HAL open science

High-dimensional regression with unknown variance

Christophe Giraud, Sylvie Huet, Nicolas Verzelen

▶ **To cite this version:**

Christophe Giraud, Sylvie Huet, Nicolas Verzelen. High-dimensional regression with unknown variance. 2012. hal-00626630v2

HAL Id: hal-00626630

https://hal.science/hal-00626630v2

Preprint submitted on 17 Feb 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Submitted to the Statistical Science

High-dimensional regression with unknown variance

Christophe Giraud, Sylvie Huet and Nicolas Verzelen

École Polytechnique and Institut National de Recherche en Agronomie

Abstract. We review recent results for high-dimensional sparse linear regression in the practical case of unknown variance. Different sparsity settings are covered, including coordinate-sparsity, group-sparsity and variation-sparsity. The emphasis is put on non-asymptotic analyses and feasible procedures. In addition, a small numerical study compares the practical performance of three schemes for tuning the Lasso estimator and some references are collected for some more general models, including multivariate regression and nonparametric regression.

AMS 2000 subject classifications: 62J05, 62J07, 62G08, 62H12.
Key words and phrases: linear regression, high-dimension, unknown variance.

1. INTRODUCTION

In the present paper, we mainly focus on the linear regression model

$$Y = \mathbf{X}\beta_0 + \varepsilon\,, \tag{1}$$

where Y is a n-dimensional response vector, $\mathbf{X}$ is a fixed $n \times p$ design matrix, and the vector ε is made of n i.i.d Gaussian random variables with $\mathcal{N}(0, \sigma^2)$ distribution. In the sequel, $\mathbf{X}^{(i)}$ stands for the i-th row of $\mathbf{X}$. Our interest is on the high-dimensional setting, where the dimension p of the unknown parameter β_0 is large, possibly larger than n.

The analysis of the high-dimensional linear regression model has attracted a lot of attention in the last decade. Nevertheless, there is a longstanding gap between the theory where the variance σ^2 is generally assumed to be known and the practice where it is often unknown. The present paper is mainly devoted to review recent results on linear regression in high-dimensional settings with *unknown* variance σ^2. A few additional results for multivariate regression and the nonparametric regression model

$$Y_i = f(\mathbf{X}^{(i)}) + \varepsilon_i, \qquad i = 1, \ldots, n\,, \tag{2}$$

will also be mentioned.

CMAP, UMR CNRS 7641, Ecole Polytechnique, Route de Saclay, 91128 Palaiseau Cedex, FRANCE. (e-mail: christophe.giraud@polytechnique.edu) UR341 MIA, INRA, F-78350 Jouy-en-Josas, FRANCE (e-mail: sylvie.huet@jouy.inra.fr) UMR729 MISTEA, INRA, F-34060 Montpellier, FRANCE Montpellier (e-mail: nicolas.verzelen@supagro.inra.fr)

1.1 Sparsity assumptions

In a high-dimensional linear regression model, accurate estimation is unfeasible unless it relies on some special properties of the parameter β_0. The most common assumption on β_0 is that it is sparse in some sense. We will consider in this paper the three following classical sparsity assumptions.

Coordinate-sparsity. Most of the coordinates of β_0 are assumed to be zero (or approximately zero). This is the most common acceptation for sparsity in linear regression.

Structured-sparsity. The pattern of zero(s) of the coordinates of β_0 is assumed to have an a priori known structure. For instance, in group-sparsity [80], the covariates are clustered into M groups and when the coefficient $\beta_{0,i}$ corresponding to the covariate $\mathbf{X}_i$ (the i-th column of $\mathbf{X}$) is non-zero, then it is likely that all the coefficients $\beta_{0,j}$ with variables $\mathbf{X}_j$ in the same cluster as $\mathbf{X}_i$ are non-zero.

Variation-sparsity. The $p-1$-dimensional vector β_0^V of variation of β_0 is defined by $\beta_{0,j}^V = \beta_{0,j+1} - \beta_{0,j}$. Sparsity in variation means that most of the components of β_0^V are equal to zero (or approximately zero). When $p = n$ and $\mathbf{X} = I_n$, variation-sparse linear regression corresponds to signal segmentation.

1.2 Statistical objectives

In the linear regression model, there are roughly two kinds of estimation objectives. In the *prediction problem*, the goal is to estimate $\mathbf{X}\beta_0$, whereas in the *inverse problem* it is to estimate β_0. When the vector β_0 is sparse, a related objective is to estimate the *support* of β_0 (model identification problem) which is the set of the indices j corresponding to the non zero coefficients $\beta_{0,j}$. Inverse problems and prediction problems are not equivalent in general. When the Gram matrix $\mathbf{X}\mathbf{X}^*$ is poorly conditioned, the former problems can be much more difficult than the latter. Since there are only a few results on inverse problems with unknown variance, we will focus on the prediction problem, the support estimation problem being shortly discussed in the course of the paper.

In the sequel, $\mathbb{E}_{\beta_0}[.]$ stands for the expectation with respect to $Y \sim \mathcal{N}(\mathbf{X}\beta_0, \sigma^2 I_n)$ and $\|.\|_2$ is the euclidean norm. The prediction objective amounts to build estimators $\widehat{\beta}$ so that the risk

$$\mathcal{R}[\widehat{\beta}; \beta_0] := \mathbb{E}_{\beta_0}[\|\mathbf{X}(\widehat{\beta} - \beta_0)\|_2^2] \tag{3}$$

is as small as possible.

1.3 Approaches

Most procedures that handle high-dimensional linear models [22, 26, 62, 72, 73, 81, 83, 85] rely on tuning parameters whose optimal value depends on σ. For example, the results of Bickel et al. [17] suggest to choose the tuning parameter λ of the Lasso of the order of $2\sigma\sqrt{2\log(p)}$. As a consequence, all these procedures cannot be directly applied when σ^2 is unknown.

A straightforward approach is to replace σ^2 by an estimate of the variance in the optimal value of the tuning parameter(s). Nevertheless, the variance σ^2 is difficult to estimate in high-dimensional settings, so a plug-in of the variance does not necessarily yield good results. There are basically two approaches to build on this amount of work on high-dimensional estimation with known variance.

1. **Ad-hoc estimation.** There has been some recent work [16, 68, 71] to modify procedures like the Lasso in such a way that the tuning parameter does not depend anymore on σ^2 (see Section 4.2). The challenge is to find a smart modification of the procedure, so that the resulting estimator $\widehat{\beta}$ is computationally feasible and has a risk $\mathcal{R}\left[\widehat{\beta};\beta_0\right]$ as small as possible.
2. **Estimator selection.** Given a collection $(\widehat{\beta}_\lambda)_{\lambda\in\Lambda}$ of estimators, the objective of estimator selection is to pick an index $\widehat{\lambda}$ such that the risk of $\widehat{\beta}_{\widehat{\lambda}}$ is as small as possible; ideally as small as the risk $\mathcal{R}[\widehat{\beta}_{\lambda^*};\beta_0]$ of the so-called *oracle* estimator
$$\widehat{\beta}_{\lambda^*} := \underset{\{\widehat{\beta}_\lambda,\ \lambda\in\Lambda\}}{\operatorname{argmin}}\ \mathcal{R}\left[\widehat{\beta}_\lambda;\beta_0\right]\ . \tag{4}$$
Efficient estimator selection procedures can then be applied to tune the aforementioned estimation methods [22, 26, 62, 72, 73, 81, 83, 85]. Among the most famous methods for estimator selection, we mention V-fold cross-validation (Geisser [32]), AIC (Akaike [1]) and BIC (Schwarz [64]) criteria.

The objective of this survey is to describe state-of-the-art procedures for high-dimensional linear regression with unknown variance. We will review both automatic tuning methods and ad-hoc methods. There are some procedures that we will let aside. For example, Baraud [11] provides a versatile estimator selection scheme, but the procedure is computationally intractable in large dimensions. Linear or convex aggregation of estimators are also valuable alternatives to estimator selection when the goal is to perform *estimation*, but only a few theoretical works have addressed the aggregation problem when the variance is unknown [35, 33]. For these reasons, we will not review these approaches in the sequel.

1.4 Why care about non-asymptotic analyses ?

AIC [1], BIC [64] and V-fold Cross-Validation [32] are probably the most popular criteria for estimator selection. The use of these criteria relies on some classical asymptotic optimality results. These results focus on the setting where the collection of estimators $(\widehat{\beta}_\lambda)_{\lambda\in\Lambda}$ and the dimension p are fixed and consider the limit behavior of the criteria when the sample size n goes to infinity. For example, under some suitable conditions, Shibata [67], Li [53] and Shao [66] prove that the risk of the estimator selected by AIC or V-fold CV (with $V = V_n \to \infty$) is asymptotically equivalent to the oracle risk $\mathcal{R}[\widehat{\beta}_{\lambda^*};\beta_0]$. Similarly, Nishii [59] shows that the BIC criterion is consistent for model selection.

All these asymptotic results can lead to misleading conclusions in modern statistical settings where the sample size remains small and the parameter's dimension becomes large. For instance it is proved in [12, Sect.3.3.2] and illustrated in [12, Sect.6.2] that BIC (and thus AIC) can strongly overfit and should not be used for p larger than n. Additional examples are provided in Appendix A. A non-asymptotic analysis takes into account all the characteristics of the selection problem (sample size n, parameter dimension p, number of models per dimension, etc). It treats n and p as they are and it avoids to miss important features hidden in asymptotic limits. For these reasons, we will restrict in this review on non-asymptotic results.

1.5 Organization of the paper

In Section 2, we investigate how the ignorance of the variance affects the minimax risk bounds. In Section 3, some "generic" estimators selection schemes are presented. The coordinate-sparse setting is addressed Section 4 : some theoretical results are collected and a small numerical experiment compares different Lasso-based procedures. The group-sparse and variation-sparse settings are reviewed in Section 5 and 6, and Section 7 is devoted to some more general models such as multivariate regression or nonparametric regression.

In the sequel, C, C_1,... refer to numerical constants whose value may vary from line to line, while $\|\beta\|_0$ stands for the number of non zero components of β and $|\mathcal{J}|$ for the cardinality of a set $\mathcal{J}$.

2. THEORETICAL LIMITS

The goal of this section is to address the intrinsic difficulty of a coordinate-sparse linear regression problem. We will answer the following questions: Which range of p can we reasonably consider? When the variance is unknown, can we hope to do as well as when the variance is known?

2.1 Minimax adaptation

A classical way to assess the performance of an estimator $\widehat{\beta}$ is to measure its maximal risk over a class $\boldsymbol{B} \subset \mathbb{R}^p$. This is the minimax point of view. As we are interested in coordinate-sparsity for β_0, we will consider the sets $\boldsymbol{B}[k,p]$ of vectors that contain at most k non zero coordinates for some $k > 0$.

Given an estimator $\widehat{\beta}$, the *maximal prediction risk* of $\widehat{\beta}$ over $\boldsymbol{B}[k,p]$ for a fixed design $\mathbf{X}$ and a variance σ^2 is defined by $\sup_{\beta_0 \in \boldsymbol{B}[k,p]} \mathcal{R}[\widehat{\beta};\beta_0]$ where the risk function $\mathcal{R}[.,\beta_0]$ is defined by (3). Taking the infimum of the maximal risk over all possible estimators $\widehat{\beta}$, we obtain the *minimax risk*

$$\mathbf{R}[k,\mathbf{X}] = \inf_{\widehat{\beta}} \sup_{\beta_0 \in \boldsymbol{B}[k,p]} \mathcal{R}[\widehat{\beta};\beta_0] \ . \tag{5}$$

Minimax bounds are convenient results to assess the range of problems that are statistically feasible and the optimality of particular procedures. Below, we say that an estimator $\widehat{\beta}$ is "minimax" over $\boldsymbol{B}[k,p]$ if its maximal prediction risk is close to the minimax risk.

In practice, the number of non-zero coordinates of β_0 is unknown. The fact that an estimator $\widehat{\beta}$ is minimax over $\boldsymbol{B}[k,p]$ for some specific $k > 0$ does not imply that $\widehat{\beta}$ estimates well vectors β_0 that are less sparse. A good estimation procedure $\widehat{\beta}$ should not require the knowledge of the sparsity k of β_0 and should perform as well as if this sparsity were known. An estimator $\widehat{\beta}$ that nearly achieves the minimax risk over $\boldsymbol{B}[k,p]$ for a range of k is said to be *adaptive* to the sparsity. Similarly, an estimator $\widehat{\beta}$ 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 the following subsections, we review sharp bounds on the minimax prediction risks for both known and unknown sparsity, known and unknown variance. The big picture is summed up in Figure 1. Roughly, it says that adaptation is possible as long as $2k\log(p/k) < n$. In contrast, the situation becomes more complex for

the ultra-high-dimensional[1] setting where $2k\log(p/k) \geq n$. The rest of this section is devoted to explain this big picture.

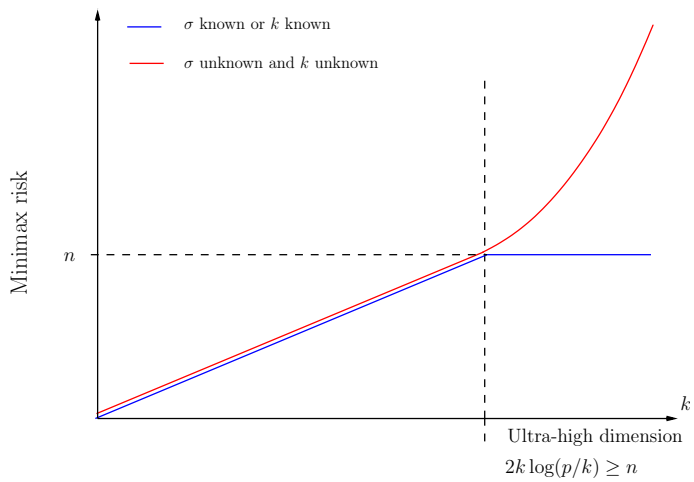

FIGURE 1. *Minimal prediction risk over $\boldsymbol{B}[k,p]$ as a function of k.*

2.2 Minimax risks under known sparsity and known variance

The minimax risk $\mathbf{R}[k,\mathbf{X}]$ depends on the form of the design $\mathbf{X}$. In order to grasp this dependency, we define for any $k > 0$, the largest and the smallest sparse eigenvalues of order k of $\mathbf{X}^*\mathbf{X}$ by

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

.

PROPOSITION 2.1. *Assume that k and σ are known. There exist positive numerical constants C_1, C_1', C_2, and C_2' such that the following holds. For any (k,n,p) such that $k \leq n/2$ and any design $\mathbf{X}$, we have*

$$C_1 \frac{\Phi_{2k,-}(\mathbf{X})}{\Phi_{2k,+}(\mathbf{X})} k\log\left(\frac{p}{k}\right)\sigma^2 \leq \mathbf{R}[k,\mathbf{X}] \leq C_1'\left[k\log\left(\frac{p}{k}\right)\wedge n\right]\sigma^2 \ , \tag{6}$$

For any (k,n,p) such that $k \leq n/2$, we have

$$C_2\left[k\log\left(\frac{p}{k}\right)\wedge n\right]\sigma^2 \leq \sup_{\mathbf{X}} \mathbf{R}[k,\mathbf{X}] \leq C_2'\left[k\log\left(\frac{p}{k}\right)\wedge n\right]\sigma^2 \ . \tag{7}$$

The minimax lower bound (6) has been first proved in [61, 62, 79] while (7) is stated in [77]. Let us first comment the bound (7). If the vector β_0 has k-non zero components and if these components are *a priori* known, then one may build estimators that achieve a risk bound of the order k. In a (non-ultra) high-dimensional

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

setting ($2k\log(p/k) \le n$), the minimax risk is of the order $k\log(p/k)\sigma^2$. The logarithmic term is the price to pay to cope with the fact that we do not know the position of the non zero components in β_0. The situation is quite different in an ultra-high-dimensional setting ($2k\log(p/k) > n$). Indeed, the minimax risk remains of the order of $n\sigma^2$, which corresponds to the minimax risk of estimation of the vector $\mathbf{X}\beta_0$ without any sparsity assumption (see the blue curve in Figure 1). In other terms, the sparsity index k does not play a role anymore.

Dependency of $\mathbf{R}[k,\mathbf{X}]$ on the design $\mathbf{X}$. It follows from (6) that $\sup_{\mathbf{X}} \mathbf{R}[k,\mathbf{X}]$ is nearly achieved by designs $\mathbf{X}$ satisfying $\Phi_{2k,-}(\mathbf{X})/\Phi_{2k,+}(\mathbf{X}) \approx 1$, when the setting is not ultra-high dimensional. For some designs such that $\Phi_{2k,-}(\mathbf{X})/\Phi_{2k,+}(\mathbf{X})$ is small, the minimax prediction risk $\mathbf{R}[k,\mathbf{X}]$ is possibly faster (see [77] for a discussion). In a ultra-high dimensional, the form of the minimax risk ($n\sigma^2$) is related to the fact that no designs can satisfy $\Phi_{2k,-}(\mathbf{X})/\Phi_{2k,+}(\mathbf{X}) \approx 1$ (see e.g. [10]). The lower bound $\mathbf{R}[k,\mathbf{X}] \ge C\,[k\log(p/k)\wedge n]\,\sigma^2$ in (7) is for instance achieved by realizations of a standard Gaussian design, that is designs $\mathbf{X}$ whose components follow independent standard normal distributions. See [77] for more details.

