloni, A., Chernozhukov, V. and Wang, L. (2011). Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika* **98** 791–806. MR2860324

[7] Bonnet, A., Gassiat, E. and Lévy-Leduc, C. (2015). Heritability estimation in high dimensional sparse linear mixed models. *Electron. J. Stat.* **9** 2099–2129. MR3400534

[8] Cai, T., Liu, W. and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. MR2847973

[9] Cai, T.T. and Guo, Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *Ann. Statist.* **45** 615–646. MR3650395

[10] Cai, T.T. and Low, M.G. (2005). Nonquadratic estimators of a quadratic functional. *Ann. Statist.* **33** 2930–2956. MR2253108

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

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

[13] Dicker, L.H. (2014). Variance estimation in high-dimensional linear models. *Biometrika* **101** 269–284. MR3215347

[14] Dicker, L.H. and Erdogdu, M.A. (2016). Maximum likelihood for variance estimation in high-dimensional linear models. In *Proceedings of the* 19*th International Conference on Artificial Intelligence and Statistics* 159–167.

[15] Donoho, D.L. and Nussbaum, M. (1990). Minimax quadratic estimation of a quadratic functional. *J. Complexity* **6** 290–323. MR1081043

[16] Fan, J., Guo, S. and Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 37–65. MR2885839

[17] Goldstein, D.B. (2009). Common genetic variation and human traits. *N. Engl. J. Med.* **360** 1696–1698.

[18] Guo, Z., Wang, W., Cai, T. and Li, H. (2016). Optimal estimation of co-heritability in high-dimensional linear models. arXiv preprint, arXiv:1605.07244.

[19] Ingster, Yu.I. and Suslina, I.A. (2003). *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*. *Lecture Notes in Statistics* **169**. New York: Springer. MR1991446

[20] Ingster, Y.I., Tsybakov, A.B. and Verzelen, N. (2010). Detection boundary in sparse regression. *Electron. J. Stat.* **4** 1476–1526. MR2747131

[21] Janson, L., Barber, R.F. and Candès, E. (2015). Eigenprism: Inference for high-dimensional signal-to-noise ratios. arXiv preprint, arXiv:1505.02097.

[22] Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15** 2869–2909. MR3277152

[23] Javanmard, A. and Montanari, A. (2015). De-biasing the lasso: Optimal sample size for Gaussian designs. arXiv preprint, arXiv:1508.02757.

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

[25] Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature* **456** 18–21.

[26] Nickl, R. and van de Geer, S. (2013). Confidence sets in sparse regression. *Ann. Statist.* **41** 2852–2876. MR3161450

[27] Steen, R.G., Mull, C., Mcclure, R., Hamer, R.M. and Lieberman, J.A. (2006). Brain volume in first-episode schizophrenia. *Br. J. Psychiatry* **188** 510–518.

[28] Stein, J.L., Medland, S.E., Vasquez, A.A., Hibar, D.P., Senstad, R.E., Winkler, A.M., Toro, R., Appel, K., Bartecek, R. and Bergmann, Ø. (2012). Identification of common variants associated with human hippocampal and intracranial volumes. *Nat. Genet.* **44** 552–561.

[29] Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99** 879–898. MR2999166

[30] Toro, R., Poline, J.-B., Huguet, G., Loth, E., Frouin, V., Banaschewski, T., Barker, G.J., Bokde, A., Büchel, C., Carvalho, F., Conrod, P., Fauth-Bühler, M., Flor, H., Gallinat, J., Garavan, H., Gowloan, P., Heinz, A., Ittermann, B., Lawrence, C., Lemaître, H., Mann, K., Nees, F., Paus, T., Pausova, Z., Rietschel, M., Robbins, T., Smolka, M., Ströhle, A., Schumann, G. and Bourgeron, T. (2015). Genomic architecture of human neuroanatomical diversity. *Mol. Psychiatry* **20** 1011–1016.

[31] van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285

[32] Verzelen, N. and Gassiat, E. (2016). Adaptive estimation of high-dimensional signal-to-noise ratios (version 1). arXiv preprint, arXiv:1602.08006v1.

[33] Verzelen, N. and Gassiat, E. (2017). Supplement to "Adaptive estimation of high-dimensional signal-to-noise ratios." DOI:10.3150/17-BEJ975SUPP.

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

[35] Zhang, C.-H. and Zhang, S.S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940

Received March 2017 and revised July 2017