High-dimensional regression with unknown variance

Christophe Giraud, Sylvie Huet, Nicolas Verzelen

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 emphasize 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 [69], 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 [20, 23, 53, 62, 63, 70, 72, 74] rely on tuning parameters, whose optimal value depends on σ. For example, Bickel et al. [15] state that under some assumptions on $\mathbf{X}$, the tuning parameter λ of the Lasso should be chosen of the order of $\sigma\sqrt{2\log(p)}$. 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 [14, 58, 61] 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 with 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 [20, 23, 53, 62, 63, 70, 72, 74]. Among the most famous methods for estimator selection, we mention V-fold cross-validation (Geisser [29]), AIC (Akaike [1]) and BIC (Schwartz [54]) 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 [9] 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 [32, 30]. For these reasons, we will not review these approaches in the sequel.

1.4 Why care on non-asymptotic analyses ?

AIC [1], BIC [54] and V-fold Cross-Validation [29] 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 [57], Li [44] and Shao [56] 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 [50] shows that the BIC criterion is consistent for model selection.

All these asymptotic results can lead to misleading conclusion in modern statistical settings where the sample size remains small and the parameters dimension becomes large. For instance it is proved in [10, Sect.3.3.2] and illustrated in [10, 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 the appendix. 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, 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 $\sup_{\beta_0 \in \boldsymbol{B}[k,p]} \mathcal{R}[\widehat{\beta}; \beta_0]/\sigma^2$ 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*

$$\mathbf{R}[k, \mathbf{X}] = \inf_{\widehat{\beta}} \sup_{\beta_0 \in \boldsymbol{B}[k,p]} \frac{\mathcal{R}[\widehat{\beta}; \beta_0]}{\sigma^2} \,. \tag{5}$$

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. For some statistical problems, adaptation is in fact impossible and there is an unavoidable loss when the variance or the sparsity parameter is unknown. In such situations, it is interesting to quantify this loss.

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.

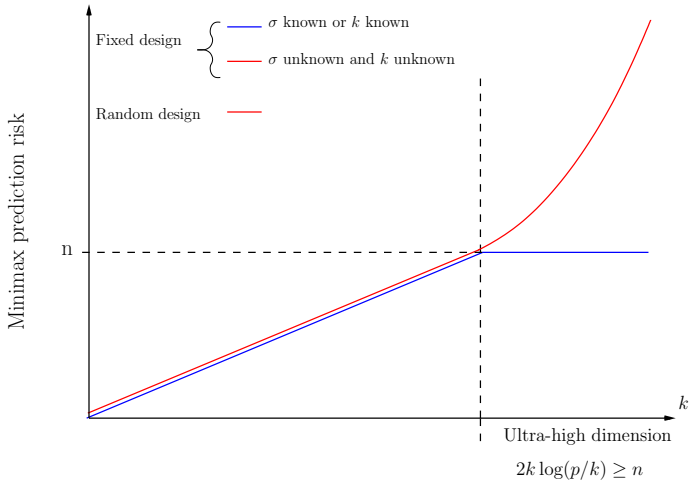

FIGURE 1. *Minimal prediction risk over $\boldsymbol{B}[k,p]$ as a function of k.*

2.2 Differences between known and unknown variance

Known variance. When the variance σ^2 is known we have the following bounds.

PROPOSITION 2.1. *For any (k,n,p) such that $k \leq n/2$, we have*

$$C_1 k\log\left(\frac{p}{k}\right) \wedge n \leq \sup_{\mathbf{X}} \mathbf{R}[k,\mathbf{X}] \leq C_1' k\log\left(\frac{p}{k}\right) \wedge n \ . \tag{6}$$

In addition, there exist procedures that are adaptive to $k \in \{1,\dots,\lfloor n/2\rfloor\}$.

