Estimator selection in the Gaussian setting

Yannick Baraud, Christophe Giraud, Sylvie Huet

Submitted to the Annals of Statistics

ESTIMATOR SELECTION IN THE GAUSSIAN SETTING

By Yannick Baraud, Christophe Giraud and Sylvie Huet

Université de Nice Sophia Antipolis, Ecole Polytechnique and INRA

We consider the problem of estimating the mean f of a Gaussian vector Y the components of which are independent with a common variance that we assume to be unknown. Our estimation procedure is based on estimator selection. More precisely, we start with a collection $\mathbb{F}$ of estimators of f based on Y and, with the same data Y, we aim at selecting an estimator among $\mathbb{F}$ with the smallest Euclidean risk. We allow the cardinality of $\mathbb{F}$ to be very large (possibly infinite) and also the dependency of the estimators with respect to the data to be possibly unknown. We establish a non-asymptotic risk bound for the selected estimator. When $\mathbb{F}$ consists of linear estimators, we derive from this bound an oracle-type inequality. For illustration, we carry out two simulation studies. One aims at comparing our procedure to cross-validation for choosing a tuning parameter. The other shows how to implement our approach to solve the problem of variable selection in practice.

1. Introduction.

We consider the Gaussian regression framework

$$Y_i = f_i + \varepsilon_i,\ i = 1, \ldots, n$$

where $f = (f_1, \ldots, f_n)$ is an unknown vector of $\mathbb{R}^n$ and the ε_i are independent centered Gaussian random variables with common variance σ^2. Throughout the paper, σ^2 is assumed to be unknown. Our aim is to estimate f from the observation of Y. For specific forms of f, this setting allows to deal simultaneously with the following problems.

Example 1 (Signal denoising). *The vector f is of the form*

$$f = (F(x_1), \ldots, F(x_n))$$

where $x_1, \ldots, x_n$ are distinct points of a set $\mathcal{X}$ and F is an unknown mapping from $\mathcal{X}$ into $\mathbb{R}$.

AMS 2000 subject classifications: Primary 62J05; secondary 62J07, 62G05, 62G08, 62F07

Keywords and phrases: Gaussian linear regression, Estimator selection, Model selection, Variable selection, Linear estimator, Kernel estimator, Ridge regression, Lasso, Elastic net, Random Forest, PLS1 regression

EXAMPLE 2 (Linear regression). *The vector f is assumed to be of the form*

$$f = X\beta \tag{1}$$

where X is a $n \times p$ matrix, β is an unknown p-dimensional vector and p some integer larger than 1 (and possibly larger than n). The columns of the matrix X are usually called predictors. When p is large, one may assume that the decomposition (1) is sparse in the sense that only few β_j are non-zero. Estimating f or finding the predictors associated to the non-zero coordinates of β are usually problems of interest. This last one is called variable selection.

Our estimation strategy is based on estimator selection. More precisely, we start with a collection $\mathbb{F} = \{\hat{f}_\lambda,\ \lambda \in \Lambda\}$ of estimators of f based on Y and aim at selecting the one with the smallest Euclidean risk by using the same observation Y. The way the estimators $\hat{f}_\lambda$ depends on Y may be arbitrary and could even be unknown to the statistician. Throughout the paper, $\mathbb{F}$ is assumed to be finite, mostly for the sake of simplicity since the reader can check that our theoretical results would remain unchanged if $\mathbb{F}$ were not.

The problem of choosing some best estimator among a family of candidate ones is central in statistics. For illustration, let us present some examples.

EXAMPLE 3 (Choosing a tuning parameter). *Many statistical procedures depend on a (possibly multi-dimensional) parameter λ that needs to be tuned in view of obtaining an estimator with the best possible performance. For example, in the context of linear regression as described in Example 2, the Lasso estimator (see Tibshirani (1996) and Chen et al. (1998)) defined by $\hat{f}_\lambda = X\widehat{\beta_\lambda}$ with*

$$\widehat{\beta_\lambda} = \arg\min_{\beta \in \mathbb{R}^p} \left[\|Y - X\beta\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right]$$

depends on the choice of the parameter $\lambda \geq 0$. Selecting this parameter among a grid $\Lambda \subset \mathbb{R}_+$ amounts to selecting a (suitable) estimator among the family $\mathbb{F} = \{\hat{f}_\lambda,\ \lambda \in \Lambda\}$.

Another dilemma for statisticians is the choice of a procedure to solve a given problem. In the context of Example 3, there exist many competitors to the Lasso estimator and one may alternatively choose ridge or PLS1 estimators for example. Similarly, for the problem of signal denoising as described in

Example 1, popular approaches include spline smoothing, wavelet decompositions and kernel estimators, the choice of a suitable kernel being possibly tricky.

EXAMPLE 4 (Choosing a kernel). *Consider the problem described in Example 1 with $\mathcal{X} = \mathbb{R}$. For a kernel K and a bandwidth $h > 0$, the Nadaraya-Watson estimator (see Nadaraya (1964) and Watson (1964)) $\hat{f}_{K,h} \in \mathbb{R}^n$ is defined as*

$$\hat{f}_{K,h} = \left(\hat{F}_{K,h}(x_1), \ldots, \hat{F}_{K,h}(x_n)\right)$$

where for $x \in \mathbb{R}$

$$\hat{F}_{K,h}(x) = \frac{\sum_{j=1}^n K\left(\frac{x-x_j}{h}\right) Y_j}{\sum_{j=1}^n K\left(\frac{x-x_j}{h}\right)}.$$

There exist many possible choices for the kernel K, such as the Gaussian kernel $K(x) = e^{-x^2/2}$, the uniform kernel $K(x) = \mathbf{1}_{|x|<1}$, etc. Given a (finite) family $\mathcal{K}$ of candidate kernels K and a grid $\mathcal{H} \subset \mathbb{R}_+^*$ of possible values of h, one may consider the problem of selecting the best kernel estimator among the family $\mathbb{F} = \{\hat{f}_\lambda,\ \lambda = (K,h) \in \mathcal{K} \times \mathcal{H}\}$.

A common way to address the above issues is to use some cross-validation scheme such as leave-one-out or V-fold. Even though these resampling techniques are widely used in practice, little is known on the theoretical performances of the resulting estimator. We refer to Arlot and Celisse (2010) for a survey of the performances of cross-validation for model selection. Compared to these approaches, the procedure we propose is less time consuming and easier to implement. Moreover, it does not require to know how the estimators depend on the data Y and can therefore handle the following problem.

EXAMPLE 5 (Selecting among mute experts). *A statistician is given a collection $\mathbb{F} = \{\hat{f}_\lambda,\ \lambda \in \Lambda\}$ of estimators from a family Λ of experts λ, each of which keeping secret the way its estimator $\hat{f}_\lambda$ depends on the observation Y. The problem is then to find which expert λ is the closest to the truth.*

Given a selection rule among $\mathbb{F}$, an important issue is to compare the risk of the selected estimator to those of the candidate ones. Results in this direction are available in the context of model selection, which can be seen as a particular case of estimator selection. More precisely, for the purpose of selecting a suitable model one starts with a collection $\mathbb{S}$ of those, typically

linear spaces chosen for their approximation properties with respect to f, and one associates to each model $S \in \mathbb{S}$ a suitable estimator $\hat{f}_S$ with values in S. Selecting a model then amounts to selecting an estimator among the collection $\mathbb{F} = \{\hat{f}_S,\ S \in \mathbb{S}\}$. For this problem, selection rules based on the minimization of a penalized criterion have been proposed in the regression setting by Yang (1999), Baraud (2000), Birgé and Massart (2001) and Baraud *et al* (2009). Another way, usually called Lepski's method, appears in a series of papers by Lepski (1990; 1991; 1992a; 1992b) and was originally designed to perform model selection among collections of nested models. Finally, we mention that other procedures based on resampling have interestingly emerged from the work of Arlot (2007; 2009) and Célisse (2008). A common feature of those approaches lies in the fact that the proposed selection rules apply to specific collections of estimators only.