2.3 Adaptation to the sparsity and to the variance

Adaptation to the sparsity when the variance is known. When σ^2 is known, there exist both model selection and aggregation procedures that achieve this $[k\log(p/k)\wedge n]\sigma^2$ risk simultaneously for all k and for all designs $\mathbf{X}$. Such procedures derive from the work of Birgé and Massart [18] and Leung and Barron [52]. However, these methods are intractable for large p except for specific forms of the design. We refer to Appendix B.1 for more details.

Simultaneous adaptation to the sparsity and the variance. We first restrict to the non-ultra high-dimensional setting, where the number of non-zero components k is unknown but satisfies $2k\log(p/k) < n$. In this setting, some procedures based on penalized log-likelihood [12] are simultaneous adaptive to the unknown sparsity and to the unknown variance and this for all designs $\mathbf{X}$. Again such procedures are intractable for large p. See Appendix B.2 for more details. If we want to cover all k (including ultra-high dimensional settings), the situation is different as shown in the next proposition (from [77]).

PROPOSITION 2.2 (Simultaneous adaptation is impossible). *There exist positive constants C, C', C_1, C_2, C_3, C_1', C_2', and C_3', such that the following holds. Consider any $p \ge n \ge C$ and $k \le p^{1/3}\wedge n/2$ such that $k\log(p/k) \ge C'n$. There exist designs $\mathbf{X}$ of size $n\times p$ such that for any estimator $\widehat{\beta}$, we have either*

$$\sup_{\sigma^2>0} \frac{\mathcal{R}[\widehat{\beta};0_p]}{\sigma^2} > C_1 n\ , \qquad \textit{or}$$

$$\sup_{\beta_0\in\boldsymbol{B}[k,p]\ ,\ \sigma^2>0} \frac{\mathcal{R}[\widehat{\beta};\beta_0]}{\sigma^2} > C_2 k\log\left(\frac{p}{k}\right)\exp\left[C_3\frac{k}{n}\log\left(\frac{p}{k}\right)\right]\ .$$

Conversely, there exist two estimators $\widehat{\beta}^{(n)}$ and $\widehat{\beta}^{BGH}$ (defined in Appendix B.2) that respectively satisfy

$$\sup_{\mathbf{X}} \sup_{\beta_0 \in \mathbb{R}^p,\ \sigma^2>0} \frac{\mathcal{R}[\widehat{\beta}^{(n)};\beta_0]}{\sigma^2} \leq C_1' n \ ,$$

$$\sup_{\mathbf{X}} \sup_{\beta_0 \in \boldsymbol{B}[k,p]\ ,\ \sigma^2>0} \frac{\mathcal{R}[\widehat{\beta}^{BGH};\beta_0]}{\sigma^2} \leq C_2' k \log\left(\frac{p}{k}\right) \exp\left[C_3' \frac{k}{n} \log\left(\frac{p}{k}\right)\right] \ ,$$

for all $1 \leq k \leq [(n-1) \wedge p]/4$.

As a consequence, simultaneous adaptation to the sparsity and to the variance is impossible in an ultra-high dimensional setting. Indeed, any estimator $\widehat{\beta}$ that does not rely on σ^2 has to pay at least one of these two prices:

1. The estimator $\widehat{\beta}$ does not use the sparsity of the true parameter β_0 and its risk for estimating $\mathbf{X}0_p$ is of the same order as the minimax risk over $\mathbb{R}^n$.
2. For any $1 \leq k \leq p^{1/3}$, the risk of $\widehat{\beta}$ fulfills

$$\sup_{\sigma>0} \sup_{\beta_0 \in \boldsymbol{B}[k,p]} \frac{\mathcal{R}[\widehat{\beta};\beta_0]}{\sigma^2} \geq C_1 k \log(p) \exp\left[C_2 \frac{k}{n} \log(p)\right] \ .$$

It follows that the maximal risk of $\widehat{\beta}$ is blowing up in an ultra-high-dimensional setting (red curve in Figure 1), while the minimax risk is stuck to n (blue curve in Figure 1). The designs that satisfy the minimax lower bounds of Proposition 2.2 include realizations of a standard Gaussian design.

In an ultra-high dimensional setting, the prediction problem becomes extremely difficult under unknown variance because the variance estimation itself is inconsistent as shown in the next proposition (from [77]).

PROPOSITION 2.3. *There exist positive constants C, C_1, and C_2 such that the following holds. Assume that $p \geq n \geq C$. For any $1 \leq k \leq p^{1/3}$, there exist designs $\mathbf{X}$ such that*

$$\inf_{\widehat{\sigma}} \sup_{\sigma>0,\ \beta_0 \in \mathbf{B}[k,p]} \mathbb{E}_{\beta_0}\left[\left|\frac{\sigma^2}{\widehat{\sigma}^2} - \frac{\widehat{\sigma}^2}{\sigma^2}\right|\right] \geq C_1 \frac{k}{n} \log\left(\frac{p}{k}\right) \exp\left[C_2 \frac{k}{n} \log\left(\frac{p}{k}\right)\right] \ .$$

2.4 What should we expect from a good estimation procedure?

Let us consider an estimator $\widehat{\beta}$ that does not depend on σ^2. Relying on the previous minimax bounds, we will say that $\widehat{\beta}$ achieves an *optimal* risk bound (with respect to the sparsity) if

$$\mathcal{R}[\widehat{\beta};\beta_0] \leq C_1 \|\beta_0\|_0 \log(p) \sigma^2 \ , \tag{8}$$

for any $\sigma > 0$ and any vector $\beta_0 \in \mathbb{R}^p$ such that $1 \leq \|\beta_0\|_0 \log(p) \leq C_2 n$. Such risk bounds prove that the estimator is approximately (up to a possible $\log(\|\beta_0\|_0)$ additional term) minimax adaptive to the unknown variance and the unknown sparsity. The condition $\|\beta_0\|_0 \log(p) \leq C_2 n$ ensures that the setting is not ultra-high-dimensional. As stated above, some procedures achieve (8) for all designs $\mathbf{X}$ but they are intractable for large p (see Appendix B). One purpose of this review

is to present fast procedures that achieve this kind of bounds under possible restrictive assumptions on the design matrix $\mathbf{X}$.

For some procedures, (8) can be improved into a bound of the form

$$\mathcal{R}[\widehat{\beta};\beta_0] \leq C_1 \inf_{\beta\neq 0}\left\{\|\mathbf{X}(\beta-\beta_0)\|_2^2 + \|\beta\|_0 \log(p)\sigma^2\right\} \ , \tag{9}$$

with C_1 close to one. Again, the dimension $\|\beta_0\|_0$ is restricted to be smaller than $Cn/\log(p)$ to ensure that the setting is not ultra-high dimensional. This kind of bound makes a clear trade-off between a bias and a variance term. For instance, when β_0 contains many components that are nearly equal to zero, the bound (9) can be much smaller than (8).

2.5 Other statistical problems in an ultra-high-dimensional setting

We have seen that adaptation becomes impossible for the prediction problem in a ultra-high dimensional setting. For other statistical problems, including the prediction problem with random design, the inverse problem (estimation of β_0), the variable selection problem (estimation of the support of β_0), the dimension reduction problem [77, 78, 46], the minimax risks are blowing up in a ultra-high dimensional setting. This kind of phase transition has been observed in a wide range of random geometry problems [29], suggesting some universality in this limitation. In practice, the sparsity index k is not known, but given (n,p) we can compute $k^* := \max\{k : \ 2k\log(p/k) \geq n\}$. One may interpret that the problem is still reasonably difficult as long as $k \leq k^*$. This gives a simple rule of thumb to know what we can hope from a given regression problem. For example, setting $p = 5000$ and $n = 50$ leads to $k^* = 3$, implying that the prediction problem becomes extremely difficult when there are more than 4 relevant covariates (see the simulations in [77]).

3. SOME GENERIC SELECTION SCHEMES

Among the selection schemes not requiring the knowledge of the variance σ^2, some are very specific to a particular algorithm, while some others are more generic. We describe in this section three versatile selection principles and refer to the examples for the more specific schemes.

3.1 Cross-Validation procedures

The cross-validation schemes are nearly universal in the sense that they can be implemented in most statistical frameworks and for most estimation procedures. The principle of the cross-validation schemes is to split the data into a *training* set and a *validation* set : the estimators are built on the *training* set and the *validation* set is used for estimating their prediction risk. This training / validation splitting is eventually repeated several times. The most popular cross-validation schemes are :

- *Hold-out* [57, 27] which is based on a single split of the data for *training* and *validation*.
- *V-fold* CV [32]. The data is split into V subsamples. Each subsample is successively removed for *validation*, the remaining data being used for *training*.
- *Leave-one-out* [69] which corresponds to n-fold CV.

- *Leave-q-out* (also called *delete-q-CV*) [65] where every possible subset of cardinality q of the data is removed for *validation*, the remaining data being used for *training*.

We refer to Arlot and Célisse [6] for a review of the cross-validation schemes and their theoretical properties.

3.2 Penalized empirical loss

Penalized empirical loss criteria form another class of versatile selection schemes, yet less universal than CV procedures. The principle is to select among a family $(\widehat{\beta}_\lambda)_{\lambda\in\Lambda}$ of estimators by minimizing a criterion of the generic form

$$\mathrm{Crit}(\lambda) = \mathcal{L}_{\mathbf{X}}(Y, \widehat{\beta}_\lambda) + \mathrm{pen}(\lambda), \tag{10}$$

where $\mathcal{L}_{\mathbf{X}}(Y, \widehat{\beta}_\lambda)$ is a measure of the distance between Y and $\mathbf{X}\widehat{\beta}_\lambda$, and pen is a function from Λ to $\mathbb{R}^+$. The penalty function sometimes depends on data.

Penalized log-likelihood. The most famous criteria of the form (10) are AIC and BIC. They have been designed to select among estimators $\widehat{\beta}_\lambda$ obtained by maximizing the likelihood of (β, σ) with the constraint that β lies on a linear space S_λ (called *model*). In the Gaussian case, these estimators are given by $\mathbf{X}\widehat{\beta}_\lambda = \Pi_{S_\lambda} Y$, where Π_{S_λ} denotes the orthogonal projector onto the model S_λ. For AIC and BIC, the function $\mathcal{L}_{\mathbf{X}}$ corresponds to twice the negative log-likelihood $\mathcal{L}_{\mathbf{X}}(Y, \widehat{\beta}_\lambda) = n \log(\|Y - \mathbf{X}\widehat{\beta}_\lambda\|_2^2)$ and the penalties are $\mathrm{pen}(\lambda) = 2\dim(S_\lambda)$ and $\mathrm{pen}(\lambda) = \dim(S_\lambda)\log(n)$ respectively. We recall that these two criteria can perform very poorly in a high-dimensional setting.

In the same setting, Baraud *et al.* [12] propose alternative penalties built from a non-asymptotic perspective. The resulting criterion can handle the high-dimensional setting where p is possibly larger than n and the risk of the selection procedure is controlled by a bound of the form (9), see Theorem 2 in [12].

Plug-in criteria. Many other penalized-empirical-loss criteria have been developed in the last decades. Several selection criteria [14, 18] have been designed from a non-asymptotic point of view to handle the case where the variance is known. These criteria usually involve the residual least-square $\mathcal{L}_{\mathbf{X}}(Y, \widehat{\beta}_\lambda) = \|Y - \mathbf{X}\widehat{\beta}_\lambda\|_2^2$ and a penalty $\mathrm{pen}(\lambda)$ depending on the variance σ^2. A common practice is then to plug in the penalty an estimate $\hat{\sigma}^2$ of the variance in place of the variance. For linear regression, when the design matrix $\mathbf{X}$ has a rank less than n, a classical choice for $\hat{\sigma}^2$ is

$$\hat{\sigma}^2 = \frac{\|Y - \Pi_{\mathbf{X}} Y\|_2^2}{n - \mathrm{rank}(\mathbf{X})},$$

with $\Pi_{\mathbf{X}}$ the orthogonal projector onto the range of $\mathbf{X}$. This estimator $\hat{\sigma}^2$ has the nice feature to be independent of $\Pi_{\mathbf{X}} Y$ on which usually rely the estimators $\widehat{\beta}_\lambda$. Nevertheless, the variance of $\hat{\sigma}^2$ is of order $\sigma^4/(n - \mathrm{rank}(\mathbf{X}))$ which is small only when the sample size n is quite large in front of the rank of $\mathbf{X}$. This situation is unfortunately not likely to happen in a high-dimensional setting where p can be larger than n.

3.3 Approximation versus complexity penalization : LinSelect

The criterion proposed by Baraud *et al.* [12] can handle high-dimensional settings but it suffers from two rigidities. First, it can only handle *fixed* collections

of models $(S_\lambda)_{\lambda\in\Lambda}$. In some situations, the size of Λ is huge (e.g. for complete variable selection) and the estimation procedure can then be computationally intractable. In this case, we may want to work with a subcollection of models $(S_\lambda)_{\lambda\in\widehat{\Lambda}}$, where $\widehat{\Lambda}\subset\Lambda$ may depend on data. For example, for complete variable selection, the subset $\widehat{\Lambda}$ could be generated by efficient algorithms like LARS [30]. The second rigidity of the procedure of Baraud *et al.* [12] is that it can only handle constrained-maximum-likelihood estimators. This procedure then does not help for selecting among arbitrary estimators such as the Lasso or Elastic-Net.

These two rigidities have been addressed recently by Baraud *et al.* [13]. They propose a selection procedure, LinSelect, which can handle both data-dependent collections of models and arbitrary estimators $\widehat{\beta}_\lambda$. The procedure is based on a collection $\mathbb{S}$ of linear spaces which gives a collection of possible "approximative" supports for the estimators $(\mathbf{X}\widehat{\beta}_\lambda)_{\lambda\in\Lambda}$. A measure of complexity on $\mathbb{S}$ is provided by a weight function $\Delta:\mathbb{S}\to\mathbb{R}^+$. We refer to Sections 4.1 and 5 for examples of collection $\mathbb{S}$ and weight Δ in the context of coordinate-sparse and group-sparse regression. We present below a simplified version of the LinSelect procedure. For a suitable, possibly data-dependent, subset $\widehat{\mathbb{S}}\subset\mathbb{S}$ (depending on the statistical problem), the estimator $\widehat{\beta}_{\widehat{\lambda}}$ is selected by minimizing the criterion

$$\mathrm{Crit}(\widehat{\beta}_\lambda)=\inf_{S\in\widehat{\mathbb{S}}}\left[\|Y-\Pi_S\mathbf{X}\widehat{\beta}_\lambda\|_2^2+\frac{1}{2}\|\mathbf{X}\widehat{\beta}_\lambda-\Pi_S\mathbf{X}\widehat{\beta}_\lambda\|_2^2+\mathrm{pen}(S)\,\widehat{\sigma}_S^2\right], \tag{11}$$

where Π_S is the orthogonal projector onto S,

$$\widehat{\sigma}_S^2=\frac{\|Y-\Pi_S Y\|_2^2}{n-\dim(S)},$$

and $\mathrm{pen}(S)$ is a penalty depending on Δ. In the cases we will consider here, the penalty $\mathrm{pen}(S)$ is roughly of the order of $\Delta(S)$ and therefore it penalizes S according to its complexity. We refer to the Appendix C for a precise definition of this penalty and more details on its characteristics. We emphasize that the Criterion (11) and the family of estimators $\{\widehat{\beta}_\lambda,\ \lambda\in\Lambda\}$ are based on the *same* data Y and $\mathbf{X}$. In other words, there is no data-splitting occurring in the LinSelect procedure. The first term in (11) quantifies the fit of the projected estimator to the data, the second term evaluates the approximation quality of the space S and the last term penalizes S according to its complexity. We refer to Proposition C.1 in Appendix C and Theorem 1 in [12] for risk bounds on the selected estimator. Instantiations of the procedure and more specific risks bounds are given in Sections 4 and 5 in the context of coordinate-sparsity and group-sparsity.

From a computational point of view, the algorithmic complexity of LinSelect is at most proportional to $|\Lambda|\times|\widehat{\mathbb{S}}|$ and in many cases there is no need to scan the whole set $\widehat{\mathbb{S}}$ for each $\lambda\in\Lambda$ to minimize (11). In the examples of Sections 4 and 5, the whole procedure is computationally less intensive than V-fold CV, see Table 3. Finally, we mention that for the constrained least-square estimators $\mathbf{X}\widehat{\beta}_\lambda=\Pi_{S_\lambda}Y$, the LinSelect procedure with $\widehat{\mathbb{S}}=\{S_\lambda:\lambda\in\Lambda\}$ simply coincides with the procedure of Baraud *et al.* [12].

4. COORDINATE-SPARSITY

In this section, we focus on the high-dimensional linear regression model $Y=\mathbf{X}\beta_0+\varepsilon$ where the vector β_0 itself is assumed to be sparse. This setting has

attracted a lot of attention in the last decade, and many estimation procedures have been developed. Most of them require the choice of tuning parameters which depend on the unknown variance σ^2. This is for instance the case for the Lasso [72, 24], Dantzig Selector [22], Elastic Net [85], MC+ [81], aggregation techniques [21, 26], etc.

We first discuss how the generic schemes introduced in the previous section can be instantiated for tuning these procedures and for selecting among them. Then, we pay a special attention to the calibration of the Lasso. Finally, we discuss the problem of support estimation and present a small numerical study.

4.1 Automatic tuning methods

Cross-validation. Arguably, V-fold Cross-Validation is the most popular technique for tuning the above-mentioned procedures. To our knowledge, there are no other theoretical results for V-fold CV in large dimensional settings.

In practice, V-fold CV seems to give rather good results. The problem of choosing the best V has not yet been solved [6, Section 10], but it is often reported that a good choice for V is between 5 and 10. Indeed, the statistical performance does not increase for larger values of V, and averaging over 10 splits remains computationally feasible [41, Section 7.10].