This bound has been proved in [66] (see also [52, 53, 68]). Let us first comment this bound when (k,p) is not too large, say when $2k\log(p/k) < n$. 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 Proposition 2.1, the minimax risk is of the order $k\log(p/k)$. 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. In a ultra-high dimensional setting (when $2k\log(p/k) \geq n$), the situation is quite different. Indeed, the minimax risk remains of the order of n and the sparsity index k does not play a role anymore.

Unknown variance. When the sparsity index k is known, the minimax prediction risks are of the same order for known and unknown variance. Furthermore, adaptation to the sparsity *and* to the variance σ^2 is possible as long as the sparsity

[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 $\delta > 0$.

index k satisfies $2k\log(p/k) \leq n$, see [66]. The situation is different in a ultra-high dimensional setting as shown in the next proposition (from [66]).

PROPOSITION 2.2. *Consider any $p \geq n \geq C$ and $k \leq p^{1/3} \wedge n/2$ such that $k\log(p/k) \geq C'n$. There exists a design $\mathbf{X}$ of size $n \times p$ such that for any estimator $\widehat{\beta}$, we have either*

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

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

As a consequence, 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

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

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). We conclude that adaptation to the sparsity is impossible when the variance is unknown.

2.3 Aspects of ultra-high dimensionality

The previous results have illustrated the existence of a phase transition when k and p are very large ($2k\log(p/k) \geq n$) : the prediction problem with unknown variance and unknown sparsity becomes extremely difficult in ultra-high dimensional settings. Similar phenomenons occur for some 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, etc [66, 67, 41]. This kind of phase transition has been observed in a wide range of random geometry problems [26], suggesting some universality of this limitation.

Finally, where lie the limits of accurate high-dimensional sparse estimation? In practice, the sparsity index k is not known, but given (n, p) we can compute $k^* := \max\{k : \ 2k\log(p/k) \geq n\}$. As a rule of thumb, 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 [66]).

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. In the sequel, we will say that $\widehat{\beta}$ achieves an *optimal* risk bound (with respect to the sparsity) if

$$\mathcal{R}[\widehat{\beta}; \beta_0] \leq C_1 \|\beta_0\|_0 \log(p)\sigma^2 \ , \tag{7}$$

for any $\sigma > 0$ and any vector $\beta_0 \in \mathbb{R}^p$ such that $\|\beta_0\|_0 \log(p) \leq C_2 n$. Such risk bounds prove that the estimator is approximately (up to a possible $\log(k)$ factor) minimax adaptive to the unknown variance and the unknown sparsity in non ultra-high dimensional setting. Fast procedures such as those presented in Section 4.2 achieve this kind of bounds under some restrictive assumptions on the design matrix $\mathbf{X}$.

For some procedure, (7) can be improved into a bound of the form

$$\mathcal{R}[\widehat{\beta};\beta_0] \leq C_1 \inf_{k \leq Cn/\log p} \left\{ \inf_{\beta,\ \|\beta\|_0 = k} \|\mathbf{X}(\beta - \beta_0)\|_2^2 + k \log(p) \sigma^2 \right\} . \tag{8}$$

This kind of bound makes clear a 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 (8) can be much smaller than (7).

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* [49, 24] which is based on a single split of the data for *training* and *validation*.
- *V-fold* CV [29]. The data is split into V subsamples. Each subsample is successively removed for *validation*, the remaining data being used for *training*.
- *Leave-one-out* [59] which corresponds to n-fold CV.
- *Leave-p-out* (also called *delete-p-CV*) [55] where every possible subset of cardinality p 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

$$\mathrm{Crit}(\lambda) = \mathcal{L}_{\mathbf{X}}(Y, \widehat{\beta}_\lambda) + \mathrm{pen}(\lambda), \tag{9}$$

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 (9) 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.* [10] 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 in terms of the risk $\mathcal{R}[\widehat{\beta}_{\lambda^*}; \beta_0]$ of the oracle (4), see Theorem 2 in [10].