An alternative to *estimator selection* is *aggregation* which aims at designing a suitable combination of given estimators in order to outperform each of these separately (and even the best combination of these) up to a remaining term. Aggregation techniques can be found in Catoni (1997; 2004), Juditsky and Nemirovski (2000), Nemirovski (2000), Yang (2000a), (2000b), (2001), Tsybakov (2003), Wegkamp (2003), Birgé (2006), Rigollet and Tsybakov (2007), Bunea, Tsybakov and Wegkamp (2007) and Goldenshluger (2009) for $\mathbb{L}_p$-losses. Most of the aggregation procedures are based on a sample splitting, one part of the data being used for building the estimators, the remaining part for selecting among these. Such a device requires that the observations be i.i.d. or at least that one has at disposal two independent copies of the data. From this point of view our approach differs from *aggregation* since we use the whole data Y to build and select. In the Gaussian regression setting we consider, we mention the results of Leung and Barron (2006) for the problem of mixing least-squares estimators. Their procedure uses the same data Y to estimate and to aggregate but requires the variance to be known. Giraud (2008) extends their results to the case where it is unknown.

The main idea underlying our approach is the following. We introduce a collection $\mathbb{S}$ of linear subspaces of $\mathbb{R}^n$ in view of approximating the estimators considered in $\mathbb{F}$ and use a penalized criterion in order to compare them. Similar ideas appeared in Baraud (2010) where the estimators were compared pair by pair by means of a testing procedure. Unfortunately, the selection rule proposed there was computationally intractable. In contrast, the procedure we present is easy to implement, an `R`-package being available soon on `http://w3.jouy.inra.fr/unites/miaj/public/perso/SylvieHuet_en.html`.

The paper is organized as follows. In Section 2 we present our selection rule

and the theoretical properties of the resulting estimator. We explain how one can use our procedure for variable selection in Section 3, and for selecting among a collection of linear estimators in Section 4. Section 5 is devoted to two simulation studies. One aims at comparing the performance of our procedure to the classical V-fold in order to select a tuning parameter among a grid. In the other, we evaluate the performance of the variable selection procedure we propose to classical ones such as the Lasso, random forest, and others based on ridge and PLS regression, among others. Finally, the proofs are postponed to Section 6.

Throughout the paper C denotes a constant that may vary from line to line.

2. The general estimation procedure and the main result.

2.1. *Description of the general procedure and main assumptions.* Given a collection $\mathbb{F} = \{\widehat{f}_\lambda, \lambda \in \Lambda\}$ of estimators of f based on Y, the selection rule we propose is based on the choices of a family $\mathbb{S}$ of linear subspaces of $\mathbb{R}^n$, a collection $\{\mathbb{S}_\lambda,\ \lambda \in \Lambda\}$ of (possibly random) subsets of $\mathbb{S}$, a weight function Δ and a penalty function pen, both from $\mathbb{S}$ into $\mathbb{R}_+$. We introduce those objects below and refer to Sections 3 and 4 for examples.

2.1.1. *The collection of estimators* $\mathbb{F}$. The collection $\mathbb{F}$ is assumed to be finite, which corresponds to the practical case. Nevertheless, the reader can check that the results would remain unchanged by considering a countable collection $\mathbb{F}$ and even a collection with the cardinality of the continuum (provided that the final estimator remains a measurable function of the observation).

2.1.2. *The families* $\mathbb{S}$ *and* $\mathbb{S}_\lambda$. The collection $\mathbb{S}$ should contain linear subspaces with good approximation properties with respect to the elements of $\mathbb{F}$. We assume

ASSUMPTION 1. *For all* $S \in \mathbb{S}$, $\dim(S) \leq n - 2$.

For each $\lambda \in \Lambda$, we extract a subset $\mathbb{S}_\lambda$ of $\mathbb{S}$ the elements of which possess good approximation properties with respect to the estimator $\hat{f}_\lambda$ specifically. One may choose $\mathbb{S}_\lambda = \mathbb{S}$. However, when $\mathbb{S}$ is large it may be wise to consider a smaller subset of $\mathbb{S}$, mainly for computational reasons. The choice of $\mathbb{S}_\lambda$ may be made on the basis of the observation Y.

2.1.3. *The weight function* Δ *and the associated function* pen_Δ. We consider a function Δ from $\mathbb{S}$ into $\mathbb{R}_+$ and assume

ASSUMPTION 2.

$$\Sigma = \sum_{S\in\mathbb{S}} e^{-\Delta(S)} < +\infty. \tag{2}$$

Whenever $\mathbb{S}$ is finite, inequality (2) automatically holds true. However, in practice Σ should be kept to a reasonable size. When $\Sigma = 1$, $e^{-\Delta(.)}$ can be interpreted as a prior distribution on $\mathbb{S}$ and gives thus a Bayesian flavor to the procedure we propose. To the weight function Δ, we associate the function pen_Δ mapping $\mathbb{S}$ into $\mathbb{R}_+$ and defined by

$$\mathbb{E}\left[\left(U - \frac{\mathrm{pen}_\Delta(S)}{n-\dim(S)}V\right)_+\right] = e^{-\Delta(S)} \tag{3}$$

where x_+ denotes the positive part of $x \in \mathbb{R}$ and U, V two independent χ^2 random variables with respectively $\dim(S)+1$ and $n-\dim(S)-1$ degrees of freedom. This function can be easily computed from the quantiles of the Fisher distribution as we shall see in Section 7.1. From a more theoretical point of view, it is shown in Baraud *et al* (2009) that when $\dim(S) \vee \Delta(S) \leq \kappa n$ for some $\kappa < 1$, then for some constant C depending on κ only,

$$\mathrm{pen}_\Delta(S) \leq C(\dim(S) \vee \Delta(S)). \tag{4}$$

2.1.4. *The selection criterion.* The selection procedure we propose involves a penalty function pen from $\mathbb{S}$ into $\mathbb{R}_+$ satisfying

ASSUMPTION 3. *There exists some* $K > 1$ *such that*

$$\mathrm{pen}(S) \geq K\mathrm{pen}_\Delta(S) \quad \text{for all} \quad S \in \mathbb{S}. \tag{5}$$

Whenever equality holds in (5), it derives from (4) that $\mathrm{pen}(S)$ measures the complexity of the model S in terms of dimension and weight. Let us denote by Π_S the projection operator onto a linear space $S \subset \mathbb{R}^n$. Given the families $\mathbb{S}_\lambda$, the penalty function pen and some positive number α, we select the estimator $\hat{f}_{\hat{\lambda}}$ which minimizes over $\mathbb{F}$ the criterion

$$\mathrm{crit}_\alpha(\hat{f}_\lambda) = \inf_{S\in\mathbb{S}_\lambda}\left[\left\|Y - \Pi_S\hat{f}_\lambda\right\|^2 + \alpha\left\|\hat{f}_\lambda - \Pi_S\hat{f}_\lambda\right\|^2 + \mathrm{pen}(S)\,\hat{\sigma}_S^2\right], \tag{6}$$

with

$$\hat{\sigma}_S^2 = \frac{\|Y - \Pi_S Y\|^2}{n-\dim(S)}. \tag{7}$$

2.1.5. *Practical suggestions.* In practice, we suggest to take $\alpha = 1/2$ and $\mathrm{pen}(.) = K\mathrm{pen}_\Delta(.)$ with $K = 1.1$. These choices are based on the simulation study carried out in Section 5. The families $\mathbb{S}$, $\mathbb{S}_\lambda$ and the weight function Δ should be chosen accordingly to the statistical problem to be solved as we shall see on examples.