LinSelect. The procedure LinSelect can be used for selecting among a collection $(\widehat{\beta}_\lambda)_{\lambda\in\Lambda}$ of sparse regressors as follows. For $\mathcal{J} \subset \{1,\dots,p\}$, we define $\mathbf{X}_\mathcal{J}$ as the matrix $[\mathbf{X}_{ij}]_{i=1,\dots,n,\ j\in\mathcal{J}}$ obtained by only keeping the columns of $\mathbf{X}$ with index in $\mathcal{J}$. We recall that the collection $\mathbb{S}$ gives some possible "approximative" supports for the estimators $(\mathbf{X}\widehat{\beta}_\lambda)_{\lambda\in\Lambda}$. For sparse linear regression, a possible collection $\mathbb{S}$ and measure of complexity Δ are

$$\begin{aligned} \mathbb{S} &= \left\{ S = \mathrm{range}(\mathbf{X}_\mathcal{J}),\ \mathcal{J} \subset \{1,\dots,p\}\,,\ 1 \le |\mathcal{J}| \le n/(3\log p) \right\} \\ \text{and} \quad \Delta(S) &= \log\binom{p}{\dim(S)} + \log(\dim(S)). \end{aligned}$$

Let us introduce the spaces $\widehat{S}_\lambda = \mathrm{range}\left(\mathbf{X}_{\mathrm{supp}(\widehat{\beta}_\lambda)}\right)$ and the subcollection of $\mathbb{S}$

$$\widehat{\mathbb{S}} = \left\{ \widehat{S}_\lambda,\ \lambda \in \widehat{\Lambda} \right\}, \quad \text{where } \widehat{\Lambda} = \left\{ \lambda \in \Lambda \ :\ \widehat{S}_\lambda \in \mathbb{S} \right\}.$$

The following proposition gives a risk bound when selecting $\widehat{\lambda}$ with LinSelect with the above choice of $\widehat{\mathbb{S}}$ and Δ.

PROPOSITION 4.1. *There exists a numerical constant $C > 1$ such that for any minimizer $\widehat{\lambda}$ of the Criterion (11), we have*

$$\begin{aligned} &\mathcal{R}\left[\widehat{\beta}_{\widehat{\lambda}};\beta_0\right] \\ &\le\ C\ \mathbb{E}\left[\inf_{\lambda\in\Lambda}\left\{\|\mathbf{X}\widehat{\beta}_\lambda - \mathbf{X}\beta_0\|_2^2 + \inf_{S\in\widehat{\mathbb{S}}}\left\{\|\mathbf{X}\widehat{\beta}_\lambda - \Pi_S\mathbf{X}\widehat{\beta}_\lambda\|_2^2 + \dim(S)\log(p)\sigma^2\right\}\right\}\right] \\ &\le\ C\ \mathbb{E}\left[\inf_{\lambda\in\widehat{\Lambda}}\left\{\|\mathbf{X}\widehat{\beta}_\lambda - \mathbf{X}\beta_0\|_2^2 + \|\widehat{\beta}_\lambda\|_0\log(p)\sigma^2\right\}\right]. \end{aligned} \tag{12}$$

Proposition 4.1 is a simple corollary of Proposition C.1 in Appendix C. The first bound involves three terms: the loss of the estimator $\widehat{\beta}_\lambda$, an approximation loss, and a variance term. Hence, LinSelect chooses an estimator $\widehat{\beta}_\lambda$ that achieves a trade-off between the loss of $\widehat{\beta}_\lambda$ and the closeness of $\mathbf{X}\widehat{\beta}_\lambda$ to some small dimensional subspace S. The bound (12) cannot be formulated in the form (9) due to the random nature of the set $\widehat{\Lambda}$. Nevertheless, a bound similar to (8) can be deduced from (12) when the estimators $\widehat{\beta}_\lambda$ are least-squares estimators, see Corollary 4 in [13]. Furthermore, we note that increasing the size of Λ leads to a better risk bound for $\widehat{\beta}_{\widehat{\lambda}}$. It is then advisable to consider a family of candidate estimators $\{\widehat{\beta}_\lambda,\ \lambda \in \Lambda\}$ as large as possible. The Proposition 4.1 is valid for any family of estimators $\{\widehat{\beta}_\lambda,\ \lambda \in \Lambda\}$, for the specific family of Lasso estimators $\{\widehat{\beta}^L_\lambda,\ \lambda > 0\}$ we provide a refined bound in Proposition 4.3, Section 4.3.

4.2 Lasso-type estimation under unknown variance

The Lasso is certainly one of the most popular methods for variable selection in a high-dimensional setting. Given $\lambda > 0$, the Lasso estimator $\widehat{\beta}^L_\lambda$ is defined by $\widehat{\beta}^L_\lambda := \operatorname{argmin}_{\beta\in\mathbb{R}^p} \|Y-\mathbf{X}\beta\|_2^2+\lambda\|\beta\|_1$. A sensible choice of λ must be homogeneous with the square-root of the variance σ^2. As explained above, when the variance σ^2 is unknown, one may apply V-fold CV or LinSelect to select λ. Some alternative approaches have also been developed for tuning the Lasso. Their common idea is to modify the ℓ_1 criterion so that the tuning parameter becomes pivotal with respect to σ^2. This means that the method remains valid for any $\sigma > 0$ and that the choice of the tuning parameter does not depend on σ. For the sake of simplicity, we assume throughout this subsection and the next one that the columns of $\mathbf{X}$ are normalized to one.

ℓ_1-penalized log-likelihood. In low-dimensional regression, it is classical to consider a penalized log-likelihood criterion instead of a penalized least-square criterion to handle the unknown variance. Following this principle, Städler et al. [68] propose to minimize the ℓ_1-penalized log-likelihood criterion

$$\widehat{\beta}^{LL}_\lambda, \widehat{\sigma}^{LL}_\lambda := \operatorname*{argmin}_{\beta\in\mathbb{R}^p,\sigma'>0} \left[n\log(\sigma') + \frac{\|Y-\mathbf{X}\beta\|_2^2}{2\sigma'^2} + \lambda\frac{\|\beta\|_1}{\sigma'} \right]. \tag{13}$$

By reparametrizing (β,σ), Städler et al. [68] obtain a convex criterion that can be efficiently minimized. Interestingly, the penalty level λ is pivotal with respect to σ. Under suitable conditions on the design matrix $\mathbf{X}$, Sun and Zhang [70] show that the choice $\lambda = c\sqrt{2\log p}$, with $c > 1$ yields optimal risk bounds in the sense of (8).

Square-root Lasso and scaled Lasso. Sun and Zhang [71], following an idea of Antoniadis [3], propose to minimize a penalized Huber's loss [44, page 179]

$$\widehat{\beta}^{SR}_\lambda, \widehat{\sigma}^{SR}_\lambda := \operatorname*{argmin}_{\beta\in\mathbb{R}^p,\sigma'>0} \left[\frac{n\sigma'}{2} + \frac{\|Y-\mathbf{X}\beta\|_2^2}{2\sigma'} + \lambda\|\beta\|_1 \right]. \tag{14}$$

This convex criterion can be minimized with roughly the same computational complexity as a Lars-Lasso path [30]. Interestingly, their procedure (called the scaled Lasso in [71]) is equivalent to the square-root Lasso estimator previously introduced by Belloni et al. [16]. The square-root Lasso of Belloni et al. is obtained

by replacing the residual sum of squares in the Lasso criterion by its square-root

$$\widehat{\beta}^{SR}_{\lambda} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left[\sqrt{\|Y - \mathbf{X}\beta\|_2^2} + \frac{\lambda}{\sqrt{n}} \|\beta\|_1 \right] . \tag{15}$$

The equivalence between the two definitions follows from the minimization of the criterion in (14) with respect to σ'. In (14) and (15), the penalty level λ is again pivotal with respect to σ. Sun and Zhang [71] state sharp oracle inequalities for the estimator $\widehat{\beta}^{SR}_{\lambda}$ with $\lambda = c\sqrt{2\log(p)}$, with $c > 1$ (see Proposition 4.2 below). Their empirical results suggest that the criterion (15) provides slightly better results than the ℓ^1-penalized log-likelihood. In the sequel, we shall refer to $\widehat{\beta}^{SR}_{\lambda}$ as the square-root Lasso estimator.

Bayesian Lasso. The Bayesian paradigm allows to put prior distributions on the variance σ^2 and the tuning parameter λ, as in the Bayesian Lasso [60]. Bayesian procedures straightforwardly handle the case of unknown variance, but no frequentist analysis of these procedures are so far available.

4.3 Risk bounds for square-root Lasso and Lasso-LinSelect

Let us state a bound on the prediction error for the square-root Lasso (also called scaled Lasso). For the sake of conciseness, we only present a simplified version of Theorem 1 in [71]. Consider some number $\xi > 0$ and some subset $T \subset \{1, \ldots, p\}$. The compatibility constant $\kappa[\xi, T]$ is defined by

$$\kappa[\xi, T] = \min_{u \in \mathcal{C}(\xi,T)} \left\{ \frac{|T|^{1/2} \|\mathbf{X}u\|_2}{\|u_T\|_1} \right\}, \quad \text{where } \mathcal{C}(\xi, T) = \{u : \ \|u_{T^c}\|_1 < \xi \|u_T\|_1\}.$$

Proposition 4.2. *There exist positive numerical constants C_1, C_2, and C_3 such that the following holds. Let us consider the square-root Lasso with the tuning parameter $\lambda = 2\sqrt{2\log(p)}$. If we assume that*

1. $p \geq C_1$
2. $\|\beta_0\|_0 \leq C_2\, \kappa^2[4, \operatorname{supp}(\beta_0)] \, \frac{n}{\log(p)}$,

then, with high probability,

$$\|\mathbf{X}(\widehat{\beta}^{SR} - \beta_0)\|_2^2 \quad \leq \quad \inf_{\beta \neq 0} \left\{ \|\mathbf{X}(\beta_0 - \beta)\|_2^2 + C_3 \frac{\|\beta\|_0 \log(p)}{\kappa^2[4, \operatorname{supp}(\beta)]} \sigma^2 \right\} .$$

This bound is comparable to the general objective (9) stated in Section 2.4. Interestingly, the constant before the bias term $\|\mathbf{X}(\beta_0 - \beta)\|_2^2$ equals one. If $\|\beta_0\|_0 = k$, the square-root Lasso achieves the minimax loss $k \log(p)\sigma^2$ as long as $k\log(p)/n$ is small and $\kappa[4, \operatorname{supp}(\beta_0)]$ is away from zero. This last condition ensures that the design $\mathbf{X}$ is not too far from orthogonality on the cone $\mathcal{C}(4, \operatorname{supp}(\beta_0))$. State of the art results for the classical Lasso with known variance [17, 48, 74] all involve this condition.

We next state a risk bound for the Lasso-LinSelect procedure. For $\mathcal{J} \subset \{1, \ldots, p\}$, we define $\phi_{\mathcal{J}}$ as the largest eigenvalue of $X_{\mathcal{J}}^T X_{\mathcal{J}}$. The following proposition involves the restricted eigenvalue $\phi_* = \max\{\phi_{\mathcal{J}} : \operatorname{Card}(\mathcal{J}) \leq n/(3\log p)\}$.

PROPOSITION 4.3. *There exist positive numerical constants C, C_1, C_2, and C_3 such that the following holds. Take $\Lambda = \mathbb{R}^+$ and assume that*

$$\|\beta_0\|_0 \leq C\,\frac{\kappa^2[5, \mathrm{supp}(\beta_0)]}{\phi_*} \times \frac{n}{\log(p)}.$$

Then, with probability at least $1 - C_1 p^{-C_2}$, the Lasso estimator $\widehat{\beta}^L_{\widehat{\lambda}}$ selected according to the LinSelect *procedure described in Section 4.1 fulfills*

$$\left\|\mathbf{X}(\beta_0 - \widehat{\beta}^L_{\widehat{\lambda}})\right\|_2^2 \leq C_3 \inf_{\beta \neq 0}\left\{\|\mathbf{X}(\beta_0 - \beta)\|_2^2 + \frac{\phi_* \|\beta\|_0 \log(p)}{\kappa^2[5, \mathrm{supp}(\beta)]}\,\sigma^2\right\}. \tag{16}$$

The bound (16) is similar to the bound stated above for the square-root Lasso, the most notable differences being the constant larger than 1 in front of the bias term and the quantity ϕ_* in front of the variance term. We refer to the Appendix E for a proof of Proposition 4.3.

4.4 Support estimation and inverse problem

Until now, we only discussed estimation methods that perform well in prediction. Little is known when the objective is to infer β_0 or its support under unknown variance.

Inverse problem. The square-root Lasso [71, 16] is proved to achieve near optimal risk bound for the inverse problems under suitable assumptions on the design $\mathbf{X}$.

Support estimation. Up to our knowledge, there are no non-asymptotic results on support estimation for the aforementioned procedures in the unknown variance setting. Nevertheless, some related results and heuristics have been developed for the cross-validation scheme. If the tuning parameter λ is chosen to minimize the prediction error (that is take $\lambda = \lambda^*$ as defined in (4)), the Lasso is not consistent for support estimation (see [51, 56] for results in a random design setting). One idea to overcome this problem, is to choose the parameter λ that minimizes the risk of the so-called Gauss-Lasso estimator $\widehat{\beta}^{GL}_\lambda$ which is the least square estimator over the support of the Lasso estimator $\widehat{\beta}^L_\lambda$

$$\widehat{\beta}^{GL}_\lambda := \underset{\beta \in \mathbb{R}^p : \mathrm{supp}(\beta) \subset \mathrm{supp}(\widehat{\beta}^L_\lambda)}{\mathrm{argmin}} \|Y - \mathbf{X}\beta\|_2^2. \tag{17}$$

When the objective is support estimation, some numerical simulations [62] suggest that it may be more advisable not to apply the selection schemes based on prediction risk (such as V-fold CV or LinSelect) to the Lasso estimators but rather to the Gauss-Lasso estimators. Similar remarks also apply for the Dantzig Selector [22].

4.5 Numerical Experiments

We present two numerical experiments to illustrate the behavior of some of the above mentioned procedures for high-dimensional sparse linear regression. The first one concerns the problem of tuning the parameter λ of the Lasso algorithm for estimating $\mathbf{X}\beta_0$. The procedures will be compared on the basis of the prediction risk. The second one concerns the problem of support estimation with

Lasso-type estimators. We will focus on the false discovery rates (FDR) and the proportion of true discoveries (Power).

Simulation design. The simulation design is the same as the one described in Sections 6.1, and 8.2 of [13], except that we restrict to the case $n = p = 100$. Therefore, 165 examples are simulated. They are inspired by examples found in [72, 85, 84, 42] and cover a large variety of situations. The simulation were carried out with `R (www.r-project.org)`, using the library `elasticnet`.

Experiment 1 : tuning the Lasso for prediction.
In the first experiment, we compare 10-fold CV [32], LinSelect [13] and the square-root Lasso [16, 71] (also called scaled Lasso) for tuning the Lasso. Concerning the square-root Lasso, we set $\lambda = 2\sqrt{2\log(p)}$ (as suggested in [71]) and we compute the estimator using the algorithm described in Sun and Zhang [71].

For each tuning procedure $\ell \in \{\text{10-fold CV}, \text{LinSelect}, \text{ square-root Lasso}\}$, we focus on the prediction risk $\mathcal{R}\left[\widehat{\beta}^L_{\hat{\lambda}_\ell}; \beta_0\right]$ of the selected Lasso estimator $\widehat{\beta}^L_{\hat{\lambda}_\ell}$.

For each simulated example $e = 1, \ldots, 165$, we estimate on the basis of 400 runs

- the risk of the oracle (4) : $\mathcal{R}_e = \mathcal{R}\left[\widehat{\beta}_{\lambda^*}; \beta_0\right]$,
- the risk when selecting λ with procedure ℓ : $\mathcal{R}_{\ell,e} = \mathcal{R}\left[\widehat{\beta}_{\hat{\lambda}_\ell}; \beta_0\right]$.

The comparison between the procedures is based on the comparison of the means, standard deviations and quantiles of the risk ratios $\mathcal{R}_{\ell,e}/\mathcal{R}_e$ computed over all the simulated examples $e = 1, \ldots, 165$. The results are displayed in Table 1.

			quantiles				
procedure	mean	std-err	0%	50%	75%	90%	95%
Lasso 10-fold CV	1.13	0.08	1.03	1.11	1.15	1.19	1.24
Lasso LinSelect	1.19	0.48	0.97	1.03	1.06	1.19	2.52
Square-root Lasso	5.15	6.74	1.32	2.61	3.37	11.2	17

TABLE 1

For each procedure ℓ, mean, standard-error and quantiles of the ratios $\{\mathcal{R}_{\ell,e}/\mathcal{R}_e, e = 1, \ldots, 165\}$.

For 10-fold CV and LinSelect, the risk ratios are close to one. For 90% of the examples, the risk of the Lasso-LinSelect is smaller than the risk of the Lasso-CV, but there are a few examples where the risk of the Lasso-LinSelect is significantly larger than the risk of the Lasso-CV. For the square-root Lasso procedure, the risk ratios are clearly larger than for the two others. An inspection of the results reveals that the square-root Lasso selects estimators with supports of small size. This feature can be interpreted as follows. Due to the bias of the Lasso-estimator, the residual variance tends to over-estimate the variance, leading the square-root Lasso to select a Lasso estimator $\widehat{\beta}^L_\lambda$ with large λ. Consequently the risk is high.