Plug-in criteria. Many other penalized-empirical-loss criteria have been developed in the last decades. Several selection criteria [12, 16] 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 $\sigma^2/(n - \mathrm{rank}(\mathbf{X}))$ 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.* [10] 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 [27]. The second rigidity of the procedure of Baraud *et al.* [10] 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.* [11]. 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. For a suitable, possibly random, subset $\mathbb{S}_\lambda \subset \mathbb{S}$ (depending on the statistical problem), the parameter $\widehat{\lambda}$ is selected by minimizing the criterion

$$\text{Crit}(\lambda) = \inf_{S \in \mathbb{S}_\lambda} \left[\|Y - \Pi_S \mathbf{X}\widehat{\beta}_\lambda\|_2^2 + \frac{1}{2}\|\mathbf{X}\widehat{\beta}_\lambda - \Pi_S \mathbf{X}\widehat{\beta}_\lambda\|_2^2 + \text{pen}_\Delta(S)\,\widehat{\sigma}_S^2 \right], \tag{10}$$

where Π_S is the orthogonal projector onto S, pen_Δ is a penalty depending on Δ and

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

We refer to Section 2.1 in [11] for more details on the procedure, Theorem 1 in [11] for risk bounds and Sections 4 and 5 for examples. The algorithmic complexity of LinSelect is proportional to $\sum_{\lambda\in\Lambda} |\mathbb{S}_\lambda|$ and the whole procedure is less intensive than V-fold CV for suitable choices of $\{\mathbb{S}_\lambda,\ \lambda \in \Lambda\}$. Finally, we mention that for the constrained least-square estimators $\mathbf{X}\widehat{\beta}_\lambda = \Pi_{S_\lambda} Y$, the LinSelect procedure with $\mathbb{S}_\lambda = \{S_\lambda\}$ simply coincides with the procedure of Baraud *et al.* [10].

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 [62, 22], Dantzig Selector [20], Elastic Net [74], MC+ [70], aggregation techniques [19, 23], 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. In the specific case of the relaxed-Lasso, Meinshausen [47] ensures that V-fold CV selects a parameter λ that performs almost as well as the oracle (4). 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 [37, 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 suitable collection $\mathbb{S}$

and measure of complexity Δ are

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

Let us introduce the spaces $\widehat{S}_\lambda = \text{range}\left(\mathbf{X}_{\text{supp}(\widehat{\beta}_\lambda)}\right)$ 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 $\mathbb{S}_\lambda = \widehat{\mathbb{S}}$ for all $\lambda \in \Lambda$ and the above choice of Δ.

PROPOSITION 4.1 (Theorem 1 in [11]). *There exists a constant $C > 1$ such that for any minimizer $\widehat{\lambda}$ of the Criterion (10), we have*

$$\mathcal{R}\left[\widehat{\beta}_{\widehat{\lambda}}; \beta_0\right] \leq C\ \mathbb{E}\left[\inf_{\lambda \in \widehat{\Lambda}} \left\{ \|\mathbf{X}\widehat{\beta}_\lambda - \mathbf{X}\beta_0\|_2^2 + \|\widehat{\beta}_\lambda\|_0 \log(p)\sigma^2 \right\}\right]. \tag{11}$$

The bound (11) cannot be formulated in the form (8) due to the random nature of the set $\widehat{\Lambda}$. Nevertheless, a bound similar to (7) can be deduced from (11) when the estimators $\widehat{\beta}_\lambda$ fulfill $\mathbf{X}\widehat{\beta}_\lambda = \Pi_{\widehat{S}_\lambda} Y$, see Corollary 4 in [11].

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 $\widehat{\beta}^L_\lambda := \text{argmin}_{\beta \in \mathbb{R}^p} \|Y - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1$. 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 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. [58] propose to minimize the ℓ_1-penalized log-likelihood criterion

$$\widehat{\beta}^{LL}_\lambda, \widehat{\sigma}^{LL}_\lambda := \underset{\beta \in \mathbb{R}^p, \sigma' > 0}{\text{argmin}} \left[n\log(\sigma') + \frac{\|Y - \mathbf{X}\beta\|_2^2}{2\sigma'^2} + \lambda \frac{\|\beta\|_1}{\sigma'} \right]. \tag{12}$$

By reparametrizing (β, σ), Städler et al. [58] 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 [60] show that the choice $\lambda = C\sqrt{2\log p}$, with $C > 1$ yields optimal risk bounds in the sense of (7).