2.2. *The main result.* The following result holds.

THEOREM 1. *Under Assumptions 1 to 3, the estimator $\hat{f}_{\hat{\lambda}}$ minimizing (6) among the elements of $\mathbb{F} = \{\hat{f}_\lambda,\ \lambda \in \Lambda\}$ satisfies*

$$\begin{aligned} C\mathbb{E}\left[\left\|f - \hat{f}_{\hat{\lambda}}\right\|^2\right] &\leq \mathbb{E}\left[\inf_{\lambda\in\Lambda}\left\{\left\|f - \hat{f}_\lambda\right\|^2 + A(\hat{f}_\lambda, \mathbb{S}_\lambda)\right\}\right] + \sigma^2\Sigma \\ &\leq \inf_{\lambda\in\Lambda}\left\{\mathbb{E}\left[\left\|f - \hat{f}_\lambda\right\|^2\right] + \mathbb{E}\left[A(\hat{f}_\lambda, \mathbb{S}_\lambda)\right]\right\} + \sigma^2\Sigma \end{aligned} \tag{8}$$

where C is a constant given by (18) which depends on K and α only and

$$A(\hat{f}_\lambda, \mathbb{S}_\lambda) = \inf_{S\in\mathbb{S}_\lambda}\left[\left\|\hat{f}_\lambda - \Pi_S\hat{f}_\lambda\right\|^2 + \mathrm{pen}(S)\,\hat{\sigma}_S^2\right] \quad \text{for all } \lambda \in \Lambda.$$

Comments. It follows from (8) that the risk bound we get is decreasing (for the inclusion) with respect to Λ. This means that if one adds a new estimator to the collection $\mathbb{F}$ (without changing neither $\mathbb{S}$ nor the families $\mathbb{S}_\lambda$ associated to the former estimators), the risk bound for $\hat{f}_{\hat{\lambda}}$ can only be improved. The best possible risk bound (up to a universal constant) is

$$\inf_{S\in\mathbb{S}}\left\{\|f - \Pi_S f\|^2 + \mathrm{pen}(S)\sigma^2\right\} + \sigma^2\Sigma.$$

It is achieved when $\mathbb{F}$ is, or at least contains, the family of projection estimators $\{\Pi_S Y,\ S \in \mathbb{S}\}$ (by taking $\mathbb{S}_\lambda = \{S\}$ when $\hat{f}_\lambda = \Pi_S Y$). When $\mathbb{S}$ is too large, selecting among such a family becomes intractable. The idea of the present paper is to solve the computational issue by considering a set $\mathbb{F}$ of smaller cardinality and still maintain good performances for $\hat{f}_{\hat{\lambda}}$ by using more sophisticated estimators $\hat{f}_\lambda$ than just projection ones.

The quantity $A(\hat{f}_\lambda, \mathbb{S}_\lambda)$ corresponds to a cost for approximating $\hat{f}_\lambda$ by $\bigcup_{S\in\mathbb{S}_\lambda} S$. This cost is small as soon as there exists a good approximation model S in $\mathbb{S}_\lambda$ associated to a small value of $\mathrm{pen}(S)$. When the quantities $\mathbb{E}[A(\hat{f}_\lambda, \mathbb{S}_\lambda)]$ are small compared to $\mathbb{E}[\|f - \hat{f}_\lambda\|^2]$ for all $\lambda \in \Lambda$, we derive from (8) the oracle-type inequality

$$C\mathbb{E}\left[\left\|f - \hat{f}_{\hat{\lambda}}\right\|^2\right] \leq \inf_{\lambda\in\Lambda}\mathbb{E}\left[\left\|f - \hat{f}_\lambda\right\|^2\right] + \sigma^2\Sigma.$$

As we shall see, we obtain a risk bound of this flavor when $\mathbb{F}$ consists of arbitrary linear estimators provided that the $\mathbb{S}_\lambda$ are chosen in a suitable way.

The quantity $A(\hat{f}_\lambda, \mathbb{S}_\lambda)$ is minimum for the choice $\mathbb{S}_\lambda = \mathbb{S}$. Nevertheless, the computation of $\mathrm{crit}_\alpha(\hat{f}_\lambda)$ given by (6) may be unpractical whenever $\mathbb{S}_\lambda$ is too large. Consequently, if one knows a good (possibly random) approximation model $\widehat{S}_\lambda \in \mathbb{S}$ for $\hat{f}_\lambda$, it is convenient to choose $\mathbb{S}_\lambda = \{\widehat{S}_\lambda\}$ in practice. The case where $\hat{f}_\lambda$ belongs to $\widehat{S}_\lambda$ with probability one is of special interest and leads to the following corollary.

COROLLARY 1. *Assume that Assumptions 1 to 3 hold and that there exists some $\kappa \in (0,1)$ such that $1 \le \dim(S) \vee \Delta(S) \le \kappa n$ for all $S \in \mathbb{S}$. Besides, assume that for all $\lambda \in \Lambda$, there exists a (possibly random) set $\widehat{S}_\lambda \in \mathbb{S}_\lambda$ such that $\hat{f}_\lambda \in \widehat{S}_\lambda$ with probability 1. If* pen *is chosen to achieve equality in (5), then for some constant C depending on κ, α, K and Σ, the estimator $\hat{f}_{\hat{\lambda}}$ satisfies*

$$C\mathbb{E}\left[\left\|f - \hat{f}_{\hat{\lambda}}\right\|^2\right] \le \inf_{\lambda\in\Lambda}\left[\mathbb{E}\left[\left\|f - \hat{f}_\lambda\right\|^2\right] + \mathbb{E}\left[\dim(\widehat{S}_\lambda) \vee \Delta(\widehat{S}_\lambda)\right]\sigma^2\right]. \tag{9}$$

The assumption that $\hat{f}_\lambda$ belongs to $\widehat{S}_\lambda$ with probability 1 for all λ, is usually met in the context of model selection. In this situation, as already explained, one starts with a collection of models $\mathbb{S} = \{S_m,\ m \in \mathcal{M}\}$ and associate to each S_m an estimator $\hat{f}_m$ with values in S_m. To perform model selection, one may apply our selection procedure to the collection $\mathbb{F} = \{\hat{f}_m,\ m \in \mathcal{M}\}$ (here $\Lambda = \mathcal{M}$) with $\mathbb{S}_m = \{S_m\}$ for all $m \in \mathcal{M}$.