Experiment 2 : variable selection with Gauss-Lasso and square-root Lasso.
We consider now the problem of support estimation, sometimes referred as the problem of variable selection. We implement three procedures. The Gauss-Lasso procedure tuned by either 10-fold CV or LinSelect and the square-root Lasso. The support of β_0 is estimated by the support of the selected estimator.

For each simulated example, the FDR and the Power are estimated on the basis of 400 runs. The results are given on Table 2.

False Discovery rate

procedure	mean	std-err	quantiles 0%	25%	50%	75%	90%
Gauss-Lasso 10-fold CV	0.28	0.26	0	0.08	0.22	0.35	0.74
Gauss-Lasso LinSelect	0.12	0.25	0	0.002	0.02	0.13	0.33
Square-root Lasso	0.13	0.26	0	0.009	0.023	0.07	0.32

Power

procedure	mean	std-err	quantiles 0%	25%	50%	75%	90%
Gauss-Lasso 10-fold CV	0.67	0.18	0.4	0.52	0.65	0.71	1
Gauss-Lasso LinSelect	0.56	0.33	0.002	0.23	0.56	0.93	1
Square-root Lasso	0.59	0.28	0.013	0.41	0.57	0.80	1

TABLE 2

For each procedure ℓ, mean, standard-error and quantiles of FDR and Power values.

It appears that the Gauss-Lasso CV procedure gives greater values of the FDR than the two others. The Gauss-Lasso LinSelect and the square-root Lasso behave similarly for the FDR, but the values of the power are more variable for the LinSelect procedure.

Computation time.
Let us conclude this numerical section with the comparison of the computation times between the methods. For all methods the computation time depends on the maximum number of steps in the lasso algorithm and for the LinSelect method, it depends on the cardinality of $\mathbb{S}$ or equivalently on the maximum number of non-zero components of $\widehat{\beta}$. The results are shown at Table 3. The square-root Lasso is the less time consuming method, closely followed by the Lasso LinSelect method. The V-fold CV carried out with the function `cv.enet` of the `R` package `elasticnet`, pays the price of several calls to the lasso algorithm.

n	p	`max.steps`	$k_{\max}$	Lasso 10-fold CV	Lasso LinSelect	Square-root Lasso
100	100	100	21	4 s	0.21 s	0.18 s
100	500	100	16	4.8 s	0.43 s	0.4 s
500	500	500	80	300 s	11 s	6.3 s

TABLE 3

For each procedure computation time for different values of n and p. The maximum number of steps in the lasso algorithm, is taken as `max.steps` $= \min\{n, p\}$. *For the LinSelect procedure, the maximum number of non-zero components of $\widehat{\beta}$, denoted $k_{\max}$ is taken as* $k_{\max} = \min\{p, n/\log(p)\}$.

5. GROUP-SPARSITY

In the previous section, we have made no prior assumptions on the form of β_0. In some applications, there are some known structures between the covariates. As an example, we treat the now classical case of group sparsity. The covariates are assumed to be clustered into M groups and when the coefficient $\beta_{0,i}$ corresponding to the covariate $\mathbf{X}_i$ is non-zero then it is likely that all the coefficients $\beta_{0,j}$ with variables $\mathbf{X}_j$ in the same group as $\mathbf{X}_i$ are non-zero. We refer to the introduction of [8] for practical examples of this so-called group-sparsity assumption.

Let $G_1, \dots, G_M$ form a given partition of $\{1, \dots, p\}$. For $\lambda = (\lambda_1, \dots, \lambda_M)$, the group-Lasso estimator $\widehat{\beta}_\lambda$ is defined as the minimizer of the convex optimization criterion

$$\|Y - \mathbf{X}\beta\|_2^2 + \sum_{k=1}^{M} \lambda_k \|\beta^{G_k}\|_2 \,, \tag{18}$$

where $\beta^{G_k} = (\beta_j)_{j \in G_k}$. The Criterion (18) promotes solutions where all the coordinates of β^{G_k} are either zero or non-zero, leading to group selection [80]. Under some assumptions on $\mathbf{X}$, Huang and Zhang [43] or Lounici *et al.* [54] provide a suitable choice of $\lambda = (\lambda_1, \dots, \lambda_M)$ that leads to near optimal prediction bounds. As expected, this choice of $\lambda = (\lambda_1, \dots, \lambda_M)$ is proportional to σ.

As for the Lasso, V-fold CV is widely used in practice to tune the penalty parameter $\lambda = (\lambda_1, \dots, \lambda_M)$. To our knowledge, there is not yet any extension of the procedures described in Section 4.2 to the group Lasso. An alternative to cross-validation is to use LinSelect.

Tuning the group-Lasso with LinSelect. For any $\mathcal{K} \subset \{1, \dots, M\}$, we define the submatrix $\mathbf{X}_{(\mathcal{K})}$ of $\mathbf{X}$ by only keeping the columns of $\mathbf{X}$ with index in $\bigcup_{k \in \mathcal{K}} G_k$. We also write $\mathbf{X}_{G_k}$ for the submatrix of $\mathbf{X}$ built from the columns with index in G_k. The collection $\mathbb{S}$ and the function Δ are given by

$$\mathbb{S} = \left\{ \text{range}(\mathbf{X}_{(\mathcal{K})}) : \ 1 \le |\mathcal{K}| \le n/(3\log(M)) \text{ and } \sum_{k \in \mathcal{K}} |G_k| \le n/2 - 1 \right\}$$

and $\Delta\big(\text{range}(\mathbf{X}_{(\mathcal{K})})\big) = \log\left[|\mathcal{K}| \binom{|\mathcal{K}|}{M}\right]$. For a given $\Lambda \subset \mathbb{R}_+^M$, similarly to Section 4.1, we define $\widehat{\mathcal{K}}_\lambda = \left\{ k : \ \|\widehat{\beta}_\lambda^{G_k}\|_2 \neq 0 \right\}$ and

$$\widehat{\mathbb{S}} = \left\{ \text{range}(\mathbf{X}_{(\widehat{\mathcal{K}}_\lambda)}), \ \lambda \in \widehat{\Lambda} \right\}, \quad \text{with } \ \widehat{\Lambda} = \left\{ \lambda \in \Lambda, \ \text{range}(\mathbf{X}_{(\widehat{\mathcal{K}}_\lambda)}) \in \mathbb{S} \right\}.$$

Proposition C.1 in Appendix C ensures that we have for some constant $C > 1$

$$\mathcal{R}\left[\widehat{\beta}_{\widehat{\lambda}}; \beta_0\right] \le C \ \mathbb{E}\left[\inf_{\lambda \in \widehat{\Lambda}} \left\{ \|\mathbf{X}\widehat{\beta}_\lambda - \mathbf{X}\beta_0\|_2^2 + \left(\|\widehat{\beta}_\lambda\|_0 \vee |\widehat{\mathcal{K}}_\lambda| \log(M) \right) \sigma^2 \right\} \right].$$

In the following, we provide a more explicit bound. For simplicity, we restrict to the specific case where each group G_k has the same cardinality T. For $\mathcal{K} \subset \{1, \dots, M\}$, we define $\phi_{(\mathcal{K})}$ as the largest eigenvalue of $\mathbf{X}_{(\mathcal{K})}^T \mathbf{X}_{(\mathcal{K})}$ and we set

$$\phi_* = \max\left\{ \phi_{(\mathcal{K})} : \quad 1 \le |\mathcal{K}| \le \frac{n-2}{2T \vee 3\log(M)} \right\}. \tag{19}$$

We assume that all the columns of $\mathbf{X}$ are normalized to 1 and following Lounici *et al.* [54], we introduce for $1 \le s \le M$

$$\kappa_G[\xi, s] = \min_{1 \le |\mathcal{K}| \le s} \ \min_{u \in \Gamma(\xi, \mathcal{K})} \frac{\|\mathbf{X}u\|_2}{\|u_{(\mathcal{K})}\|_2} \tag{20}$$

where $\Gamma(\xi, \mathcal{K})$ is the cone of vectors $u \in \mathbb{R}^M \setminus \{0\}$ such that $\sum_{k \in \mathcal{K}^c} \lambda_k \|u^{G_k}\|_2 \le \xi \sum_{k \in \mathcal{K}} \lambda_k \|u^{G_k}\|_2$. In the sequel, $\mathcal{K}_0$ stands for the set of groups containing non-zero components of β_0.

PROPOSITION 5.1. *There exist positive numerical constants C, C_1, C_2, and C_3 such that the following holds. Assume that Λ contains $\bigcup_{\lambda\in\mathbb{R}_+}\{(\lambda,\dots,\lambda)\}$, that $T \le (n-2)/4$ and that*

$$1 \le |\mathcal{K}_0| \le C\,\frac{\kappa_G^2[3,|\mathcal{K}_0|]}{\phi_*} \times \frac{n-2}{\log(M)\vee T}\,.$$

Then, with probability larger than $1 - C_1 M^{-C_2}$, we have

$$\|\mathbf{X}\widehat{\beta}_{\widehat{\lambda}} - \mathbf{X}\beta_0\|_2^2 \le C_3\,\frac{\phi_*}{\kappa_G^2[3,|\mathcal{K}_0|]}\;|\mathcal{K}_0|\;\;(T\vee\log(M))\;.$$

This proposition provides a bound comparable to the bounds of Lounici *et al.* [54], without requiring the knowledge of the variance. Its proof can be found in Appendix E.

6. VARIATION-SPARSITY

We focus in this section on the *variation-sparse* regression. We recall that the vector $\beta^V \in \mathbb{R}^{p-1}$ of the variations of β has for coordinates $\beta_j^V = \beta_{j+1} - \beta_j$ and that the variation-sparse setting corresponds to the setting where the vector of variations β_0^V is coordinate-sparse. In the following, we restrict to the case where $n = p$ and $\mathbf{X}$ is the identity matrix. In this case, the problem of variation-sparse regression coincides with the problem of segmentation of the mean of the vector $Y = \beta_0 + \varepsilon$.

For any subset $\mathcal{I} \subset \{1,\dots,n-1\}$, we define $S_{\mathcal{I}} = \left\{\beta \in \mathbb{R}^n \,:\, \mathrm{supp}(\beta^V) \subset \mathcal{I}\right\}$ and $\widehat{\beta}_{\mathcal{I}} = \Pi_{S_{\mathcal{I}}} Y$. For any integer $q \in \{0,\dots,n-1\}$, we define also the "best" subset of size q by

$$\widehat{\mathcal{I}}_q = \operatorname*{argmin}_{|\mathcal{I}|=q} \|Y - \widehat{\beta}_{\mathcal{I}}\|_2^2.$$

Though the number of subsets $\mathcal{I} \subset \{1,\dots,n-1\}$ of cardinality q is of order n^{q+1}, this minimization can be performed using dynamic programming with a complexity of order n^2 [39]. To select $\widehat{\mathcal{I}} = \widehat{\mathcal{I}}_{\hat{q}}$ with $\hat{q}$ in $\{0,\dots,n-1\}$, any of the generic selection schemes of Section 3 can be applied. Below, we instantiate these schemes and present some alternatives.

6.1 Penalized empirical loss

When the variance σ^2 is known, penalized log-likelihood model selection amounts to select a subset $\widehat{\mathcal{I}}$ which minimizes a criterion of the form $\|Y - \widehat{\beta}_{\mathcal{I}}\|_2^2 + \mathrm{pen}(\mathrm{Card}(\mathcal{I}))$. This is equivalent to select $\widehat{\mathcal{I}} = \widehat{\mathcal{I}}_{\hat{q}}$ with $\hat{q}$ minimizing

$$\mathrm{Crit}(q) = \|Y - \widehat{\beta}_{\widehat{\mathcal{I}}_q}\|_2^2 + \mathrm{pen}(q). \tag{21}$$

Following the work of Birgé and Massart [18], Lebarbier [50] considers the penalty

$$\mathrm{pen}(q) = (q+1)\left(c_1 \log(n/(q+1)) + c_2\right)\sigma^2$$

and determines the constants $c_1 = 2, c_2 = 5$ by extensive numerical experiments (see also Comte and Rozenholc [25] for a similar approach in a more general

setting). With this choice of the penalty, the procedure satisfies a bound of the form

$$\mathcal{R}\left[\widehat{\beta}_{\widehat{\mathcal{I}}}, \beta_0\right] \;\leq\; C \inf_{\mathcal{I}\subset\{1,\dots,n-1\}} \left\{ \|\widehat{\beta}_{\mathcal{I}} - \beta_0\|_2^2 + (1+|\mathcal{I}|)\log(n/(1+|\mathcal{I}|))\,\sigma^2 \right\}. \quad (22)$$

When σ^2 is unknown, several approaches have been proposed.

Plug-in estimator. The idea is to replace σ^2 in $\mathrm{pen}(q)$ by an estimator of the variance such as $\widehat{\sigma}^2 = \sum_{i=1}^{n/2}(Y_{2i}-Y_{2i-1})^2/n$, or one of the estimators proposed by Hall and al. [40]. No theoretical results are proved in a non-asymptotic framework.

Estimating the variance by the residual least-squares. Baraud et al. [12] Section 5.4.2 propose to select q by minimizing a penalized log-likelihood criterion. This criterion can be written in the form $\mathrm{Crit}(q) = \|Y - \widehat{\beta}_{\widehat{\mathcal{I}}_q}\|_2^2(1+K\mathrm{pen}(q))$, with $K>1$ and the penalty $\mathrm{pen}(q)$ solving

$$\mathbb{E}\left[(U-\mathrm{pen}(q)V)_+\right] = \frac{1}{(q+1)\binom{n-1}{q}}\ ,$$

where $(.)_+ = \max(.,0)$, and U, V are two independent χ^2 variables with respectively $q+2$ and $n-q-2$ degrees of freedom. The resulting estimator $\widehat{\beta}_{\widehat{\mathcal{I}}}$, with $\widehat{\mathcal{I}} = \widehat{\mathcal{I}}_{\hat{q}}$, satisfies a non asymptotic risk bound similar to (22) for all $K>1$. The choice $K=1.1$ is suggested for the practice.

Slope heuristic. Lebarbier [50] implements the slope heuristic introduced by Birgé and Massart [19] for handling the unknown variance σ^2. The method consists in calibrating the penalty directly, without estimating $\widehat{\sigma}^2$. It is based on the following principle. First, there exists a so-called *minimal* penalty $\mathrm{pen}_{\min}(q)$ such that choosing $\mathrm{pen}(q) = K\mathrm{pen}_{\min}(q)$ in (21) with $K<1$ can lead to a strong overfit, whereas for $K>1$ the bound (22) is met. Second, it can be shown that there exists a *dimension jump* around the minimal penalty, allowing to estimate $\mathrm{pen}_{\min}(q)$ from the data. The slope heuristic then proposes to select q by minimizing the criterion $\mathrm{Crit}(q) = \|Y - \widehat{\beta}_{\widehat{\mathcal{I}}_q}\|_2^2 + 2\,\widehat{\mathrm{pen}}_{\min}(q)$. Arlot and Massart [7] provide a non asymptotic risk bound for this procedure. Their results are proved in a general regression model with heteroscedatic and non Gaussian errors, but with a constraint on the number of models per dimension which is not met for the family of models $(S_{\mathcal{I}})_{\mathcal{I}\subset\{1,\dots,n-1\}}$. Nevertheless, the authors indicate how to generalize their results for the problem of signal segmentation.

Finally, for practical issues, different procedures for estimating the minimal penalty are compared and implemented in Baudry et al. [15].

6.2 CV procedure

In a recent paper, Arlot and Célisse [5] consider the problem of signal segmentation using cross-validation. Their results apply in the heteroscedastic case. They consider several CV-methods, the leave-one-out, leave-p-out and V-fold CV for estimating the quadratic loss. They propose two cross-validation schemes. The first one, denoted *Procedure 5*, aims to estimate directly $\mathbb{E}\left[\|\beta_0 - \beta_{\widehat{I}_q}\|_2^2\right]$, while the second one, denoted *Procedure 6*, relies on two steps where the cross-validation is used first for choosing the best partition of dimension q, then the best dimension

q. They show that the leave-p-out CV method can be implemented with a complexity of order n^2, and they give a control of the expected CV risk. The use of CV leads to some restrictions on the subsets $\mathcal{I}$ that compete for estimating β_0. This problem is discussed in [5], Section 3 of the supplemental material.

6.3 Alternative for very high-dimensional settings

When n is very large, the dynamic programming optimization can become computationally too intensive. An attractive alternative is based on the fused Lasso proposed by Tibshirani et al. [73]. The estimator $\widehat{\beta}_\lambda^{TV}$ is defined by minimizing the convex criterion

$$\|Y - \beta\|_2^2 + \lambda \sum_{j=1}^{n-1} |\beta_{j+1} - \beta_j|,$$

where the total-variation norm $\sum_j |\beta_{j+1} - \beta_j|$ promotes solutions which are variation-sparse. The family $(\widehat{\beta}_\lambda^{TV})_{\lambda \geq 0}$ can be computed very efficiently with the LARS-algorithm, see Vert and Bleakley [75]. A sensible choice of the parameter λ must be proportional to σ. When the variance σ^2 is unknown, the parameter λ can be selected either by V-fold CV or by LinSelect (see Section 5.1 in [13] for details).