Scaled Lasso. Alternatively, Antoniadis [3] suggests to minimize a penalized Huber's loss [39, page 179]

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

This convex criterion can be minimized with roughly the same computational complexity as a Lars-Lasso path [27]. Furthermore, Sun and Zhang [61] state sharp oracle inequalities for this estimator with $\lambda = C\sqrt{2\log(p)}$, with $C > 1$. Their empirical results suggest that the criterion (13) provides slightly better results than the ℓ^1-penalized log-likelihood.

Square-root Lasso. Belloni et al. [14] propose to replace the residual sum of squares in the Lasso criterion by its square-root

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

Interestingly, the penalty level λ is again pivotal. For $\lambda = C\sqrt{2\log(p)}$ with $C > 1$, the square-root Lasso estimator also achieves optimal risk bounds under suitable assumptions on the design $\mathbf{X}$.

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 [51]. Bayesian procedures straightforwardly handle the case of unknown variance, but no frequentist analysis of these procedures are yet available.

4.3 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 scaled Lasso [61] and the square root Lasso [14] are 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 developped 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 [43, 48]. 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 $\widehat{\beta}^{L}_{\lambda}$

$$(14) \qquad \widehat{\beta}^{GL}_{\lambda} := \underset{\beta\in\mathbb{R}^p:\operatorname{supp}(\beta)\subset\operatorname{supp}(\widehat{\beta}^{L}_{\lambda})}{\operatorname{argmin}} \|Y-\mathbf{X}\beta\|_2^2 .$$

When the objective is support estimation, we advise to apply LinSelect or the V-fold CV scheme to the Gauss-Lasso estimators instead of the Lasso estimators. Similar remarks also apply for the Dantzig Selector [20].

4.4 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 [11], except that we restrict to the case $n = p = 100$. Therefore, 165 examples are simulated. They are inspired by examples found in [62, 74, 73, 38] 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 [29], LinSelect [11] and the scaled-Lasso [61] (with $\lambda = \sqrt{2\log(p)}$) for tuning the Lasso. For each tuning procedure $\ell \in \{\text{10-fold CV}, \text{LinSelect}, \text{scaled-Lasso}\}$, 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 V-fold CV	1.19	0.09	1.07	1.16	1.27	1.33	1.36
Lasso LinSelect	1.19	0.58	1.02	1.05	1.09	1.11	2.24
Scaled Lasso	5.77	7.35	1.30	2.75	5.85	13	21

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 scaled-lasso procedure, the risk ratios are clearly larger than for the two others. An inspection of the results reveals that the scaled-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 scaled-lasso to choose a large λ. Consequently the risk is high. We illustrate this phenomenon in Table 2, where we consider apart *easy examples* with ratio $\mathcal{R}\left[\widehat{\beta}_{\lambda^*}; \beta_0\right]/n\sigma^2$ smaller than

0.1 and *difficult examples* with ratio $\mathcal{R}\left[\widehat{\beta}_{\lambda^*};\beta_0\right]/n\sigma^2$ greater than 0.8. The bias of the Lasso estimator is small for easy examples, whereas it is large for difficult examples. The difference of behavior between the scaled-Lasso and the two other procedures is significant.

	easy examples		*difficult examples*	
	mean risk ratio	$\|\widehat{\beta}^L_{\hat{\lambda}_\ell}\|_0$	mean risk ratio	$\|\widehat{\beta}^L_{\hat{\lambda}_\ell}\|_0$
Oracle	1	(11)	1	(43)
Lasso V-fold CV	1.35	(13)	1.12	(45)
Lasso LinSelect	1.06	(11)	1.71	(40)
Scaled Lasso	1.89	(8)	15	(3.5)