In the particular case where $\hat{f}_m = \Pi_{S_m} Y$ for all $m \in \mathcal{M}$, this selection rule turns out to be exactly the same as that described in Baraud *et al* (2009). It is proved there that under the assumptions of Corollary 1,

$$C\mathbb{E}\left[\left\|f - \hat{f}_{\hat{m}}\right\|^2\right] \le \inf_{m\in\mathcal{M}}\left[\mathbb{E}\left[\left\|f - \hat{f}_m\right\|^2\right] + (\dim(S_m) \vee \Delta(S_m))\,\sigma^2\right]. \tag{10}$$

Inequality (9) generalizes (10) in the sense that the family $\mathcal{M}$ is now allowed to be random depending on Y. Moreover, Corollary 1 shows that the result is not only true for collections of projection estimators but also, more generally, for all families $\mathbb{F}$ such that $\hat{f}_m$ belongs to S_m for all $m \in \mathcal{M}$.

3. Variable selection. Throughout this section, we consider the problem of variable selection introduced in Example 2. When p is small enough

(say smaller than 20), this problem can be solved by using a suitable variable selection procedure exploring all the subsets of $\{1,\dots,p\}$. For example, one may use the penalized criterion introduced in Birgé and Massart (2001) when the variance is known and in Baraud *et al* (2009) when it is not. When p is larger, such an approach can no longer be applied since it becomes numerically intractable. To overcome this problem, algorithms based on the minimization of convex criteria have been proposed among which the Lasso, the Dantzig selector of Candès and Tao (2007), the elastic net of Zou and Hastie (2005). An alternative to those criteria is the forward-backward algorithm described in Zhang (2008), among others. Since there seems to be no evidence that one of these procedures outperforms all the others, it may be reasonable to mix them all and let the data decide which is the more appropriate to solve the problem at hand. As enlarging $\mathbb{F}$ can only improve the risk bound of our estimator, only the CPU resources should limit the number of candidate estimators.

The procedure we propose could not only be used to select among those candidate procedures but also to select the tuning parameters they depend on. From this point of view, it provides an alternative to the cross-validation techniques which are quite popular but offer little theoretical guarantees.

3.1. *Implementation roadmap.* Start by choosing a family $\mathcal{L}$ of variable selection procedures. Examples of such procedures are the Lasso, the Dantzig selector, the elastic net, among others. If necessary, associate to each $\ell \in \mathcal{L}$ a family of tuning parameters H_ℓ. For example, in order to use the Lasso procedure one needs to choose a tuning parameter $h > 0$ among a grid $H_{\mathrm{Lasso}} \subset \mathbb{R}_+$. If a selection procedure ℓ requires no choice of tuning parameters, then one may take $H_\ell = \{0\}$. Let us denote by $\widehat{m}(\ell, h)$ the subset of $\{1,\dots,p\}$ corresponding to the predictors selected by the procedure ℓ for the choice of the tuning parameter h and for $m \subset \{1,\dots,p\}$, let S_m be the linear span of the column vectors $X_{\cdot,j}$ for $j \in m$ (with the convention $S_\varnothing = \{0\}$). For $\ell \in \mathcal{L}$ and $h \in H_\ell$, associate to the subset $\widehat{m}(\ell, h)$ an estimator $\hat{f}_{(\ell,h)}$ of f with values in $S_{\widehat{m}(\ell,h)}$ (for example, the projection estimator). Finally, consider the family $\mathbb{F} = \{\hat{f}_\lambda,\ \lambda \in \Lambda\}$ of those estimators by taking $\Lambda = \bigcup_{\ell\in\mathcal{L}}(\{\ell\} \times H_\ell)$.

3.1.1. *The approximation spaces and the weight function.* Throughout, we shall restrict ourselves to subsets of predictors with cardinality not larger than some $D_{\max} \leq n-2$. In view of approximating the estimators $\hat{f}_\lambda$, we

suggest the collection $\mathbb{S}$ given by

$$\mathbb{S} = \bigcup \left\{ S_m \middle|\ m \subset \{1, \ldots, p\}, \mathrm{card}(m) \leq D_{\max} \right\}. \tag{11}$$

We associate to $\mathbb{S}$ the weight function Δ defined for $S \in \mathbb{S}$ by

$$\Delta(S) = \log\left[\binom{D}{p}\right] + \log(1+D) \quad \text{with} \quad D = \dim(S). \tag{12}$$

Since

$$\begin{aligned} \sum_{S \in \mathbb{S}} e^{-\Delta(S)} &= \sum_{D=0}^{p} \sum_{\substack{S \in \mathbb{S} \\ \dim(S) = D}} e^{-\Delta(S)} \\ &\leq \sum_{D=0}^{p} e^{-\log(1+D)} \leq 1 + \log(1+p), \end{aligned}$$

Assumption 2 is satisfied with $\Sigma = 1 + \log(1+p)$.

Let us now turn to the choices of the $\mathbb{S}_\lambda \subset \mathbb{S}$. The criterion given by (6) cannot be computed when $\mathbb{S}_\lambda = \mathbb{S}$ for all λ as soon as p is too large. In such a case, one must consider a smaller subset of $\mathbb{S}$ and we suggest for $\lambda = (\ell, h) \in \Lambda$

$$\mathbb{S}_{(\ell,h)} = \left\{ S_{\widehat{m}(\ell,h')},\ h' \in H_\ell \right\}$$

(where the S_m are defined in Section 3.1), or preferably

$$\mathbb{S}_{(\ell,h)} = \left\{ S_{\widehat{m}(\ell',h')},\ \ell' \in \mathcal{L}, h' \in H_\ell \right\}$$

whenever this latter family is not too large. Note that these two families are random.

3.1.2. *The case of projection estimators.* When $\mathbb{F}$ is the family of projection estimators $\Pi_{S_{\widehat{m}(\ell,h)}} Y$ associated to the family of random subsets

$$\widehat{\mathcal{M}} = \left\{ \widehat{m}(\ell,h), (\ell,h) \in \mathcal{L} \times H_\ell \right\},$$

one can merely take $\mathbb{S}_{(\ell,h)} = \{S_{\widehat{m}(\ell,h)}\}$ for $(\ell,h) \in \mathcal{L} \times H_\ell$. In this case, selecting among $\mathbb{F}$ by minimizing (6) amounts to minimizing over $\widehat{\mathcal{M}}$

$$\mathrm{crit}(m) = \|Y - \Pi_{S_m} Y\|^2 + K\, \mathrm{pen}_\Delta(S_m) \hat{\sigma}^2_{S_m}, \tag{13}$$

where pen_Δ is given by (3). An example of family $\widehat{\mathcal{M}}$ of interest is described in Section 7.3. It consists of data-driven subsets $\widehat{m}(\ell,h)$ obtained from the Lasso, ridge regression, elastic net, PLS1 regression and random forest.