7. EXTENSIONS

7.1 Gaussian design and graphical models

Assume that the design $\mathbf{X}$ is now random and that the n rows $\mathbf{X}^{(i)}$ are independent observations of a Gaussian vector with mean 0_p and unknown covariance matrix Σ. This setting is mainly motivated by applications in compressed sensing [28] and in Gaussian graphical modeling. Indeed, Meinshausen and Bühlmann [56] have proved that it is possible to estimate the graph of a Gaussian graphical model by studying linear regression with Gaussian design and unknown variance. If we work conditionally on the observed $\mathbf{X}$ design, then all the results and methodologies described in this survey still apply. Nevertheless, these prediction results do not really take into account the fact that the design is random. In this setting, it is more natural to consider the integrated prediction risk $\mathbb{E}\big[\|\Sigma^{1/2}(\widehat{\beta} - \beta_0)\|_2^2\big]$ rather than the risk (3). Some procedures [34, 76] have been proved to achieve optimal risk bounds with respect to this risk but they are computationally intractable in a high-dimensional setting. In the context of Gaussian graphical modeling, the procedure GGMselect [38] is designed to select among any collection of graph estimators and it is proved to achieve near optimal risk bounds in terms of the integrated prediction risk.

7.2 Non Gaussian noise

A few results do not require that the noise ε follows a Gaussian distribution. The Lasso-type procedures such as the square-root Lasso [71, 16] do not require the normality of the noise and extend to other distributions. In practice, it seems that cross-validation procedures still work well for other distributions of the noise.

7.3 Multivariate regression

Multivariate regression deals with T simultaneous linear regression models $y_k = \mathbf{X}\beta_k + \varepsilon_k$, $k = 1, \ldots, T$. Stacking the y_k's in a $n \times T$ matrix Y, we obtain the

model $Y = \mathbf{X}B_0 + E$, where B_0 is a $p \times T$ matrix with columns given by β_k and E is a $n \times T$ matrix with i.i.d. entries. The classical structural assumptions on B_0 are either that most rows of B_0 are identically zero, or the rank of B_0 is small. The first case is a simple case of group sparsity and can be handled by the group-lasso as in Section 5. The second case, first considered by Anderson [2] and Izenman [45], is much more non-linear. Writing $\|.\|_F$ for the Frobenius (or Hilbert-Schmidt) norm, the problem of selecting among the estimators

$$\widehat{B}_r = \underset{B:\mathrm{rank}(B)\leq r}{\mathrm{argmin}} \|Y - \mathbf{X}B\|_F^2, \quad r \in \{1, \dots, \min(T, \mathrm{rank}(\mathbf{X}))\}$$

has been investigated recently from a non-asymptotic point of view by Bunea *et al.* [20] and Giraud [36]. The prediction risk of $\widehat{B}_r$ is of order of

$$\mathbb{E}\left[\|\mathbf{X}\widehat{B}_r - \mathbf{X}B_0\|_F^2\right] \asymp \sum_{k\geq r} s_k^2(\mathbf{X}B_0) + r\left(n + \mathrm{rank}(X)\right)\sigma^2,$$

where $s_k(M)$ denotes the k-th largest singular value of the matrix M. Therefore, a sensible choice of r depends on σ^2. The first selection criterion introduced by Bunea *et al.* [20] requires the knowledge of the variance σ^2. To handle the case of unknown variance, Bunea *et al.* [20] propose to plug an estimate of the variance in their selection criterion (which works when $\mathrm{rank}(\mathbf{X}) < n$), whereas Giraud [36] introduces a penalized log-likelihood criterion independent of the variance. Both papers provide oracle risk bounds for the resulting estimators showing rate-minimax adaptation.

Several recent papers [9, 58, 63, 20, 48] have investigated another strategy for the low-rank setting. For a positive λ, the matrix B_0 is estimated by

$$\widehat{B}_\lambda \in \underset{B\in\mathbb{R}^{p\times T}}{\mathrm{argmin}} \left\{\|Y - \mathbf{X}B\|_F^2 + \lambda \sum_k s_k(B)\right\}.$$

Translating the work on trace regression of Koltchinskii *et al.* [48] into the setting of multivariate regression provides (under some conditions on $\mathbf{X}$) an oracle bound on the risk of $\widehat{B}_{\lambda^*}$ with $\lambda^* = 3s_1(X)\big(\sqrt{T} + \sqrt{\mathrm{rank}(X)}\,\big)\sigma$. We also refer to Giraud [37] for a slight variation of this result requiring no condition on the design $\mathbf{X}$. Again, the value of λ^* is proportional to σ. To handle the case of unknown variance, Klopp [47] adapts the concept of the square-root Lasso [16] to this setting and provides an oracle risk bound for the resulting procedure.

7.4 Nonparametric regression

In the nonparametric regression model (2), classical estimation procedures include local-polynomial estimators, kernel estimators, basis-projection estimators, k-nearest neighbors etc. All these procedures depend on one (or several) tuning parameter(s), whose optimal value(s) scales with the variance σ^2. V-fold CV is widely used in practice for choosing these parameters, but little is known on its theoretical performance.

The class of linear estimators (including spline smoothing, Nadaraya estimators, k-nearest neighbors, low-pass filters, kernel ridge regression, etc) has attracted some attention in the last years. Some papers have investigated the tuning of some specific family of estimators. For example, Cao and Golubev [23] provides

a tuning procedure for spline smoothing and Zhang [82] analyses in depth kernel ridge regression. Recently, two papers have focused on the tuning of arbitrary linear estimators when the variance σ^2 is unknown. Arlot and Bach [4] generalize the slope heuristic to symmetric linear estimators with spectrum in $[0, 1]$ and prove an oracle bound for the resulting estimator. Baraud *et al.* [13] Section 4 shows that LinSelect can be used for selecting among a (almost) completely arbitrary collection of linear estimators (possibly non-symmetric and/or singular). Corollary 2 in [13] provides an oracle bound for the selected estimator under the mild assumption that some effective dimension of the linear estimators is not larger than a fraction of n.

REFERENCES

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*. Akadémiai Kiadó, Budapest, 267–281. MR0483125 (58 #3144)

[2] Anderson, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Ann. Math. Statistics 22*, 327–351. MR0042664 (13,144f)

[3] Antoniadis, A. (2010). Comments on: ℓ_1-penalization for mixture regression models [mr2677722]. *TEST* **19**, 2, 257–258. http://dx.doi.org/10.1007/s11749-010-0198-y. MR2677723

[4] Arlot, S. and Bach, F. (2009). Data-driven calibration of linear estimators with minimal penalties. In *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, Eds. 46–54.

[5] Arlot, S. and Célisse, A. (2010a). Segmentation of the mean of heteroscedastic data via cross-validation. *Stat. Comput.* **21**, 4, 1–20. 10.1007/s11222-010-9196-x, http://dx.doi.org/10.1007/s11222-010-9196-x.

[6] Arlot, S. and Célisse, A. (2010b). A survey of cross-validation procedures for model selection. *Stat. Surv. 4*, 40–79. http://dx.doi.org/10.1214/09-SS054. MR2602303 (2011g:62111)

[7] Arlot, S. and Massart, P. (2010). Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res. 10*, 245–279.

[8] Bach, F. (2008a). Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res. 9*, 1179–1225. MR2417268 (2010a:68132)

[9] Bach, F. (2008b). Consistency of trace norm minimization. *J. Mach. Learn. Res. 9*, 1019–1048. MR2417263

[10] 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

[11] Baraud, Y. (2011). Estimator selection with respect to hellinger-type risks. *Probab. Theory Related Fields* **151**, 1–2, 353–401. 10.1007/s00440-010-0302-y, http://dx.doi.org/10.1007/s00440-010-0302-y.

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

[13] Baraud, Y., Giraud, C., and Huet, S. (2010). Estimator selection in the gaussian setting. arXiv:1007.2096v2.

[14] Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113**, 3, 301–413. http://dx.doi.org/10.1007/s004400050210. MR1679028 (2000k:62049)

[15] Baudry, J.-P., Maugis, C., and Michel, B. (2012). Slope heuristics: Overview and implementation. *Statist. Comput.* **22**, 2, 455–470.

[16] Belloni, A., Chernozhukov, V., and Wang, L. (2011). Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika* **98**, 4, 791–806.

[17] 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

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

[19] 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)

[20] Bunea, F., She, Y., and Wegkamp, M. H. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *Ann. Stat.* **39**, 2, 1282–1309.

[21] Bunea, F., Tsybakov, A., and Wegkamp, M. (2007). Aggregation for Gaussian regression. *Ann. Statist.* **35**, 4, 1674–1697. MR2351101

[22] 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

[23] Cao, Y. and Golubev, Y. (2006). On oracle inequalities related to smoothing splines. *Math. Methods Statist.* **15**, 4, 398–414 (2007). MR2301659 (2008i:62039)

[24] Chen, S., Donoho, D., and Saunders, M. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.* **20**, 1, 33–61. http://dx.doi.org/10.1137/S1064827596304010. MR1639094 (99h:94013)

[25] Comte, F. and Rozenholc, Y. (2004). A new algorithm for fixed design regression and denoising. *Ann. Inst. Statist. Math.* **56**, 3, 449–473. http://dx.doi.org/10.1007/BF02530536. MR2095013 (2005e:62081)

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

[27] Devroye, L. P. and Wagner, T. J. (1979). The L_1 convergence of kernel density estimates. *Ann. Statist.* **7**, 5, 1136–1139. MR536515 (80k:62054)

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

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

[30] 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)

[31] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 456, 1348–1360. http://dx.doi.org/10.1198/016214501753382273. MR1946581 (2003k:62160)

[32] Geisser, S. (1975). The predictive sample reuse method with applications. *J. Amer. Statist. Assoc. 70*, 320–328.

[33] Gerchinovitz, S. (2011). Sparsity regret bounds for individual sequences in online linear regression. *Proceedings of COLT 2011*.

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

[35] Giraud, C. (2008b). Mixing least-squares estimators when the variance is unknown. *Bernoulli* **14**, 4, 1089–1107. http://dx.doi.org/10.3150/08-BEJ135. MR2543587 (2010k:62274)

[36] Giraud, C. (2011a). Low rank multivariate regression. *Electron. J. Stat. 5*, 775–799.

[37] Giraud, C. (2011b). A pseudo-rip for multivariate regression. Arxiv:1106.5599v1.

[38] Giraud, C., Huet, S., and Verzelen, N. (2012). Graph selection with GGMselect. *Stat. Appl. Genet. Mol. Biol.* **11**, 3, 1–50.

[39] Guthery, S. B. (1974). A transformation theorem for one-dimensional F-expansions. *J. Number Theory 6*, 201–210. MR0342484 (49 #7230)

[40] Hall, P., Kay, J. W., and Titterington, D. M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika* **77**, 3, 521–528. http://dx.doi.org/10.1093/biomet/77.3.521. MR1087842 (92d:62042)

[41] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning*, Second ed. Springer Series in Statistics. Springer, New York. Data mining, inference, and prediction, http://dx.doi.org/10.1007/978-0-387-84858-7. MR2722294

[42] Huang, J., Ma, S., and Zhang, C.-H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statist. Sinica* **18**, 4, 1603–1618. MR2469326 (2010a:62214)

[43] Huang, J. and Zhang, T. (2010). The benefit of group sparsity. *Ann. Statist.* **38**, 4, 1978–2004. http://dx.doi.org/10.1214/09-AOS778. MR2676881 (2011f:62029)

[44] Huber, P. (1981). *Robust statistics.* John Wiley & Sons Inc., New York. Wiley Series in Probability and Mathematical Statistics. MR606374 (82i:62057)

[45] Izenman, A. (1975). Reduced-rank regression for the multivariate linear model. *J. Multivariate Anal. 5*, 248–264. MR0373179 (51 #9381)

[46] Ji, P. and Jin, J. (2010). Ups delivers optimal phase diagram in high dimensional variable selection. http://arxiv.org/abs/1010.5028.

[47] Klopp, O. (2011). High dimensional matrix estimation with unknown variance of the noise. Arxiv:1112.3055v1.

[48] Koltchinski, V., Lounici, K., and Tsybakov, A. (2011). Nuclear norm penalization and optimal rates for noisy low rank matrix completion. *Annals of Statistics* **39**, 5, 2302–2329.

[49] Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28**, 5, 1302–1338.

[50] Lebarbier, E. (2005). Detecting multiple change-points in the mean of gaussian process by model selection. *Signal Processing 85*, 717–736.

[51] Leng, C., Lin, Y., and Wahba, G. (2006). A note on the lasso and related procedures in model selection. *Statist. Sinica* **16**, 4, 1273–1284. MR2327490

[52] Leung, G. and Barron, A. (2006). Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory* **52**, 8, 3396–3410. http://dx.doi.org/10.1109/TIT.2006.878172. MR2242356

[53] Li, K.-C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. *Ann. Statist.* **15**, 3, 958–975. http://dx.doi.org/10.1214/aos/1176350486. MR902239 (89c:62112)

[54] Lounici, K., Pontil, M., Tsybakov, A., and van de Geer, S. (2011). Oracle inequalities and optimal inference under group sparsity. *Annals of Statistics* **39**, 4, 2164–2204.

[55] Mallows, C. L. (1973). Some comments on c_p. *Technometrics 15*, 661–675.

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

[57] Mosteller, F. and Tukey, J. (1968). Data analysis, including statistics. In *Handbook of Social Psychology, Vol. 2*, G. Lindsey and E. Aronson, Eds. Addison-wesley.

[58] Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Statist.* **39**, 2, 1069–1097. http://dx.doi.org/10.1214/10-AOS850. MR2816348

[59] Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.* **12**, 2, 758–765. http://dx.doi.org/10.1214/aos/1176346522. MR740928 (86f:62109)

[60] Park, T. and Casella, G. (2008). The Bayesian lasso. *J. Amer. Statist. Assoc.* **103**, 482, 681–686. http://dx.doi.org/10.1198/016214508000000337. MR2524001

[61] Raskutti, G., Wainwright, M., and Yu, B. (2011). Minimax rates of estimations for high-dimensional regression over l^q balls. *IEEE Trans. Inf. Theory* **57**, 10, 6976–6994.

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

[63] Rohde, A. and Tsybakov, A. (2011). Estimation of high-dimensional low-rank matrices. *Ann. Statist.* **39**, 2, 887–930. http://dx.doi.org/10.1214/10-AOS860. MR2816342

[64] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 2, 461–464. MR0468014 (57 #7855)

[65] Shao, J. (1993). Linear model selection by cross-validation. *J. Amer. Statist. Assoc.* **88**, 422, 486–494. MR1224373 (94k:62107)

[66] Shao, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica* **7**, 2, 221–264. With comments and a rejoinder by the author. MR1466682 (99m:62104)

[67] Shibata, R. (1981). An optimal selection of regression variables. *Biometrika* **68**, 1, 45–54. http://dx.doi.org/10.1093/biomet/68.1.45. MR614940 (84a:62103a)

[68] Städler, N., Bühlmann, P., and van de Geer, S. (2010). ℓ_1-penalization for mixture regression models. *TEST* **19**, 2, 209–256. http://dx.doi.org/10.1007/s11749-010-0197-z. MR2677722

[69] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B 36*, 111–147. With discussion by G. A. Barnard, A. C. Atkinson, L. K. Chan, A. P. Dawid, F. Downton, J. Dickey, A. G. Baker, O. Barndorff-Nielsen, D. R. Cox, S. Giesser, D. Hinkley, R. R. Hocking, and A. S. Young, and with a reply by the authors. MR0356377 (50 #8847)

[70] Sun, T. and Zhang, C.-H. (2010). Comments on: ℓ_1-penalization for mixture regression models [mr2677722]. *TEST* **19**, 2, 270–275. http://dx.doi.org/10.1007/s11749-010-0201-7. MR2677726

[71] Sun, T. and Zhang, C.-H. (2011). Scaled sparse linear regression. arXiv:1104.4595.

[72] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 1, 267–288. MR1379242 (96j:62134)

[73] Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67**, 1, 91–108. http://dx.doi.org/10.1111/j.1467-9868.2005.00490.x. MR2136641

[74] van de Geer, S. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat. 3*, 1360–1392. http://dx.doi.org/10.1214/09-EJS506. MR2576316 (2011c:62231)

[75] Vert, J.-P. and Bleakley, K. (2010). Fast detection of multiple change-points shared by many signals using group lars. In *Advances in Neural Information Processing Systems 23*, J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, Eds. 2343–2351.

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

[77] Verzelen, N. (2012). Minimax risks for sparse regressions: Ultra-high-dimensional phenomenons. *Electron. J. Stat. 6*, 38–90.

[78] Wainwright, M. (2009). 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

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

[80] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68**, 1, 49–67. http://dx.doi.org/10.1111/j.1467-9868.2005.00532.x. MR2212574

[81] Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38**, 2, 894–942. http://dx.doi.org/10.1214/09-AOS729. MR2604701 (2011d:62211)

[82] Zhang, T. (2005). Learning bounds for kernel regression using effective data dimensionality. *Neural Comput.* **17**, 9, 2077–2098. http://dx.doi.org/10.1162/0899766054323008. MR2175849 (2006d:62062)

[83] Zhang, T. (2011). Adaptive Forward-Backward Greedy Algorithm for Learning Sparse Representations. *IEEE Trans. Inform. Theory* **57**, 7, 4689–4708.

[84] Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 476, 1418–1429. MR2279469 (2008d:62024)

[85] 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

APPENDIX A: A NOTE ON BIC TYPE CRITERIA

The BIC criterion has been initially introduced [64] to select an estimator among a collection of constrained maximum likelihood estimators. Nevertheless, modified versions of this criterion are often used for tuning more general estimation procedures. The purpose of this appendix is to illustrate why we advise against this approach in a high-dimensional setting.