TABLE 2

Mean risk ratio $\mathcal{R}_{\ell,e}/\mathcal{R}_e$ for each selection procedure ℓ on easy *and* difficult *examples. In parentheses : $\|\widehat{\beta}^L_{\hat{\lambda}_\ell}\|_0$. The first line gives the same values for the oracle $\widehat{\beta}^L_{\lambda^\star}$*

Experiment 2 : variable selection with Gauss-Lasso and scaled-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 scaled-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 Figure 2. It appears that the Gauss-Lasso CV procedure gives greater values of the FDR than the two others. The Gauss-Lasso LinSelect and the scaled-Lasso behave similarly for the FDR, but the values of the power are more variable for the LinSelect procedure.

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 Bach [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

$$\|Y-\mathbf{X}\beta\|_2^2+\sum_{k=1}^{M}\lambda_k\|\beta^{G_k}\|_2, \tag{15}$$

where $\beta^{G_k}=(\beta_j)_{j\in G_k}$. The Criterion (15) promotes solutions where all the coordinates of β^{G_k} are either zero or non-zero, leading to group selection [69]. Under some assumptions on $\mathbf{X}$, Lounici *et al.* [45] provides 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 proportionnal 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

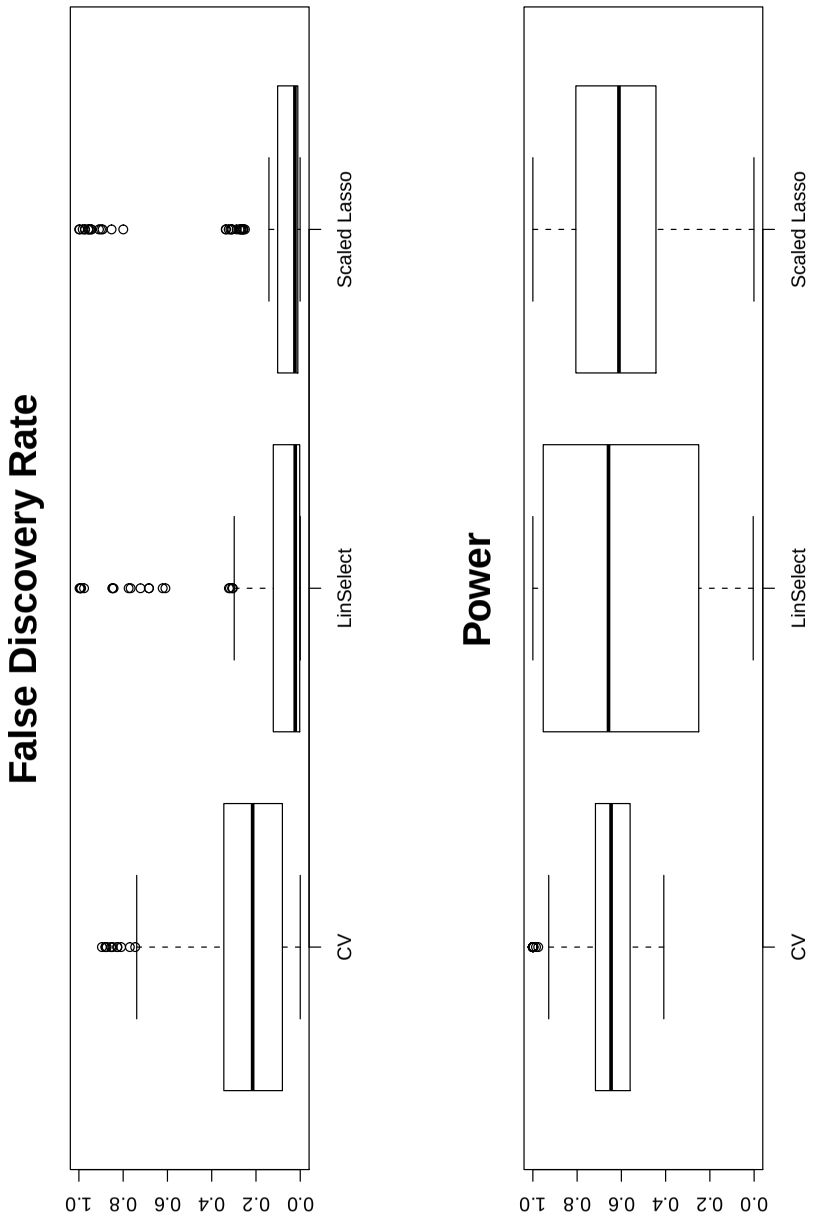

FIGURE 2. *For each procedure boxplots of FDR and Power values*

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 $\bigcup_{k\in\mathcal{K}} G_k$. The collection $\mathbb{S}$ and the function Δ are given by

$$\mathbb{S} = \left\{ \text{range}(\mathbf{X}_{(\mathcal{K})}) \ : \ \mathcal{K} \subset \{1, \dots, M\} \text{ and } 2|\mathcal{K}| \log(M) \vee \sum_{k\in\mathcal{K}} |G_k| \leq n/2 \right\}$$

and $\Delta(\text{range}(\mathbf{X}_{(\mathcal{K})})) = |\mathcal{K}| \log(M)$. 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 $\mathbb{S}_\lambda = \widehat{\mathbb{S}}$ for all $\lambda \in \Lambda$, where

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

Theorem 1 in [11] provides the following risk bound.