3.2. *Theoretical guarantees.* Our choices of Δ and $\mathbb{S}_\lambda$ ensure that $\hat{f}_\lambda \in S_{\widehat{m}(\lambda)} \in \mathbb{S}_\lambda$ for all $\lambda \in \Lambda$ and that

$$\Delta(S_{\widehat{m}(\lambda)}) \leq 2 \dim(S_{\widehat{m}(\lambda)}) \log p.$$

Hence, by applying Corollary 1 with $\widehat{S}_\lambda = S_{\widehat{m}(\lambda)}$, we get that if $D_{\max} \leq \kappa n/(2 \log p)$, the selected estimator satisfies

$$C\mathbb{E}\left[\left\|f - \hat{f}_{\hat{\lambda}}\right\|^2\right] \leq \inf_{\lambda \in \Lambda}\left[\mathbb{E}\left[\left\|f - \hat{f}_\lambda\right\|^2\right] + \mathbb{E}\left[\dim(S_{\widehat{m}(\lambda)})\right] \log(p)\sigma^2\right].$$

4. Selecting among linear estimator. In this section, we focus on the case where the estimators $\hat{f}_\lambda$ are linear, that is, are of the form $\hat{f}_\lambda = A_\lambda Y$ for some deterministic linear operator $A_\lambda : \mathbb{R}^n \to \mathbb{R}^n$. As mentioned before, this setting covers many popular estimation procedures including kernel ridge estimators, spline smoothing, Nadaraya estimators, λ-nearest neighbors, projection estimators, low-pass filters, etc. In some cases A_λ is symmetric (e.g. kernel ridge, spline smoothing, projection estimators), in some others A_λ is non-symmetric and non-singular (as for Nadaraya estimators) and sometimes A_λ can be both singular and non-symmetric (low pass filters, λ-nearest neighbors).

All the procedures mentioned above have a tuning parameter (possibly multidimensional) and their practical performance strongly depends on the value of this parameter. A series of papers have investigated the calibration of some of these procedures. To mention a few of them, Cao and Golubev (2006) focus on spline smoothing, Zhang (2005) on kernel ridge regression, Goldenshluger and Lepski (2009) on kernel estimators and Arlot and Bach (2009) propose a procedure to select among symmetric linear estimator with spectrum in $[0, 1]$. The procedure we present can handle all these cases in an unified framework. As in Section 3 for the problem of variable selection, one may index the estimators $\hat{f}_\lambda$ by a pair $\lambda = (\ell, h)$ corresponding to the choice of a procedure ℓ and a tuning parameter h.

4.1. *The approximation families* $\mathbb{S}_\lambda$. To apply our selection procedure, we need to associate to each A_λ a suitable collection of approximation spaces $\mathbb{S}_\lambda$. To do so, we introduce below an approximation space S_λ which plays a key role in our analysis.

For the sake of simplicity, let us first consider the case where A_λ is non-singular. The approximation space S_λ is then defined as the linear span of the right-singular vectors of $A_\lambda^{-1} - I$ associated to singular values smaller than

1. When A_λ is symmetric, S_λ is merely the linear span of the eigenvectors of A_λ associated to eigenvalues not smaller than $1/2$.

Let us now extend the definition of S_λ to singular operators A_λ. Let us recall that $\mathbb{R}^n = \ker(A_\lambda) \oplus \mathrm{rg}(A_\lambda^*)$ where A_λ^* stands for the transpose of A_λ and $\mathrm{rg}(A_\lambda^*)$ for its range. The operator A_λ then induces a one to one operator between $\mathrm{rg}(A_\lambda^*)$ and $\mathrm{rg}(A_\lambda)$. Write A_λ^+ for the inverse of this operator from $\mathrm{rg}(A_\lambda)$ to $\mathrm{rg}(A_\lambda^*)$. The orthogonal projection operator from $\mathbb{R}^n$ onto $\mathrm{rg}(A_\lambda^*)$ induces a linear operator from $\mathrm{rg}(A_\lambda)$ into $\mathrm{rg}(A_\lambda^*)$, denoted $\overline{\Pi}_\lambda$. The approximation space S_λ is defined as the linear span of the right-singular vectors of $A_\lambda^+ - \overline{\Pi}_\lambda$ associated to singular values smaller than 1. When A_λ is non-singular or symmetric, we recover the definition of S_λ given above.

For each $\lambda \in \Lambda$, we choose $\mathbb{S}_\lambda \supset \{S_\lambda\}$. For example, we may take $\mathbb{S}_\lambda = \{S_\lambda\}$ or alternatively $\mathbb{S}_\lambda = \{S_\lambda^1, \ldots, S_\lambda^{n-2}\}$ where S_λ^k is the linear span of the right-singular vectors associated to the k smallest singular values of $A_\lambda^+ - \overline{\Pi}_\lambda$.

4.2. *Collection* $\mathbb{S}$ *and weights* Δ. The minimal choice for $\mathbb{S}$ is $\mathbb{S} = \bigcup_{\lambda\in\Lambda} \mathbb{S}_\lambda$. As to the weight function Δ, we suggest to take it of the form

$$\Delta(S) = a \dim(S) \quad \text{for all } S \in \mathbb{S}$$

where a is a suitable positive number for which the value of Σ is not too large (say not larger than 1).

4.3. *Theoretical garanties.* The following holds.

COROLLARY 2. *Assume that there exists* $\kappa \in (0,1)$ *such that*

$$1 \le \dim(S_\lambda) \le (a^{-1}\kappa n) \wedge \kappa n \wedge (n-2) \quad \textit{for all } \lambda \in \Lambda.$$

If pen *achieves equality in (5) then*

$$C\mathbb{E}\left[\left\|f - \hat{f}_{\hat{\lambda}}\right\|^2\right] \le (a \vee 1) \inf_\lambda \mathbb{E}\left[\left\|f - \hat{f}_\lambda\right\|^2\right],$$

for some C *depending on* K, α, Σ *and* κ *only.*

Risk bounds for the problem of selecting among a family of linear estimators have also been obtained by Arlot and Bach (2009) in the Gaussian regression framework, and Goldenshluger and Lepski (2009) in the multidimensional Gaussian white noise model. Arlot and Bach propose a penalized

procedure based on random penalties in view of selecting among families of linear estimators. Unlike ours, their approach requires that the operators be symmetric with eigenvalues in $[0, 1]$ and that the cardinality of Λ is at most polynomial with respect to n. Goldenshluger and Lepski propose a selection rule among families of kernel estimators to solve the problem of structural adaptation. Their approach requires suitable assumptions on the kernels while ours requires almost nothing. Nevertheless, we restrict to the case of the Euclidean loss whereas Goldenshluger and Lepski consider more general $\mathbb{L}_p$ ones.

5. Simulation study. In the linear regression setting described in Example 2, we carry out a simulation study to evaluate the performances of our procedure to solve the two following problems.

We first consider the problem, described in Example 3, of tuning the smoothing parameter of the Lasso procedure for estimating f. The performances of our procedure are compared with those of the V-fold cross-validation method. Secondly, we consider the problem of variable selection. We solve it by using our criterion in view of selecting among a family $\mathcal{L}$ of candidate variable selection procedures.

Our simulation study is based on a large number of examples which have been chosen in view of covering a large variety of situations. Most of these have been found in the literature in the context of Example 2 either for estimation or variable selection purposes when the number p of predictors is large.

The section is organized as follows. The simulation design is given in the following section. Then, we describe how our procedure is applied for tuning the Lasso and performing variable selection. Finally, we give the results of the simulation study.

5.1. *Simulation design.* One example is determined by the number of observations n, the number of variables p, the $n \times p$ matrix X, the values of the parameters β, and the ratio signal/noise ρ. It is denoted by $\mathrm{ex}(n, p, X, \beta, \rho)$, and the set of all considered examples is denoted $\mathcal{E}$. For each example, we carry out 400 simulations of Y as a Gaussian random vector with expectation $f = X\beta$ and variance $\sigma^2 I_n$, where I_n is the $n \times n$ identity matrix, and $\sigma^2 = \|f\|^2/n\rho$.