DEFINITION A.1. **A Modified BIC criterion**. *Suppose we are given a collection* $(\widehat{\beta}_\lambda)_{\lambda\in\Lambda}$ *of estimators depending on a tuning parameter* $\lambda \in \Lambda$. *For any* $\lambda \in \Lambda$, *we consider* $\widehat{\sigma}^2_\lambda = \|Y - \mathbf{X}\widehat{\beta}_\lambda\|_2^2/n$, *and define the modified* BIC

$$\widehat{\lambda} \in \operatorname*{argmin}_{\lambda\in\widehat{\Lambda}} \left\{ -2\mathbf{L}_n(\widehat{\beta}_\lambda, \widehat{\sigma}_\lambda) + \log(n)\|\widehat{\beta}_\lambda\|_0 \right\}, \tag{A.1}$$

where $\mathbf{L}_n$ *is the log-likelihood and* $\widehat{\Lambda} = \left\{ \lambda \in \Lambda \,:\, \|\widehat{\beta}_\lambda\|_0 \le n/2 \right\}$.

Sometimes, the $\log(n)$ term is replaced by $\log(p)$. Replacing Λ by $\widehat{\Lambda}$ allows to avoid trivial estimators. First, we would like to emphasize that there is *no* theoretical warranty that the selected estimator does not overfit in a high-dimensional setting. In practice, using this criterion often leads to overfitting. Let us illustrate this with a simple experiment.

Setting. We consider the model

$$Y_i = \beta_{0,i} + \varepsilon_i,\ i = 1, \dots, n\ , \tag{A.2}$$

with $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ so that $p = n$ and $\mathbf{X} = I_n$. Here, we fix $n = 10000$, $\sigma = 1$ and $\beta_0 = 0_n$.

Methods. We apply the modified BIC criterion to tune the Lasso [72], SCAD [31] and the hard thresholding estimator. The hard thresholding estimator $\widehat{\beta}^{HT}_\lambda$ is defined for any $\lambda > 0$ by $[\widehat{\beta}^{HT}_\lambda]_i = Y_i \mathbf{1}_{|Y_i|\ge\lambda}$. Given $\lambda > 0$ and $a > 2$, the SCAD estimator $\widehat{\beta}^{SCAD}_{\lambda,a}$ is defined as the minimizer of the penalized criterion $\|Y - \mathbf{X}\beta\|_2^2 + \sum_{i=1}^n p_\lambda(|\beta_i|)$, where for $x > 0$,

$$p'_\lambda(x) = \lambda \mathbf{1}_{x\le\lambda} + (a\lambda - x)_+ \mathbf{1}_{x>\lambda}/(a-1)\ .$$

For the sake of simplicity we fix $a = 3$. We note $\widehat{\beta}^{L;\mathsf{BIC}}$, $\widehat{\beta}^{SCAD;\mathsf{BIC}}_a$, and $\widehat{\beta}^{HT;\mathsf{BIC}}$ for the Lasso, hard thresholding, and SCAD estimators selected by the modified BIC criterion.

Results. We have realized $N = 200$ experiments. For each of these experiments, the estimator $\widehat{\beta}^{L;\mathsf{BIC}}$, $\widehat{\beta}^{SCAD,\mathsf{BIC}}_a$ and $\widehat{\beta}^{HT;\mathsf{BIC}}$ are computed. The mean number of non-zero components and the estimated risk $\mathcal{R}[\widehat{\beta}^{*;\mathsf{BIC}}; 0_n]$ are reported in Table 1.

	LASSO	SCAD	Hard Thres.
$\widehat{\mathcal{R}}[\widehat{\beta}^{*;\mathsf{BIC}}; 0_p]$	4.6×10^{-2}	1.6×10^{1}	3.0×10^{2}
Mean of $\|\widehat{\beta}^{*;\mathsf{BIC}}\|_0$	0.025	86.9	28.2

Table 1: Estimated risk and Estimated number of non zero components for $\widehat{\beta}^{L;\mathsf{BIC}}$, $\widehat{\beta}^{SCAD;\mathsf{BIC}}$, and $\widehat{\beta}^{HT;\mathsf{BIC}}$.

Obviously, the SCAD and hard Thresholding methods select too many irrelevant variables when they are tuned with BIC. Moreover, their risks are quite high. Intuitively, this is due to the fact that the $\log(n)$ (or $\log(p)$) term in the BIC penalty is too small in this high-dimensional setting ($n = p$).

For the Lasso estimator, a very specific phenomenon occurs due to the soft thresholding effect. In the discussion of [30], Loubes and Massart advocate that soft thresholding estimators penalized by Mallows' C_p [55] penalties should yield good results, while hard thresholding estimators penalized by Mallows' C_p are known to highly overfit. This strange behavior is due to the bias of the soft thresholding estimator. Nevertheless, Loubes and Massart' arguments have been developed for an orthogonal design. In fact, there is no non-asymptotic justification that the Lasso tuned by BIC or AIC performs well for general designs $\mathbf{X}$.

Conclusion. The use of the modified BIC criterion to tune estimation procedures in a high-dimensional setting is not supported by theoretical results. It is proved to overfit in the case of thresholding estimators [12, Sect. 3.2.2]. Empirically, BIC seems to overfit except for the Lasso. We advise the practitioner to avoid BIC (and AIC) when p is at least of the same order as n. For instance, LinSelect is supported by non-asymptotic arguments and by empirical results [13] in contrast to BIC.

APPENDIX B: MINIMAX ADAPTIVE PROCEDURES

In this section, we detail procedures that are minimax adaptive to the sparsity k simultaneously for all designs $\mathbf{X}$ in the sense of (7). In most settings, these procedures are not of practical interest as they are intractable for large p. We present them as theoretical benchmarks to assess the quality of fast procedures.

Given a subspace S of $\mathbb{R}^n$, we define $\widehat{\beta}_S^{\perp}$ as a least-squares estimator of β_0 such that $\mathbf{X}\beta$ is included in S:

$$\widehat{\beta}_S^{\perp} \in \underset{\beta\in\mathbb{R}^p,\ \mathbf{X}\beta\in S}{\operatorname{argmin}} \|Y-\mathbf{X}\beta\|_2^2 \ .$$

We consider the collections of subspaces:

$$\begin{aligned}
\mathbb{S}_1 &= \Big\{S=\operatorname{range}(\mathbf{X}_{\mathcal{J}}),\ \mathcal{J}\subset\{1,\dots,p\}\setminus\{\emptyset\},\quad 2|\mathcal{J}|[1+\log(p/|\mathcal{J}|)]\le n\Big\} \\
&\qquad \bigcup \operatorname{range}(\mathbf{X}_{\{1,\dots p\}})\,, \\
\mathbb{S}_2 &= \Big\{S=\operatorname{range}(\mathbf{X}_{\mathcal{J}}),\ \mathcal{J}\subset\{1,\dots,p\}\setminus\{\emptyset\},\quad |\mathcal{J}|\le (n-1)/4\Big\}\ .
\end{aligned}$$

Finally, we note $k^* := \max\{k : 2k[1+\log(p/k)]\le n\}$. To simplify the presentation, we assume throughout this section that $n\le p$ and that $\operatorname{Rank}(\mathbf{X})>k^*$.

B.1 Known variance

A penalization strategy. The model selection paradigm aims at selecting an estimator $\widehat{\beta}_{\widehat{S}}$ with the smallest possible risk. One strategy to tackle the selection problem amounts to minimizing a least-squares criterion penalized by the "complexity" of the collection of models under consideration. We select $\widehat{S}^{BM}$ as one minimizer over $S\in\mathbb{S}_1$ of the following criterion

$$\|Y-\Pi_S Y\|_2^2 + \begin{cases} 4\dim(S)\left[4+\log\left(\frac{p}{\dim(S)}\right)\right]\sigma^2 & \text{if } \dim(S)\le k^* \\ 2n\sigma^2 & \text{if } \dim(S)=\operatorname{Rank}(\mathbf{X})\,, \end{cases}$$

We write $\widetilde{\beta}^{BM} := \widehat{\beta}^{\perp}_{\widehat{S}^{BM}}$. More general forms of penalties are discussed in [18].

An aggregation strategy. In contrast to model selection, model aggregation aims at mixing a collection of estimators. Following, Leung and Barron [52], we mix the least-squares estimators $\widehat{\beta}_S$ in the following way

$$\widetilde{\beta}^{LB} := \sum_{S \in \mathbb{S}_1} \omega_S \widehat{\beta}_S^{\perp} \ ,$$

where the weights ω_S sum to one and for any $S \in \mathbb{S}_1$, ω_S is proportional to

$$\exp\left[-\frac{\|Y - \Pi_S Y\|_2^2 + 2\sigma^2 \dim(S)}{4\sigma^2}\right] \times \begin{cases} \left[k^* \binom{\dim(S)}{p}\right]^{-1} & \text{if } \dim(S) \leq k^* \\ 1 & \text{if } \dim(S) = \mathrm{Rank}(\mathbf{X}). \end{cases}$$

We refer to [52] for more general forms of the aggregation procedures.

Risk bounds. In the next proposition, we state that $\widetilde{\beta}^{BM}$ and $\widetilde{\beta}^{LB}$ are minimax adaptive to the sparsity for all designs $\mathbf{X}$ in the sense of (7).

PROPOSITION B.1. *There exist numerical constants C_1 and C_2 such that the following holds. For any design $\mathbf{X}$, any $k \in \{1, \ldots, n\}$ and any vector β_0 such that $\|\beta_0\|_0 = k$, we have*

$$\begin{aligned} \mathcal{R}\left[\widetilde{\beta}^{BM}; \beta_0\right] &\leq C_1 \left[k\left(1 + \log\left(\frac{p}{k}\right)\right) \wedge n\right] \sigma^2 \ , \\ \mathcal{R}\left[\widetilde{\beta}^{LB}; \beta_0\right] &\leq C_2 \left[k\left(1 + \log\left(\frac{p}{k}\right)\right) \wedge n\right] \sigma^2 \ . \end{aligned}$$

These two risk bounds derive straightforwardly from the aforementioned work [18, 52].

B.2 Unknown variance

For any set $S \in \mathbb{S}_2$, we set the following measure of complexity $\Delta(S)$

$$\Delta(S) = \log\binom{p}{\dim(S)} + \log(\dim(S)) \ ,$$

and we take the same penalty term $\mathrm{pen}(S)$ as for LinSelect (see Appendix C.1). Baraud et al. [12] consider the model selection estimators $\widetilde{\beta}^{BGH} := \widehat{\beta}^{\perp}_{\widehat{S}^{BGH}}$ with

$$\widehat{S}^{BGH} := \operatorname*{argmin}_{S \in \mathbb{S}_2} \|Y - \Pi_S Y\|_2^2 \left[1 + \frac{\mathrm{pen}(S)}{n - \dim(S)}\right].$$

The first risk bound only covers the (non-ultra) high-dimensional setting.

PROPOSITION B.2. *There exists some numerical constant C such that the following holds. For any design $\mathbf{X}$ and any vector β_0, we have*

$$\mathcal{R}\left[\widetilde{\beta}^{BGH}; \beta_0\right] \leq C \inf_{\substack{\beta \in \mathbb{R}^p \\ \|\beta\|_0 \leq \frac{n}{2\log(p)}}} \left\{\|\mathbf{X}(\beta - \beta_0)\|_2^2 + \|\beta\|_0 \left[1 + \log\left(\frac{p}{\|\beta\|_0}\right)\right] \sigma^2\right\}.$$

Proposition B.2 is a straightforward consequence of Corollary 1 in [12]. It shows that simultaneous adaptation to the variance and the sparsity is possible if we restrict ourselves to a non-ultra high-dimensional setting. The next proposition complements the risk upper bound of Proposition 2.2. Consider $\widetilde{\beta}^{(n)}$ as a least-squares estimator of β_0 over $\mathbb{R}^n$.

PROPOSITION B.3. *There exist numerical constants C, C_1, and C_2 such that the following holds. For any design $\mathbf{X}$, any $\sigma > 0$, and any vector $\beta_0 \in \mathbb{R}^p$, we have*

$$\mathcal{R}\left[\widetilde{\beta}^{(n)};\beta_0\right] \quad \leq \quad Cn\sigma^2.$$

For any design $\mathbf{X}$, any $\sigma > 0$, any $k \in \{1,\ldots,(n-1)/4\}$ and any vector $\beta_0 \in \mathbb{R}^p$ such that $\|\beta_0\|_0 = k$, we have

$$\mathcal{R}\left[\widetilde{\beta}^{BGH};\beta_0\right] \quad \leq \quad C_1 k \log\left(\frac{p}{k}\right)\exp\left[C_2\frac{k}{n}\log\left(\frac{p}{k}\right)\right]\sigma^2.$$

The first bound is straightforward while the second bound derives from [12].

APPENDIX C: COMPLEMENTS ON LINSELECT

C.1 More details on the selection procedure

The penalty $\mathrm{pen}(S)$ involved in the LinSelect criterion (11) is defined by $\mathrm{pen}(S) = 1.1\,\mathrm{pen}_\Delta(S)$ where $\mathrm{pen}_\Delta(S)$ is the unique solution of

$$\mathbb{E}\left[\left(U - \frac{\mathrm{pen}_\Delta(S)}{n-\dim(S)}V\right)_+\right] = e^{-\Delta(S)}$$

where U and V are two independent chi-square random variables with $\dim(S)+1$ and $n-\dim(S)-1$ degrees of freedom respectively. It is also the solution in x of

$$e^{-\Delta(S)} = \\ (D+1)\mathbb{P}\left(F_{D+3,N-1} \geq x\frac{N-1}{N(D+3)}\right) - x\frac{N-1}{N}\mathbb{P}\left(F_{D+1,N+1} \geq x\frac{N+1}{N(D+1)}\right)$$

where $D = \dim(S)$, $N = n - \dim(S)$ and $F_{d,r}$ is a Fisher random variable with d and r degrees of freedom.

Proposition 4 in [12] ensures the following upper-bound on $\mathrm{pen}_\Delta(S)$. For any $0 < \kappa < 1$, there exists a constant $C_\kappa > 1$ such that for any $S \in \mathbb{S}$ fulfilling $1 \leq \dim(S) \vee \Delta(S) \leq \kappa n$ we have

$$\mathrm{pen}_\Delta(S) \leq C_\kappa\big(\dim(S) \vee \Delta(S)\big).$$

Conversely, Lemma D.3 in Appendix D ensures that $\mathrm{pen}_\Delta(S) \geq 2\Delta(S) + \dim(S) - C$ for some constant $C \geq 0$.

C.2 A general risk bound for LinSelect

We set

$$\Sigma = \sigma^2 \sum_{S\in\mathbb{S}} e^{-\Delta(S)}. \tag{C.1}$$

The following proposition gives a risk bound when selecting $\widehat{\lambda}$ by minimizing (11).

PROPOSITION C.1. *Assume that* $1 \leq \dim(S) \leq n/2-1$ *and* $\Delta(S) \leq 2n/3$ *for all* $S \in \mathbb{S}$. *Then, there exists a constant* $C > 1$ *such that for any minimizer* $\widehat{\lambda}$ *of the Criterion (11), we have*

$$C^{-1}\mathcal{R}\left[\widehat{\beta}_{\widehat{\lambda}};\beta_0\right] \leq \mathbb{E}\left[\inf_{\lambda\in\Lambda}\left\{\|\mathbf{X}\widehat{\beta}_\lambda - \mathbf{X}\beta_0\|_2^2 + \inf_{S\in\widehat{\mathbb{S}}}\left\{\|\mathbf{X}\widehat{\beta}_\lambda - \Pi_S\mathbf{X}\widehat{\beta}_\lambda\|_2^2 + [\Delta(S)\vee\dim(S)]\sigma^2\right\}\right\}\right] + \Sigma.$$

Furthermore, with probability larger than $1-e^{-C_0 n} - C_1\sum_{S\in\mathbb{S}} e^{-C_2[\Delta(S)\wedge n]}e^{-\Delta(S)}$, *we have for some* $C > 1$

$$C^{-1}\left\|\mathbf{X}\beta_0 - \mathbf{X}\widehat{\beta}_{\widehat{\lambda}}\right\|_2^2 \leq \inf_{\lambda\in\Lambda}\left\{\|\mathbf{X}\widehat{\beta}_\lambda - \mathbf{X}\beta_0\|_2^2 + \inf_{S\in\widehat{\mathbb{S}}}\left\{\|\mathbf{X}\widehat{\beta}_\lambda - \Pi_S\mathbf{X}\widehat{\beta}_\lambda\|_2^2 + [\Delta(S)\vee\dim(S)]\sigma^2\right\}\right\}.$$

The first part of Proposition C.1 is a slight variation of Theorem 1 in [13]. We refer to the Appendix D.1 for a sketch of the proof of this result. The second part is proved in Appendix D.2.