PROPOSITION 5.1. *Let $\widehat{\lambda}$ be any minimizer of the Criterion (10) with the above choice of $\mathbb{S}_\lambda$ and Δ. Then, there exists a constant $C > 1$ such that*

$$\mathcal{R}\left[\widehat{\beta}_{\widehat{\lambda}}; \beta_0\right] \leq C \, \mathbb{E}\left[\inf_{\lambda\in\widehat{\Lambda}} \left\{ \|\mathbf{X}\widehat{\beta}_\lambda - \mathbf{X}\beta_0\|_2^2 + \left(\|\widehat{\beta}_\lambda\|_0 \vee |\widehat{\mathcal{K}}_\lambda| \log(M) \right) \sigma^2 \right\} \right]. \tag{16}$$

In the case where each group G_k is a singleton, we have $M = p$ and we recover the result of Proposition 4.1 when we assume that $\lambda_1 = \lambda_2 = \ldots = \lambda_M$. When the cardinality of each G_k is larger than $\log(M)$, we have $\|\widehat{\beta}_\lambda\|_0 \geq |\widehat{\mathcal{K}}_\lambda| \log(M)$ with probability one, so the final estimator nearly achieves the best trade off between $\|\mathbf{X}\widehat{\beta}_\lambda - \mathbf{X}\beta_0\|_2^2$ and $\|\widehat{\beta}_\lambda\|_0 \sigma^2$ for $\lambda \in \widehat{\Lambda}$.

Let us compare Proposition 5.1 with the bounds of Lounici *et al.* [45] in the specific case of multitask learning with M tasks that is $n = Mn_0$ and $p = Mp_0$. Suppose that only m groups out of M correspond to non-zero vector $\beta_0^{G_k}$. Under suitable assumptions on the design $\mathbf{X}$, it is proved in [45] that the group Lasso estimator $\widehat{\beta}_\lambda$ with a well-chosen tuning parameter $\lambda_*(\sigma)$ satisfies

$$\|\mathbf{X}\widehat{\beta}_{\lambda_*(\sigma)} - \mathbf{X}\beta_0\|_2^2 \leq C \left[mp_0 \vee k \log(M) \right] \sigma^2 \ ,$$

and that $\|\widehat{\beta}_{\lambda_*(\sigma)}\|_0 \leq Cmp_0$ with large probability. It follows that with large probability

$$\inf_{\lambda\in\widehat{\Lambda}} \left\{ \|\mathbf{X}\widehat{\beta}_\lambda - \mathbf{X}\beta_0\|_2^2 + \left(\|\widehat{\beta}_\lambda\|_0 \vee |\widehat{\mathcal{K}}_\lambda| \log(M) \right) \sigma^2 \right\} \leq C \left[mp_0 \vee m \log(M) \right] \sigma^2 \ ,$$

suggesting that $\widehat{\beta}_{\widehat{\lambda}}$ satisfies the same kind of bounds as $\widehat{\beta}_{\lambda_*(\sigma)}$ without requiring the knowledge of σ.