The collection $\mathcal{E}$ is composed of several collections $\mathcal{E}_e$ for $e = 1, \ldots, E$ where each collection $\mathcal{E}_e$ is characterized by a vector of parameters β_e, and a set

$\mathcal{X}_e$ of matrices X:

$$\mathcal{E}_e = \{\mathrm{ex}(n,p,X,\beta,\rho) : (n,p) \in \mathcal{I}, X \in \mathcal{X}_e, \beta = \beta_e, \rho \in \mathcal{R}\}$$

where $\mathcal{R} = \{5, 10, 20\}$ and $\mathcal{I}$ consists of pairs (n,p) such that p is smaller, equal or greater than n. The examples are described in further details in Section 7.2. They are inspired by examples found in Tibshirani (1996), Zou and Hastie (2005), Zou (2006), and Huang et al. (2008) for comparing the Lasso method to the ridge, adaptive Lasso and elastic net methods. They make up a large variety of situations. They include cases where

- the covariates are not, moderately or strongly correlated,
- the covariates with zero coefficients are weakly or highly correlated with covariates with non-zero coefficients,
- the covariates with non-zero coefficients are grouped and correlated within these groups,
- the lasso method is known to be inconsistent,
- few or many effects are present.

5.2. *Tuning a smoothing parameter.* We consider here the problem of tuning the smoothing parameter of the Lasso estimator as described in Example 3. Instead of considering the Lasso estimators for a fixed grid Λ of smoothing parameters λ, we rather focus on the sequence $\{\hat{f}_1, \ldots, \hat{f}_{D_{\max}}\}$ of estimators given by the $D_{\max}$ first steps of the LARS-Lasso algorithm proposed by Efron *et al.* (2004). Hence, the tuning parameter is here the number $h \in H = \{1, \ldots, D_{\max}\}$ of steps. In our simulation study, we compare the performance of our criterion to that of the V-fold cross-validation for the problem of selecting the best estimator among the collection $\mathbb{F} = \{\hat{f}_1, \ldots, \hat{f}_{D_{\max}}\}$.

The estimator of f based on our procedure. We recall that our selection procedure relies on the choices of families $\mathbb{S}$, $\mathbb{S}_h$ for $h \in H$, a weight function Δ, a penalty function pen and two universal constants $K > 1$ and $\alpha > 0$. We choose the family $\mathbb{S}$ defined by (11). We associate to $\hat{f}_h$ the family $\mathbb{S}_h = \{S_{\widehat{m}(h')} |\ h' \in H\} \subset \mathbb{S}$ where the S_m are defined in Section 3.1 and $\widehat{m}(h') \subset \{1, \ldots, p\}$ is the set of indices corresponding to the predictors retuned by the LARS-Lasso algorithm at step $h' \in H$. We take $\mathrm{pen}(S) = K\mathrm{pen}_\Delta(S)$ with $\Delta(S)$ defined by (12) and $K = 1.1$. This value of K is consistent with what is suggested in Baraud *et al.* (2009). The choice of α is based on the following considerations. First, choosing α around one seems reasonable since it weights similarly the term $\|Y - \Pi_S \widehat{f}_\lambda\|^2$ which

			quantiles				
procedure	mean	std-err	0%	50%	75%	99%	100%
CV	1.18	0.08	1.05	1.18	1.24	1.36	1.38
pen_Δ	1.065	0.06	1.01	1.055	1.084	1.18	2.27

TABLE 1

Mean, standard-error and quantiles of the ratios $R_{\text{ex}}/O_{\text{ex}}$ calculated over all $\text{ex} \in \mathcal{E}$ *such that $O_{\text{ex}} < n\sigma^2/3$. The number of such examples equals 654, see Section 7.2.*

measures how well the estimator fits the data and the approximation term $\|\widehat{f}_\lambda - \Pi_S \widehat{f}_\lambda\|^2$ involved in our criterion (6). Second, simple calculation shows that the constant $C^{-1} = C^{-1}(1.1, \alpha)$ involved in Theorem 1 is minimum for α close to 0.6. We therefore carried out our simulations for α varying from 0.2 to 1.5. The results being very similar for α between 0.5 and 1.2, we choose $\alpha = 0.5$. We denote by $\widehat{f}_{\text{pen}_\Delta}$ the resulting estimator of f.

The estimator of f based on V-fold cross-validation. For each $h \in H$, the prediction error is estimated using a V-fold cross-validation procedure, with $V = n/10$. The estimator $\widehat{f}_{CV}$ is chosen by minimizing the estimated prediction error.

The results. The simulations were carried out with `R` (`www.r-project.org`) using the library `elasticnet`.

For each example $\text{ex} \in \mathcal{E}$, we estimate on the basis of 400 simulations the oracle risk

$$O_{\text{ex}} = \mathbb{E}\left(\min_{h \in H} \|f - \widehat{f}_h\|^2\right), \tag{14}$$

and the Euclidean risks $R_{\text{ex}}(\widehat{f}_{\text{pen}_\Delta})$ and $R_{\text{ex}}(\widehat{f}_{CV})$ of $\widehat{f}_{\text{pen}_\Delta}$ and $\widehat{f}_{CV}$ respectively.

The results presented in Table 1 show that our procedure tends to choose a better estimator than the CV in the sense that the ratios $R_{\text{ex}}(\widehat{f}_{\text{pen}_\Delta})/O_{\text{ex}}$ are closer to one than $R_{\text{ex}}(\widehat{f}_{CV})/O_{\text{ex}}$.

Nevertheless, for a few examples these ratios are larger for our procedure than for the CV. These examples correspond to situations where the Lasso estimators are highly biased.

In practice, it is worth considering several estimation procedures in order to increase the chance to have good estimators of f among the family $\mathbb{F}$. Selecting among candidate procedures is the purpose of the following simulation experiment in the variable selection context.

5.3. *Variable selection.* In this section, we consider the problem of variable selection and use the procedure and notations introduced in Section 3. To solve this problem, we deal with projection estimators. More precisely, for a subset m of $\{1,\dots,p\}$ we denote by $\hat{f}_m$ the estimator $\Pi_{S_m}Y$ and consider the family $\mathbb{F}=\{\hat{f}_{\widehat{m}(\ell,h)}|\ (\ell,h)\in\mathcal{L}\times H_\ell\}$, the descriptions of $\mathcal{L}$ and H_ℓ being postponed to Section 7.3. Let us merely mention that we choose $\mathcal{L}$ which gathers variable selection procedures based on the Lasso, ridge regression, Elastic net, PLS regression, Adaptive Lasso, Random Forest, and whenever possible, on an exhaustive research among the subsets of $\{1,\dots,p\}$.