APPENDIX D: PROOF OF PROPOSITION C.1

D.1 Proof of the first part of Proposition C.1

In this section C denotes a constant whose value may vary from line to line. We also use in this section the notations $\|.\|$ for $\|.\|_2$, $f_0 = \mathbf{X}\beta_0$ and $\widehat{f}_\lambda = \mathbf{X}\widehat{\beta}_\lambda$. Finally, for any $S \in \mathbb{S}$, we write $\overline{S}$ for the linear space generated by S and f_0. Let $(\widehat{\lambda}, S_*)$ be any minimizer over $\Lambda \times \widehat{\mathbb{S}}$ of

$$\mathrm{Crit}(\lambda, S) = \left\|Y - \Pi_S\widehat{f}_\lambda\right\|^2 + \frac{1}{2}\left\|\widehat{f}_\lambda - \Pi_S\widehat{f}_\lambda\right\|^2 + \mathrm{pen}(S)\widehat{\sigma}_S^2.$$

From $\mathrm{Crit}(\widehat{\lambda}, S_*) \leq \mathrm{Crit}(\lambda, S)$ and simple algebra, we get for any $K > 1$, $\lambda \in \widehat{\Lambda}$ and $S \in \widehat{\mathbb{S}}$

$$\begin{aligned}
&\left\|f_0 - \Pi_{S_*}\widehat{f}_{\widehat{\lambda}}\right\|^2 + \frac{1}{2}\left\|\widehat{f}_{\widehat{\lambda}} - \Pi_{S_*}\widehat{f}_{\widehat{\lambda}}\right\|^2 \\
&\leq \left\|f_0 - \Pi_S\widehat{f}_\lambda\right\|^2 + \frac{1}{2}\left\|\widehat{f}_\lambda - \Pi_S\widehat{f}_\lambda\right\|^2 + 2\mathrm{pen}(S)\widehat{\sigma}_S^2 \\
&\quad + 2\langle\varepsilon, \Pi_{S_*}\widehat{f}_{\widehat{\lambda}} - f_0\rangle - \mathrm{pen}(S_*)\widehat{\sigma}_{S_*}^2 \;+\; 2\langle\varepsilon, f_0 - \Pi_S\widehat{f}_\lambda\rangle - \mathrm{pen}(S)\widehat{\sigma}_S^2. \\
&\leq \left\|f_0 - \Pi_S\widehat{f}_\lambda\right\|^2 + \frac{1}{2}\left\|\widehat{f}_\lambda - \Pi_S\widehat{f}_\lambda\right\|^2 + 2\mathrm{pen}(S)\widehat{\sigma}_S^2 \\
&\quad + K^{-1}\left\|f_0 - \Pi_{S_*}\widehat{f}_{\widehat{\lambda}}\right\|^2 + K\left\|\Pi_{\bar{S}_*}\varepsilon\right\|^2 - \mathrm{pen}(S_*)\widehat{\sigma}_{S_*}^2 \\
&\quad + K^{-1}\left\|f_0 - \Pi_S\widehat{f}_\lambda\right\|^2 + K\left\|\Pi_{\bar{S}}\varepsilon\right\|^2 - \mathrm{pen}(S)\widehat{\sigma}_S^2,
\end{aligned}$$

the second inequality following from $2\langle f, g\rangle \leq K^{-1}\|f\|^2 + K\|g\|^2$. Introducing the notation

$$\tilde{\Sigma} = 2\sum_{S\in\mathbb{S}}\left(K\left\|\Pi_{\overline{S}}\varepsilon\right\|^2 - \frac{\mathrm{pen}(S)}{n-\dim(S)}\left\|Y - \Pi_{\overline{S}}Y\right\|^2\right)_+,$$

we can reformulate the above bound as

$$\begin{aligned}
&\left(2+\frac{1}{1-K^{-1}}\right)^{-1}\left\|f_0-\widehat{f}_{\widehat{\lambda}}\right\|^2 \\
&\leq\ (1-K^{-1})\left\|f_0-\Pi_{S_*}\widehat{f}_{\widehat{\lambda}}\right\|^2+\frac{1}{2}\left\|\widehat{f}_{\widehat{\lambda}}-\Pi_{S_*}\widehat{f}_{\widehat{\lambda}}\right\|^2 \\
&\leq\ (1+K^{-1})\left\|f_0-\Pi_S\widehat{f}_{\lambda}\right\|^2+\frac{1}{2}\left\|\widehat{f}_{\lambda}-\Pi_S\widehat{f}_{\lambda}\right\|^2+2\mathrm{pen}(S)\widehat{\sigma}_S^2+\tilde{\Sigma}.
\end{aligned} \tag{D.1}$$

For any $S\in\widehat{\mathbb{S}}$ we have $\dim(S)\leq n/2-1$ and $\Delta(S)\leq 2n/3$. Therefore, according to Proposition 4 in [12] we have $\mathrm{pen}(S)\leq C[\dim(S)\vee\Delta(S)]$ and then

$$\begin{aligned}
\mathrm{pen}(S)\widehat{\sigma}_S^2 &= \frac{\mathrm{pen}(S)}{n-\dim(S)}\|Y-\Pi_S Y\|^2 \leq \frac{\mathrm{pen}(S)}{n-\dim(S)}\|Y-\Pi_S\widehat{f}_{\lambda}\|^2 \\
&\leq 3\frac{\mathrm{pen}(S)}{n-\dim(S)}\left(\|\varepsilon\|^2+\|f_0-\widehat{f}_{\lambda}\|^2+\|\widehat{f}_{\lambda}-\Pi_S\widehat{f}_{\lambda}\|^2\right) \\
&\leq C\left([\dim(S)\vee\Delta(S)]\sigma^2+\left(\|\varepsilon\|^2-2n\sigma^2\right)_+ +\|f_0-\widehat{f}_{\lambda}\|^2+\|\widehat{f}_{\lambda}-\Pi_S\widehat{f}_{\lambda}\|^2\right),
\end{aligned}$$

where C is a positive constant. Combining this bound with (D.1) and

$$(1+K^{-1})\left\|f_0-\Pi_S\widehat{f}_{\lambda}\right\|^2+\frac{1}{2}\left\|\widehat{f}_{\lambda}-\Pi_S\widehat{f}_{\lambda}\right\|^2\leq 4\left\|f_0-\widehat{f}_{\lambda}\right\|^2+5\left\|\widehat{f}_{\lambda}-\Pi_S\widehat{f}_{\lambda}\right\|^2$$

we finally obtain that for any $\lambda\in\Lambda$ and $S\in\widehat{\mathbb{S}}$

$$C^{-1}\left\|f_0-\widehat{f}_{\widehat{\lambda}}\right\|^2\leq\|f_0-\widehat{f}_{\lambda}\|^2+\|\widehat{f}_{\lambda}-\Pi_S\widehat{f}_{\lambda}\|^2+[\dim(S)\vee\Delta(S)]\sigma^2+\tilde{\Sigma}+\left(\|\varepsilon\|^2-2n\sigma^2\right)_+ \tag{D.2}$$

for some positive constant C depending on K only. Finally, choosing $K=1.1$, we deduce the upper bound

$$\mathbb{E}\left[\tilde{\Sigma}+\left(\|\varepsilon\|^2-2n\sigma^2\right)_+\right]\leq 2\Sigma+3\sigma^2,\qquad \text{(with } \Sigma \text{ defined in (C.1))}$$

from the definition of $\mathrm{pen}_\Delta(S)$ and the fact that $\left\|Y-\Pi_{\overline{S}}Y\right\|^2$ is independent of $\left\|\Pi_{\overline{S}}\varepsilon\right\|^2$ and is stochastically larger than $\left\|\varepsilon-\Pi_{\overline{S}}\varepsilon\right\|^2$. The bound (C.2) follows.

D.2 Proof of the second part of Proposition C.1

We use the same notation as in Section D.1. By (D.2), we have

$$\begin{aligned}
C^{-1}\left\|f_0-\widehat{f}_{\widehat{\lambda}}\right\|^2 \leq\ &\inf_{\lambda\in\Lambda}\left\{\|f_0-\widehat{f}_{\lambda}\|^2+\inf_{S\in\widehat{\mathbb{S}}}\left\{\|\widehat{f}_{\lambda}-\Pi_S\widehat{f}_{\lambda}\|^2+[\dim(S)\vee\Delta(S)]\sigma^2\right\}\right\} \\
&+\tilde{\Sigma}+\left(\|\varepsilon\|^2-2n\sigma^2\right)_+
\end{aligned}$$

for some positive constant C depending on K only. Setting $K=1.02$, we shall prove that with overwhelming probability $(\|\epsilon\|^2-2n\sigma^2)_+$ and

$$\tilde{\Sigma}:=2\sum_{S\in\mathbb{S}}\left(1.02\|\Pi_{\overline{S}}\epsilon\|^2-\frac{\mathrm{pen}(S)}{n-\dim(S)}\|Y-\Pi_{\overline{S}}(Y)\|^2\right)_+$$

are non positive. Applying a classical deviation inequality for χ^2 random variables (Lemma 1 in [49]), we derive that $\mathbb{P}\left[\|\epsilon\|^2 \geq 2n\sigma^2\right] \leq e^{-n/16}$. Let us turn to $\tilde{\Sigma}$. The random variable $(n-\dim(S)-1)\|\Pi_{\overline{S}}\epsilon\|^2/\|Y-\Pi_{\overline{S}}(Y)\|^2$ is stochastically smaller than a variable F_S such that $F_S/(\dim(S)+1)$ follows a Fisher distribution with $\dim(S)+1$ and $n-\dim(S)-1$ degrees of freedom. As a consequence, we have

$$\mathbb{P}\left[\tilde{\Sigma} > 0\right] \leq \sum_{S\in\mathbb{S}} \mathbb{P}\left[F_S \geq \frac{1.1}{1.02}\frac{n-\dim(S)-1}{n-\dim(S)}\mathrm{pen}_\Delta(S)\right]. \tag{D.3}$$

In order to upper bound the right hand-side of (D.3), we control the penalty terms $\mathrm{pen}_\Delta(S)$. We have

$$\mathbb{E}\left[\left(U - \frac{n-\dim(S)}{n-\dim(S)-1}\mathrm{pen}_\Delta(S)W\right)_+\right] = e^{-\Delta(S)} ,$$

where U and $(n-\dim(S)-1)W$ are two independent χ^2 random variables with respectively $\dim(S)+1$ and $n-\dim(S)-1$ degrees of freedom. We prove in the next sections the three following technical lemmas.

Lemma D.1. *Let $F = U/W$ and $0<\alpha<1$. We have*

$$\mathbb{P}\left(F \geq \frac{1}{1-\alpha}\frac{n-\dim(S)-1}{n-\dim(S)}\mathrm{pen}_\Delta(S)\right) \leq \frac{e^{-\Delta(S)}}{\alpha(\dim(S)+1)} .$$

Lemma D.2. *Assume that $\dim(S) \leq n/2-1$. For any $u>1$ and for any $x \geq 0$, we have*

$$\mathbb{P}\left(F \geq ux\right) \leq \exp\left[-\frac{u-1}{12u}\left\{(x-\dim(S)-1)\wedge n\right\}\right]\mathbb{P}\left(F \geq x\right) .$$

Lemma D.3. *For all $S \in \mathbb{S}$, we have*

$$\frac{n-\dim(S)-1}{n-\dim(S)}\mathrm{pen}_\Delta(S) \geq 2\Delta(S) + \dim(S) - C ,$$

where C is a positive constant.

We can now complete the proof of Proposition C.1. Applying Lemma D.1 with $1/(1-\alpha) = 1.1/1.05$ and Lemma D.2 with $u = 1.05/1.02$ and

$$x_S = \frac{1.1}{1.05} \times \frac{n-\dim(S)-1}{n-\dim(S)}\mathrm{pen}_\Delta(S) ,$$

we derive from (D.3) the following upper bound.

$$\begin{aligned}
\mathbb{P}\left[\tilde{\Sigma} > 0\right] &\leq \sum_{S\in\mathbb{S}} \exp\left[-C_2\left(\{x_S - \dim(S) - 1\}\wedge n\right)\right]\mathbb{P}\left[F_S \geq x_S\right] \\
&\leq \sum_{S\in\mathbb{S}} C_1 \exp\left[-C_2\left(\{x_S - \dim(S) - 1\}\wedge n\right)\right] e^{-\Delta(S)} \\
&\leq \sum_{S\in\mathbb{S}} C_1 \exp\left[-C_2\left(\Delta(S)\wedge n\right)\right] e^{-\Delta(S)}.
\end{aligned}$$

The proof of the second part of Proposition C.1 is complete.

D.3 Proof of the technical Lemmas D.1, D.2 and D.3

D.3.1 Proof of Lemma D.1

Since U is independent of W and $x \to (1-y/x)_+$ is increasing for all $y>0$ we have

$$\begin{aligned} e^{-\Delta(S)} &= \mathbb{E}\left[U\left(1-\frac{n-\dim(S)-1}{n-\dim(S)}\mathrm{pen}_\Delta(S)W/U\right)_+\right] \\ &\geq \mathbb{E}[U]\,\mathbb{E}\left[\left(1-\frac{n-\dim(S)-1}{n-\dim(S)}\mathrm{pen}_\Delta(S)/F\right)_+\right] \\ &\geq (\dim(S)+1)\times\alpha\,\mathbb{P}\left(1-\frac{n-\dim(S)-1}{n-\dim(S)}\mathrm{pen}_\Delta(S)/F\geq\alpha\right). \end{aligned}$$

D.3.2 Proof of Lemma D.2

Note that the bound is trivial if $x \leq \dim(S)+1$. In the sequel, we assume that $x \geq \dim(S)+1$. We set $d_1=\dim(S)+1$, $d_2=n-\dim(S)-1$ and write $B(.,.)$ for the Beta function. Since d_1F follows a Fisher distribution with (d_1,d_2) degrees of freedom, we have

$$\begin{aligned} \mathbb{P}(F\geq ux) &= \int_{ux}^{+\infty}\frac{t^{d_1/2}d_2^{d_2/2}}{(t+d_2)^{(d_1+d_2)/2}tB(d_1/2,d_2/2)}dt \\ &= \int_{x}^{+\infty}\frac{(ut)^{d_1/2}d_2^{d_2/2}}{(ut+d_2)^{(d_1+d_2)/2}tB(d_1/2,d_2/2)}dt \\ &\leq u^{d_1/2}\int_x^{+\infty}\left[\frac{t+d_2}{ut+d_2}\right]^{(d_1+d_2)/2}\frac{t^{d_1/2}d_2^{d_2/2}}{(t+d_2)^{(d_1+d_2)/2}tB(d_1/2,d_2/2)}dt \\ &\leq u^{d_1/2}\left[\frac{x+d_2}{ux+d_2}\right]^{(d_1+d_2)/2}\mathbb{P}(F\geq x) \\ &\leq \left\{u^{d_1/2}\left[\frac{d_1+d_2}{ud_1+d_2}\right]^{(d_1+d_2)/2}\right\}\left\{\left[\frac{(x+d_2)(ud_1+d_2)}{(ux+d_2)(d_1+d_2)}\right]^{(d_1+d_2)/2}\right\} \\ &\times \mathbb{P}(F\geq x)\ . \end{aligned}$$

In order to conclude, we shall prove that the first term between brackets is smaller than one and we shall control the second term. The derivative of the function

$$g: u\mapsto \log\left[u^{d_1/2}\left[\frac{d_1+d_2}{ud_1+d_2}\right]^{(d_1+d_2)/2}\right] \quad \text{is} \quad g'(u)=\frac{d_1}{2}\left[\frac{1}{u}-\frac{d_1+d_2}{ud_1+d_2}\right]\ ,$$

which is non positive for any $u\geq 1$. Since $g(1)=0$, we conclude that the first term is smaller than one. Let us turn to the logarithm of the second term:

$$\begin{aligned} -\frac{d_1+d_2}{2}\log\left[\frac{ux+d_2}{x+d_2}\frac{d_1+d_2}{ud_1+d_2}\right] &= -\frac{d_1+d_2}{2}\log\left[1+\frac{d_2(u-1)(x-d_1)}{(x+d_2)(ud_1+d_2)}\right] \\ &\leq -\frac{d_1+d_2}{2}\frac{d_2(u-1)(x-d_1)}{(x+d_2)(ud_1+d_2)+d_2(u-1)(x-d_1)} \\ &\leq -\frac{(u-1)}{2u}(x-d_1)\left[\frac{x}{d_2}+1+\frac{x-d_1}{d_2+d_1}\right]^{-1} \\ &\leq -\frac{(u-1)}{4u}\left[\frac{x-d_1}{2}\wedge\frac{n}{3}\right]\ , \end{aligned}$$

where the last line is proved by considering separately $x \leq d_1+d_2$ and $x > d_1+d_2$ and by using $d_1 \leq d_2 \leq n/2$.

D.3.3 Proof of Lemma D.3
We recall that the penalty $\mathrm{pen}_\Delta(S)$ is defined by

$$\mathbb{E}\left[\left(U - \frac{\mathrm{pen}_\Delta(S)}{n-\dim(S)}V\right)_+\right] = e^{-\Delta(S)}$$

where x_+ denotes the positive part of $x \in \mathbb{R}$ and U, V are two independent χ^2 random variables with respectively $\dim(S)+1$ and $n-\dim(S)-1$ degrees of freedom. Let us lower bound this expectation applying Jensen's inequality.

$$\begin{aligned}\mathbb{E}\left[\left(U - \frac{\mathrm{pen}_\Delta(S)}{n-\dim(S)}V\right)_+\right] &\geq \mathbb{E}\left[\left(U - \frac{n-\dim(S)-1}{n-\dim(S)}\mathrm{pen}_\Delta(S)\right)_+\right] \\ &\geq \mathbb{E}\left[\mathbf{1}\left\{U > \frac{n-\dim(S)-1}{n-\dim(S)}\mathrm{pen}_\Delta(S)+1\right\}\right] ,\end{aligned}$$

where $1\{A\}$ stands for the indicator function of the event A. Hence, we get

$$\mathrm{pen}_\Delta(S) \geq \frac{n-\dim(S)}{n-\dim(S)-1}\left[\bar{\chi}^{-1}_{dim(S)+1}\left[e^{-\Delta(S)}\right]-1\right] , \tag{D.4}$$

where $\bar{\chi}^{-1}_{\dim(S)+1}(\alpha)$ is a $1-\alpha$ quantile of a χ^2 random variable with $\dim(S)+1$ degrees of freedom.