6. VARIATION-SPARSITY

We focus in this section to 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 [35]. 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

$$\mathrm{Crit}(q) = \|Y - \widehat{\beta}_{\widehat{\mathcal{I}}_q}\|_2^2 + \mathrm{pen}(q). \tag{17}$$

Following the work of Birgé and Massart [16], Lebarbier [42] 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. With this choice of the penalty, the procedure satisfies a bound of the form

$$\mathcal{R}\left[\widehat{\beta}_{\widehat{\mathcal{I}}}, \beta_0\right] \;\le\; C \inf_{\mathcal{I} \subset \{1,\dots,n-1\}} \left\{\|\widehat{\beta}_{\mathcal{I}} - \beta_0\|_2^2 + (1 + |\mathcal{I}|)\log(n/(1+|\mathcal{I}|))\,\sigma^2\right\}. \tag{18}$$

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. [36]. No theoretical results are proved in a non-asymptotic framework.

Estimating the variance by the residual least-squares. Baraud et al. [10] 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 (18) for all $K > 1$. The choice $K = 1.1$ is suggested for the practice.

Slope heuristic. Lebarbier [42] implements the slope heuristic introduced by Birgé and Massart [17] 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 (17) with $K < 1$ can lead to a strong overfitting, whereas for $K > 1$ the bound (18) 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. [13].

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. [63]. 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 [64]. 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 [11] 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 [25] and in Gaussian graphical modeling. Indeed, Meinshausen and Bühlmann [48] 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 [31, 65] 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 [34] 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 Scaled Lasso [61] or the square-root Lasso [14] 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 $Y = \mathbf{X}B + E$, where B 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 are either that most rows of B are identically zero, or the rank of B 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 [40], is much more non-linear. Writing $\|.\|_F$ for the Frobenius (or Hilbert-Schmidt) norm, the problem of selecting among the estimators

$$\widehat{B}_r = \operatorname*{argmin}_{B:\operatorname{rank}(B)\le r} \|Y - \mathbf{X}B\|_F^2, \quad r \in \{1, \ldots, \min(T, \operatorname{rank}(\mathbf{X}))\}$$

has been investigated recently from a non-asymptotic point of view by Bunea *et al.* [18] and Giraud [33]. To handle the case of unknown variance, Bunea *et al.* [18] propose to plug an estimate of the variance in their selection criterion (which works when $\operatorname{rank}(\mathbf{X}) < n$), whereas Giraud [33] introduces a penalized log-likelihood criterion independent of the variance. Both papers provide oracle risk bounds for the resulting estimators showing rate-minimax adaptation.

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 [21] provides a tuning procedure for spline smoothing and Zhang [71] 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.* [11] 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 [11] provides an oracle bound for the selected estimator under the mild assumption that the effective dimension of the linear estimators is not larger than a fraction of n. This assumption can be viewed as a "sparsity assumption" suitable for linear estimators.

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 Celisse, A. (2010a). Segmentation of the mean of heteroscedastic data viacross-validation. *Stat. Comput.*, 1–20. 10.1007/s11222-010-9196-x, http://dx.doi.org/10.1007/s11222-010-9196-x.

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

[9] Baraud, Y. (2010). Estimator selection with respect to hellinger-type risks. *Probability Theory and Related Fields*, 1–49. http://dx.doi.org/10.1007/s00440-010-0302-y.

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

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

[12] 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)

[13] Baudry, J.-P., Maugis, C., and Michel, B. (2010). Slope heuristics: Overview and implementation. http://hal.archives-ouvertes.fr/hal-00461639/fr/.