5.3.1. *Results.* The simulations were carried out with `R` (`www.r-project.org`) using the libraries `elasticnet`, `randomForest`, `pls` and the program `lm.ridge` in the library `MASS`. We first select the tuning parameters associated to the procedures ℓ in $\mathcal{L}$. More precisely, for each ℓ we select an estimator among the collection $\mathbb{F}_\ell=\{\hat{f}_{\widehat{m}(\ell,h)}|\ h\in H_\ell\}$ by minimizing Criterion (13) over $\widehat{\mathcal{M}}_\ell=\{\widehat{m}(\ell,h)|h\in H_\ell\}$. We denote by $\widehat{m}(\ell)$ the selected set and by $\hat{f}_{\widehat{m}(\ell)}$ the corresponding projection estimator. For each example ex $\in\mathcal{E}$ and each method $\ell\in\mathcal{L}$, we estimate the risk

$$R_{\mathrm{ex},\ell}=\mathbb{E}\left(\|f-\hat{f}_{\widehat{m}(\ell)}\|^2\right)$$

of $\hat{f}_{\widehat{m}(\ell)}$ on the basis of 400 simulations and we do the same to calculate that of our estimator $\hat{f}_{\widehat{m}}$,

$$R_{\mathrm{ex,all}}=\mathbb{E}\left(\|f-\hat{f}_{\widehat{m}}\|^2\right).$$

Let us now define the minimum of these risks over all methods:

$$R_{\mathrm{ex,min}}=\min\left\{R_{\mathrm{ex,all}},R_{\mathrm{ex},\ell},\ell\in\mathcal{L}\right\}.$$

We compare the ratios $R_{\mathrm{ex},\ell}/R_{\mathrm{ex,min}}$ for $\ell\in\mathcal{L}\cup\{\mathrm{all}\}$ to judge the performances of the candidate procedures on each example ex $\in\mathcal{E}$. The mean, standard deviations and quantiles of the sequence $\{R_{\mathrm{ex},\ell}/R_{\mathrm{ex,min}},\ \mathrm{ex}\in\mathcal{E}\}$ are presented in Table 2. In particular, the results show that

- none of the procedures ℓ in $\mathcal{L}$ outperforms all the others simultaneously over all examples,
- our procedure, corresponding to $\ell=$ all, achieves the smallest mean value. Besides, this value is very close to one.
- the variability of our procedure is small compared to the others
- for all examples, our procedure selects an estimator the risk of which does not exceed twice that of the oracle.

			quantiles			
method	mean	std-err	50%	75%	95%	100%
Lasso	2.82	9.40	1.12	1.33	6.38	127
ridge	1.76	1.90	1.42	1.82	2.87	36.9
pls	1.50	1.20	1.22	1.50	2.58	17
en	1.46	1.90	1.12	1.33	2.57	29
ALridge	1.20	0.31	1.15	1.26	1.51	5.78
ALpls	1.29	0.87	1.14	1.29	1.75	12.7
rFmse	4.13	9.50	1.38	2.04	19.2	118
rFpurity	3.99	10.00	1.42	2.06	15.1	138
exhaustive	22.9	45	6.30	24.5	92.9	430
all	1.16	0.16	1.12	1.25	1.47	1.95

TABLE 2

For each $\ell \in \mathcal{L} \cup \{\text{all}\}$, mean, standard-error and quantiles of the ratios $R_{\text{ex},\ell}/R_{\text{ex,min}}$ calculated over all ex $\in \mathcal{E}$. *The number of examples in the collection $\mathcal{E}$ is equal to 660.*

	$\mathcal{E}_1$	$\mathcal{E}_2$	$\mathcal{E}_3$	$\mathcal{E}_4$	$\mathcal{E}_5$	$\mathcal{E}_6$	$\mathcal{E}_7$	$\mathcal{E}_8$	$\mathcal{E}_9$	$\mathcal{E}_{10}$	$\mathcal{E}_{11}$
FDR	0.045	0.026	0.004	0.026	0.018	0.041	0.012	0.026	0.042	0.15	0.014
TDR	0.74	0.63	0.18	0.63	0.17	0.99	1	1	0.98	0.29	0.20

TABLE 3

False dicovery rate (FDR) and true discovery rate (TDR) using our method, for each example with $\rho = 10$ and $n = p = 100$.

The false discovery rate (FDR) and the true discovery rate (TDR) are also parameters of interest in the context of variable selection. These quantities are given at Table 3 for each example when $\rho = 10$ and $n = p = 100$. Except for one example, the FDR is small, while the TDR is varying a lot among the examples.

6. Proofs.

6.1. *Proof of Theorem 1.* Throughout this section, we use the following notations. For all $\lambda \in \Lambda$, $S(\lambda)$ denotes any minimizer among $S \in \mathbb{S}_\lambda$ of

$$\text{crit}_\alpha(\hat{f}_\lambda, S) = \left\| Y - \Pi_S \hat{f}_\lambda \right\|^2 + \sigma^2 \mathfrak{pen}(S) + \alpha \left\| \hat{f}_\lambda - \Pi_S \hat{f}_\lambda \right\|^2,$$

where

$$\mathfrak{pen}(S) = \text{pen}(S)\, \hat{\sigma}_S^2/\sigma^2, \quad \text{for all } S \in \mathbb{S}. \tag{15}$$

We also write $\varepsilon = Y - f$ and $\overline{S}$ for the linear space generated by S and f. It follows the facts that for all $\lambda \in \Lambda$ and $S \in \mathbb{S}_\lambda$

$$\text{crit}_\alpha(\hat{f}_{\hat{\lambda}}) \le \text{crit}_\alpha(\hat{f}_\lambda) \le \text{crit}_\alpha(\hat{f}_\lambda, S)$$

and simple algebra that

$$\begin{aligned}
&\left\|f-\Pi_{S(\hat{\lambda})}\hat{f}_{\hat{\lambda}}\right\|^2+\alpha\left\|\hat{f}_{\hat{\lambda}}-\Pi_{S(\hat{\lambda})}\hat{f}_{\hat{\lambda}}\right\|^2\\
&\quad\leq\quad\left\|f-\Pi_S\hat{f}_\lambda\right\|^2+\alpha\left\|\hat{f}_\lambda-\Pi_S\hat{f}_\lambda\right\|^2+2\sigma^2\,\mathfrak{pen}(S)\\
&\qquad+2\langle\varepsilon,\Pi_{S(\hat{\lambda})}\hat{f}_{\hat{\lambda}}-f\rangle-\sigma^2\,\mathfrak{pen}(S(\hat{\lambda}))\;+\;2\langle\varepsilon,f-\Pi_S\hat{f}_\lambda\rangle-\sigma^2\,\mathfrak{pen}(S).
\end{aligned}$$

For $\lambda\in\Lambda$ and $S\in\mathbb{S}$, let us set $u_{\lambda,S}=\left(\Pi_S\hat{f}_\lambda-f\right)/\left\|\Pi_S\hat{f}_\lambda-f\right\|$ if $\Pi_S\hat{f}_\lambda\neq f$ and $u_{\lambda,S}=0$ otherwise. For all λ and S, we have $u_{\lambda,S}\in\overline{S}$ and

$$\begin{aligned}
&\left\|f-\Pi_{S(\hat{\lambda})}\hat{f}_{\hat{\lambda}}\right\|^2+\alpha\left\|\hat{f}_{\hat{\lambda}}-\Pi_{S(\hat{\lambda})}\hat{f}_{\hat{\lambda}}\right\|^2\\
&\quad\leq\quad\left\|f-\Pi_S\hat{f}_\lambda\right\|^2+\alpha\left\|\hat{f}_\lambda-\Pi_S\hat{f}_\lambda\right\|^2+2\sigma^2\,\mathfrak{pen}(S)\\
&\qquad+2\left|\langle\varepsilon,u_{\hat{\lambda},S(\hat{\lambda})}\rangle\right|\left\|\Pi_{S(\hat{\lambda})}\hat{f}_{\hat{\lambda}}-f\right\|-\sigma^2\,\mathfrak{pen}(S(\hat{\lambda}))\\
&\qquad+2\left|\langle\varepsilon,u_{\lambda,S}\rangle\right|\left\|\Pi_S\hat{f}_\lambda-f\right\|-\sigma^2\,\mathfrak{pen}(S)\\
&\quad\leq\quad\left\|f-\Pi_S\hat{f}_\lambda\right\|^2+\alpha\left\|\hat{f}_\lambda-\Pi_S\hat{f}_\lambda\right\|^2+2\sigma^2\,\mathfrak{pen}(S)\\
&\qquad+K^{-1}\left\|f-\Pi_{S(\hat{\lambda})}\hat{f}_{\hat{\lambda}}\right\|^2+K\left\|\Pi_{\bar{S}(\hat{\lambda})}\varepsilon\right\|^2-\sigma^2\,\mathfrak{pen}(S(\hat{\lambda}))\\
&\qquad+K^{-1}\left\|f-\Pi_S\hat{f}_\lambda\right\|^2+K\left\|\Pi_{\bar{S}}\varepsilon\right\|^2-\sigma^2\,\mathfrak{pen}(S)
\end{aligned}$$