Let us note $k = \dim(S)+1$. For any positive number x, we have

$$\begin{aligned}\mathbb{P}\left[U \geq x+k\right] &= \int_{x+k}^{+\infty}\frac{t^{k/2-1}e^{-t/2}}{2^{k/2}\Gamma(k/2)}dt = e^{-(x+k)/2}\int_0^{+\infty}\frac{(t+x+k)^{k/2-1}e^{-t/2}}{2^{k/2}\Gamma(k/2)}dt \\ &\geq e^{-(x+k)/2}\frac{k^{k/2-1}}{2^{k/2}\Gamma(k/2)}\int_0^{\sqrt{k}}\exp\left[-\frac{t}{2}+\left(\frac{k}{2}-1\right)\log\left(1+\frac{t}{k}\right)\right]dt \\ &\geq e^{-(x+k)/2}\frac{k^{k/2-1}}{2^{k/2}\Gamma(k/2)}\int_0^{\sqrt{k}}\exp\left[-\frac{t}{k}-\left(\frac{k}{2}-1\right)\frac{t^2}{2k^2}\right]dt ,\end{aligned}$$

since $\log(1+t) \geq t - t^2/2$. It follows that

$$\mathbb{P}\left[U \geq x+k\right] \geq e^{-(x+k)/2}\frac{k^{k/2-1}}{2^{k/2}\Gamma(k/2)}\int_0^{\sqrt{k}} e^{-1}e^{-t/(4\sqrt{k})}dt \geq Ce^{-(x+k)/2}\frac{k^{k/2-1/2}}{2^{k/2}\Gamma(k/2)} .$$

By Stirling's expansion $\Gamma(k/2) \leq (k/2)^{k/2-1/2}e^{-k/2}\sqrt{2\pi}$ so that $\mathbb{P}\left[U \geq x+k\right] \geq Ce^{-x/2}$. It follows that

$$\bar{\chi}^{-1}_{\dim(S)+1}\left(e^{-\Delta(S)}\right) \geq 2\Delta(S)+\dim(S)+1-C .$$

APPENDIX E: PROOF OF THE SPECIFIC BOUNDS FOR LASSO-LINSELECT AND GROUP-LASSO-LINSELECT

E.1 Size of the support of the Lasso and Group-Lasso estimators

For $\mathcal{K} \subset \{1,\ldots,M\}$, we recall that $\phi_{(\mathcal{K})}$ denotes the largest eigenvalue of $\mathbf{X}^T_{(\mathcal{K})}\mathbf{X}_{(\mathcal{K})}$.

LEMMA E.1. *Let $\widehat{\mathcal{K}}_\lambda$ be the subset of groups selected by the group-Lasso estimator $\widehat{\beta}_\lambda$. Then, on the event $\mathcal{A}_\lambda = \bigcap_{k=1}^M \left\{ \|\mathbf{X}_{G_k}^T \varepsilon\|_2 \le \lambda_k/4 \right\}$ we have*

$$\sum_{k \in \widehat{\mathcal{K}}_\lambda} \lambda_k^2 \le 16\, \phi_{(\widehat{\mathcal{K}}_\lambda)}\, \|\mathbf{X}\widehat{\beta}_\lambda - \mathbf{X}\beta_0\|_2^2 \,.$$

In particular, for the Lasso estimator $\widehat{\beta}_\lambda^L$, we have the upper bound

$$\lambda^2 \|\widehat{\beta}_\lambda^L\|_0 \le 16\, \phi_{\mathrm{supp}(\widehat{\beta}_\lambda^L)}\, \|\mathbf{X}\widehat{\beta}_\lambda^L - \mathbf{X}\beta_0\|_2^2$$

on the event $\mathcal{A}_\lambda = \left\{ |\mathbf{X}^T\varepsilon|_{\ell^\infty} \le \lambda/4 \right\}$.

The proof of this lemma is delayed to the Appendix E.4. The above bounds are similar to those stated in Bickel *et al.* [17] and Lounici *et al.* [54], except that it involves the restricted eigenvalue $\phi_{(\widehat{\mathcal{K}}_\lambda)}$ instead of the largest eigenvalue $\phi_{\max}$ of X^TX. When $|\widehat{\mathcal{K}}_\lambda|$ is small compared to n the restricted eigenvalue $\phi_{(\widehat{\mathcal{K}}_\lambda)}$ can be much smaller than $\phi_{\max}$. Actually, since X^TX has at most n non-zero eigenvalues and $\mathrm{Tr}(X^TX) = p$, we always have $\phi_{\max} \ge p/n$ which can be large when $p \gg n$.

E.2 Proof of Proposition 4.3

The first step is to provide a sufficient condition for having $\|\widehat{\beta}_\lambda\|_0 \le n/(3\log(p))$. Recall that the compatibility constant $\kappa[\xi, T]$ is defined in Section 4.3.

LEMMA E.2. *Assume that $\lambda \ge 8\sigma\sqrt{\log(p)}$ and*

$$1 \le \|\beta_0\|_0 \le \frac{\kappa^2[5, \mathrm{supp}(\beta_0)]}{96\,\phi_*} \times \frac{n}{\log(p)} \,. \tag{E.1}$$

Then, on the event $\mathcal{A} = \left\{ |\mathbf{X}^T\varepsilon|_{\ell^\infty} \le 2\sigma\sqrt{\log(p)} \right\}$ we have $\|\widehat{\beta}_\lambda\|_0 \le n/(3\log(p))$.

PROOF OF LEMMA E.2. We write $\widehat{\mathcal{J}}$ for the support of $\widehat{\beta}_\lambda$. A slight variation of Theorem 14 in [48] ensures that

$$\|\mathbf{X}\widehat{\beta}_\lambda - \mathbf{X}\beta_0\|_2^2 \le \inf_{\beta \ne 0} \left\{ \|\mathbf{X}\beta_0 - \mathbf{X}\beta\|_2^2 + \frac{\lambda^2}{\kappa^2[5, \mathrm{supp}(\beta)]} \|\beta\|_0 \right\} \tag{E.2}$$

on the event $\mathcal{A}$. Combining Lemma E.1 with the bound (E.2) we obtain that

$$\mathrm{Card}(\widehat{\mathcal{J}}) \le 16\, \phi_{\widehat{\mathcal{J}}}\, \frac{\|\beta_0\|_0}{\kappa^2[5, \mathrm{supp}(\beta_0)]}.$$

Let us set $d^* = n/[3\log(p)]$. The upper-bound $\phi_{\widehat{\mathcal{J}}} \le (1 + \mathrm{Card}(\widehat{\mathcal{J}})/d^*)\phi_*$ enforces

$$\mathrm{Card}(\widehat{\mathcal{J}}) \le \frac{16\phi_* \|\beta_0\|_0}{\kappa^2[5, \mathrm{supp}(\beta_0)]} \left[1 + \frac{\mathrm{Card}(\widehat{\mathcal{J}})}{d^*} \right] \le \left(d^* + \mathrm{Card}(\widehat{\mathcal{J}}) \right)/2 \,,$$

where the last inequality follows from (E.1). □

We can now complete the proof of Proposition 4.3. We recall that the event $\mathcal{A} = \left\{ |\mathbf{X}^T \varepsilon|_{\ell^\infty} \leq 2\sigma\sqrt{\log(p)} \right\}$ has probability at least $1 - 1/p$. Let us set

$$\lambda_0 = \sqrt{16(4 \vee \phi_*)\log(p)\sigma^2} \geq 8\sigma\sqrt{\log(p)}.$$

Under the hypothesis (E.1), the combination of Lemma E.2 with Proposition C.1 ensures that with probability larger than $1 - C_1 p^{-C_2}$ we have

$$\left\| \mathbf{X}\beta_0 - \mathbf{X}\widehat{\beta}_{\widehat{\lambda}} \right\|_2^2 \quad \leq \quad C\left\{ \|\mathbf{X}\beta_0 - \mathbf{X}\widehat{\beta}_{\lambda_0}\|_2^2 + [\|\widehat{\beta}_{\lambda_0}\|_0 \vee 1]\log(p)\sigma^2 \right\}.$$

We upper bound the right-hand side by combining Lemma E.1 with (E.2)

$$\begin{aligned}
\left\| \mathbf{X}\beta_0 - \mathbf{X}\widehat{\beta}_{\widehat{\lambda}} \right\|_2^2 \quad &\leq \quad C\left(1 + \frac{16\phi_{\widehat{\mathcal{J}}}\log(p)\sigma^2}{\lambda_0^2}\right) \\
&\qquad \times \inf_{\beta \neq 0}\left\{ \|\mathbf{X}\beta_0 - X\beta\|_2^2 + \frac{\lambda_0^2}{\kappa^2[5, \mathrm{supp}(\beta)]}\|\beta\|_0 \right\} \\
&\leq \quad C' \inf_{\beta \neq 0}\left\{ \|\mathbf{X}\beta_0 - X\beta\|_2^2 + \frac{\phi_*\log(p)\sigma^2}{\kappa^2[5, \mathrm{supp}(\beta)]}\|\beta\|_0 \right\},
\end{aligned}$$

where we used in the last inequality that $\widehat{\mathcal{J}}$ (the support of $\widehat{\beta}_{\lambda_0}$) is of size at most $n/(3\log(p))$.

E.3 Proof of Proposition 5.1

The proof of Proposition 5.1 is very similar to that of Proposition 4.3. We only sketch the main lines. The first step is to provide a sufficient condition for having $|\widehat{\mathcal{K}}_\lambda| \leq (n-2)/(2T \vee 3\log(M))$. Recall that the compatibility constant $\kappa_G[\xi, s]$ is defined in (20) and ϕ_* in (19).

Lemma E.3. *Assume that*

$$\lambda_k^2 \quad = \quad 96\phi_*(T \vee 3\log(M))\sigma^2, \quad \text{for } k = 1, \ldots, M \tag{E.3}$$

$$\text{and} \quad 1 \quad \leq \quad |\mathcal{K}_0| \quad \leq \quad \frac{\kappa_G^2[3, |\mathcal{K}_0|]}{2^9\phi_*} \times \frac{n-2}{2T \vee 3\log(M)}. \tag{E.4}$$

Then we have $|\widehat{\mathcal{K}}_\lambda| \leq (n-2)/(3\log(M) \vee 2T)$, with probability at least $1 - 3/M$.

Proof of Lemma E.3. We set $k^* = (n-2)/(3\log(M) \vee 2T)$. Theorem 3.1 in [54] gives

$$\|\mathbf{X}\widehat{\beta}_\lambda - \mathbf{X}\beta_0\|_2^2 \leq \frac{16}{\kappa_G^2[3, |\mathcal{K}_0|]}\,|\mathcal{K}_0|\,\lambda_1^2\,, \tag{E.5}$$

with probability larger than $1 - 3/M$. Combining this bound with Lemma E.1 and the bound $\phi_{(\widehat{\mathcal{K}}_\lambda)} \leq [1 + |\widehat{\mathcal{K}}_\lambda|/k_*]\phi_*$, we get that with probability larger than $1 - 3/M$

$$\begin{aligned}
|\widehat{\mathcal{K}}_\lambda| \quad &\leq \quad \frac{2^8}{\kappa_G^2[3, |\mathcal{K}_0|]}\phi_{(\widehat{\mathcal{K}}_\lambda)}|\mathcal{K}_0| \\
&\leq \quad \frac{2^8}{\kappa_G^2[3, |\mathcal{K}_0|]}\left[1 + \frac{|\widehat{\mathcal{K}}_\lambda|}{k_*}\right]\phi_*|\mathcal{K}_0| \quad \leq \quad (k^* + |\widehat{\mathcal{K}}_\lambda|)/2\,,
\end{aligned}$$

where the last bound follows from (E.4). □

We complete now the proof of Proposition 5.1. Assume that (E.3) and (E.4) are satisfied. Combining Lemma E.3 with Proposition C.1 ensures that with probability larger than $1 - C_1 M^{-C_2} - 3/M$ we have

$$\begin{aligned} C^{-1} \left\| \mathbf{X}\beta_0 - \mathbf{X}\widehat{\beta}_{\widehat{\lambda}} \right\|_2^2 &\leq \|\mathbf{X}\widehat{\beta}_\lambda - \mathbf{X}\beta_0\|_2^2 + (1 \vee |\widehat{\mathcal{K}}_\lambda|)\,(T \vee \log(M))\,\sigma^2 \\ &\leq 2\left(\|\mathbf{X}\widehat{\beta}_\lambda - \mathbf{X}\beta_0\|_2^2 \vee \left[(T \vee \log(M))\,\sigma^2 \right] \right). \end{aligned}$$

Proposition 5.1 then simply follows from (E.5).

E.4 Proof of Lemma E.1

We write $\widehat{\beta}$ for $\widehat{\beta}_\lambda$, $\widehat{\mathcal{K}}$ for $\widehat{\mathcal{K}}_\lambda$ and A^+ for the Moore-Penrose pseudo-inverse of A. The optimality condition gives

$$2\mathbf{X}_{(\widehat{\mathcal{K}})}^T (Y - \mathbf{X}_{(\widehat{\mathcal{K}})}\widehat{\beta}_{(\widehat{\mathcal{K}})}) = \lambda z_{(\widehat{\mathcal{K}})} \tag{E.6}$$

where $\|z^{G_k}\|_2 = 1$ for all $k \in \widehat{\mathcal{K}}$. As a consequence we have

$$\widehat{\beta}_{(\widehat{\mathcal{K}})} = (\mathbf{X}_{(\widehat{\mathcal{K}})}^T \mathbf{X}_{(\widehat{\mathcal{K}})})^+ (\mathbf{X}_{(\widehat{\mathcal{K}})}^T Y - \lambda z_{(\widehat{\mathcal{K}})}/2)$$

and

$$\mathbf{X}\widehat{\beta} = P_{(\widehat{\mathcal{K}})} Y - \lambda \mathbf{X}_{(\widehat{\mathcal{K}})} (\mathbf{X}_{(\widehat{\mathcal{K}})}^T \mathbf{X}_{(\widehat{\mathcal{K}})})^+ z_{(\widehat{\mathcal{K}})}/2$$

where $P_{(\widehat{\mathcal{K}})}$ is the orthogonal projector onto the range of $\mathbf{X}_{(\widehat{\mathcal{K}})}$. Pythagorean equality gives

$$\begin{aligned} \|\mathbf{X}\beta_0 - \mathbf{X}\widehat{\beta}\|_2^2 &= \|\mathbf{X}\beta_0 - P_{(\widehat{\mathcal{K}})}\mathbf{X}\beta_0\|_2^2 + \|P_{(\widehat{\mathcal{K}})}\varepsilon - \lambda \mathbf{X}_{(\widehat{\mathcal{K}})} (\mathbf{X}_{(\widehat{\mathcal{K}})}^T \mathbf{X}_{(\widehat{\mathcal{K}})})^+ z_{(\widehat{\mathcal{K}})}/2\|_2^2 \\ &\geq \|\mathbf{X}_{(\widehat{\mathcal{K}})} (\mathbf{X}_{(\widehat{\mathcal{K}})}^T \mathbf{X}_{(\widehat{\mathcal{K}})})^+ (\mathbf{X}_{(\widehat{\mathcal{K}})}^T \varepsilon - \lambda z_{(\widehat{\mathcal{K}})}/2)\|_2^2. \end{aligned}$$

From (E.6) we know that the vector $\mathbf{X}_{(\widehat{\mathcal{K}})}^T \varepsilon - \lambda z_{(\widehat{\mathcal{K}})}/2$ belongs to the range of $\mathbf{X}_{(\widehat{\mathcal{K}})}^T$ and therefore (see Lemma E.4 below)

$$\phi_{(\widehat{\mathcal{K}})} \|\mathbf{X}_{(\widehat{\mathcal{K}})} (\mathbf{X}_{(\widehat{\mathcal{K}})}^T \mathbf{X}_{(\widehat{\mathcal{K}})})^+ (\mathbf{X}_{(\widehat{\mathcal{K}})}^T \varepsilon - \lambda z_{(\widehat{\mathcal{K}})}/2)\|_2^2 \geq \|\mathbf{X}_{(\widehat{\mathcal{K}})}^T \varepsilon - \lambda z_{(\widehat{\mathcal{K}})}/2\|_2^2.$$

Finally, on the event $\mathcal{A}_\lambda$ we have $\|\mathbf{X}_{G_k}^T \varepsilon - \lambda_k z^{G_k}/2\|_2 \geq \lambda_k/4$ for all $k \in \widehat{\mathcal{K}}$, so

$$\|\mathbf{X}_{(\widehat{\mathcal{K}})}^T \varepsilon - \lambda z_{(\widehat{\mathcal{K}})}/2\|_2^2 \geq \sum_{k \in \widehat{\mathcal{K}}} \lambda_k^2/16.$$

This allows to conclude.

LEMMA E.4. *Let A be any $n \times d$ real matrix. Then for any x in the range of A^T we have*

$$\|x\|_2^2 \leq \varphi_{\max}(A^T A)\, \|A(A^T A)^+ x\|_2^2$$

where $\varphi_{\max}(A^T A)$ denotes the largest eigenvalue of $A^T A$.

PROOF OF LEMMA E.4. We first note that

$$\|A(A^TA)^+x\|_2^2 = x^T(A^TA)^+A^TA(A^TA)^+x = x^T(A^TA)^+x.$$

Furthermore the range of A^T coincides with the range of A^TA, which in turn is the same as the range of $(A^TA)^+$. We then have

$$\sigma_{\mathrm{rank}((A^TA)^+)}((A^TA)^+)\|x\|_2^2 \leq x^T(A^TA)^+x$$

where $\sigma_k((A^TA)^+)$ is the k-th largest singular value of $(A^TA)^+$. The result follows from the equality

$$\left[\sigma_{\mathrm{rank}((A^TA)^+)}((A^TA)^+)\right]^{-1} = \sigma_1(A^TA) = \varphi_{\max}(A^TA).$$

□