[14] Belloni, A., Chernozhukov, V., and Wang, L. (2010). Square-root lasso: Pivotal recovery of sparse signals via conic programming. http://arxiv.org/pdf/1009.5689v2.

[15] 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. MRMR2533469

[16] Birgé, L. and Massart, P. (2001). Gaussian model selection. *J. Eur. Math. Soc. (JEMS)* **3**, 3, 203–268. MRMR1848946 (2002i:62072)

[17] 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. MRMR2288064 (2008g:62070)

[18] 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.

[19] Bunea, F., Tsybakov, A., and Wegkamp, M. (2007). Aggregation for Gaussian regression. *Ann. Statist.* **35**, 4, 1674–1697. MRMR2351101

[20] 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. MRMR2382644

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

[22] 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)

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

[24] 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)

[25] 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)

[26] 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)

[27] 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. MRMR2060166 (2005d:62116)

[28] 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)

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

[30] Gerchinovitz, S. (2011). Sparsity regret bounds for individual sequences in online linear regression. Arxiv:1101.1057.

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

[32] 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)

[33] Giraud, C. (2011). Low rank multivariate regression. *Electron. J. Stat. (to appear).* http://arxiv.org/abs/1009.5165.

[34] Giraud, C., Huet, S., and Verzelen, N. (2009). Graph selection with ggmselect. arXiv:0907.0619.

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

[36] 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)

[37] 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

[38] 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)

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

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

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

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

[43] 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

[44] 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)

[45] Lounici, K., Pontil, M., Tsybakov, A., and van de Geer, S. (2010). Oracle inequalities and optimal inference under group sparsity. Arxiv:1007.1771v3.

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

[47] Meinshausen, N. (2007). Relaxed Lasso. *Comput. Statist. Data Anal.* **52**, 1, 374–393. http://dx.doi.org/10.1016/j.csda.2006.12.019. MR2409990

[48] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34**, 3, 1436–1462. MRMR2278363 (2008b:62044)

[49] 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.

[50] 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)

[51] 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

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

[53] Rigollet, P. and Tsybakov, A. (2010). Exponential screening and optimal rates of sparse estimation. http://arxiv.org/pdf/1003.2654.

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

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

[56] 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)

[57] 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)

[58] 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

[59] 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)

[60] 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

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

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

[63] 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

[64] 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.

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

[66] Verzelen, N. (2010b). Minimax risks for sparse regressions: Ultra-high-dimensional phenomenons. `arXiv:1008.0526`.

[67] 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. MRMR2597190

[68] 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

[69] 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

[70] 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)

[71] 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)

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

[73] Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 476, 1418–1429. MRMR2279469 (2008d:62024)

[74] 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. MRMR2137327

APPENDIX A: A NOTE ON BIC TYPE CRITERIA

The BIC criterion has been initially introduced [54] 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*

$$\widehat{\lambda} \in \operatorname*{argmin}_{\lambda\in\widehat{\Lambda}} \left\{ -2\mathbf{L}_n(\widehat{\beta}_\lambda, \widehat{\sigma}_\lambda) + \log(n)\|\widehat{\beta}_\lambda\|_0 \right\}, \tag{19}$$

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

$$Y_i = \beta_{0,i} + \varepsilon_i, \ i = 1, \dots, n \ , \tag{20}$$

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 [62], SCAD [28] and the hard thresholding estimator. The hard thresholding estimator $\widehat{\beta}^{HT}_{\lambda}$ is defined for any $\lambda > 0$ by $[\widehat{\beta}^{HT}_{\lambda}]_i = Y_i \mathbf{1}_{|Y_i| \geq \lambda}$. 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 $x > 0$,

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

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 [27], Loubes and Massart advocate that soft thresholding estimators penalized by Mallows' C_p [46] 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 developped 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 [10, Sect. 3.2.2]. Empirically, BIC

seems to overfit except for the Lasso. We advise the practitionner 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 [11] in contrast to BIC.