Hence, by using (5) and (15) we get

$$\begin{aligned}
&(1-K^{-1})\left\|f-\Pi_{S(\hat{\lambda})}\hat{f}_{\hat{\lambda}}\right\|^2+\alpha\left\|\hat{f}_{\hat{\lambda}}-\Pi_{S(\hat{\lambda})}\hat{f}_{\hat{\lambda}}\right\|^2\\
&\quad\leq\quad(1+K^{-1})\left\|f-\Pi_S\hat{f}_\lambda\right\|^2+\alpha\left\|\hat{f}_\lambda-\Pi_S\hat{f}_\lambda\right\|^2+2\sigma^2\,\mathfrak{pen}(S)+\tilde{\Sigma}\\
&\quad\leq\quad 2(1+K^{-1})\left\|f-\hat{f}_\lambda\right\|^2\\
&\qquad+\left(\alpha+2(1+K^{-1})\right)\left\|\hat{f}_\lambda-\Pi_S\hat{f}_\lambda\right\|^2+2\sigma^2\,\mathfrak{pen}(S)+\tilde{\Sigma}
\end{aligned}\tag{16}$$

where

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

For each $S\in\mathbb{S}$,

$$\frac{\|Y-\Pi_SY\|^2}{n-\dim(S)}\geq\frac{\left\|Y-\Pi_{\overline{S}}Y\right\|^2}{n-\dim(S)}$$

and since the variable $\left\|Y - \Pi_{\overline{S}} Y\right\|^2$ is independent of $\left\|\Pi_{\overline{S}}\varepsilon\right\|^2$ and is stochastically larger than $\left\|\varepsilon - \Pi_{\overline{S}}\varepsilon\right\|^2$, we deduce from the definition of $\mathrm{pen}_\Delta(S)$ and (2), that on the one hand $\mathbb{E}(\tilde{\Sigma}) \leq 2K\sigma^2\Sigma$.

On the other hand, since S is arbitrary among $\mathbb{S}_\lambda$ and since

$$\left(\frac{1}{\alpha} + \frac{1}{1-K^{-1}}\right)^{-1} \left\|f - \hat{f}_{\hat{\lambda}}\right\|^2 \leq (1-K^{-1}) \left\|f - \Pi_{S(\hat{\lambda})}\hat{f}_{\hat{\lambda}}\right\|^2 + \alpha \left\|\hat{f}_{\hat{\lambda}} - \Pi_{S(\hat{\lambda})}\hat{f}_{\hat{\lambda}}\right\|^2$$

we deduce from (16) that for all $\lambda \in \Lambda$,

$$\left\|f - \hat{f}_{\hat{\lambda}}\right\|^2 \leq C^{-1}\left[\left\|f - \hat{f}_\lambda\right\|^2 + A(\hat{f}_\lambda, \mathbb{S}_\lambda) + \tilde{\Sigma}\right] \tag{17}$$

with

$$C^{-1} = C^{-1}(K,\alpha) = \frac{\left(1+\alpha-K^{-1}\right)\left(\alpha + 2(1+K^{-1})\right)}{\alpha(1-K^{-1})}, \tag{18}$$

and the result follows by taking the expectation on both sides of (17).

6.2. *Proof of Corollary 1.* We set $\hat{\sigma}^2_\lambda = \hat{\sigma}^2_{\widehat{S}_\lambda}$. It suffices to compute the expectation of $\mathrm{pen}(\widehat{S}_\lambda)\,\hat{\sigma}^2_\lambda$ for $\lambda \in \Lambda$. Since $\hat{f}_\lambda \in \widehat{S}_\lambda$ we have

$$\begin{aligned}
\mathrm{pen}(\widehat{S}_\lambda)\,\hat{\sigma}^2_\lambda &= K\frac{\mathrm{pen}_\Delta(\widehat{S}_\lambda)}{n - \dim(\widehat{S}_\lambda)} \left\|Y - \Pi_{\widehat{S}_\lambda} Y\right\|^2 \\
&\leq K\frac{\mathrm{pen}_\Delta(\widehat{S}_\lambda)}{n - \dim(\widehat{S}_\lambda)} \left\|Y - \hat{f}_\lambda\right\|^2 = K\frac{\mathrm{pen}_\Delta(\widehat{S}_\lambda)}{n - \dim(\widehat{S}_\lambda)} \left\|f + \varepsilon - \hat{f}_\lambda\right\|^2 \\
&\leq 2K\frac{\mathrm{pen}_\Delta(\widehat{S}_\lambda)}{n - \dim(\widehat{S}_\lambda)} \left[\left\|f - \hat{f}_\lambda\right\|^2 + \|\varepsilon\|^2\right] \\
&\leq 2K\frac{\mathrm{pen}_\Delta(\widehat{S}_\lambda)}{n - \dim(\widehat{S}_\lambda)} \left[\left\|f - \hat{f}_\lambda\right\|^2 + (\|\varepsilon\|^2 - 2n\sigma^2)_+ + 2n\sigma^2\right].
\end{aligned}$$

Under the assumption that for all $S \in \mathbb{S}$, $\Delta(S) \vee \dim(S) \leq \kappa n$, we deduce from (4) that for some constant C depending only on K and κ, we have that for all $\lambda \in \Lambda$

$$C\,\mathrm{pen}(\widehat{S}_\lambda)\,\hat{\sigma}^2_\lambda \leq \left\|f - \hat{f}_\lambda\right\|^2 + \left(\dim(\widehat{S}_\lambda) \vee \Delta(\widehat{S}_\lambda)\right)\sigma^2 + (\|\varepsilon\|^2 - 2n\sigma^2)_+.$$

The result follows from the fact that $\mathbb{E}[(\|\varepsilon\|^2 - 2n\sigma^2)_+] \leq 3\sigma^2$ for all n and $\dim(\widehat{S}_\lambda) \geq 1$ for all λ.

6.3. *Proof of Corollary 2.* Combining the equality

$$\mathbb{E}\left[\hat{\sigma}^2_{S_\lambda}\right] = \sigma^2 + \frac{\|f - \Pi_{S_\lambda} f\|^2}{n - \dim(S_\lambda)}$$

with the bounds (4) and (8) leads to

$$\begin{aligned} C\mathbb{E}\left[\left\|f - \hat{f}_{\hat{\lambda}}\right\|\right] \quad \leq \quad & \inf_{\lambda\in\Lambda}\left\{\mathbb{E}\left[\left\|f - \hat{f}_\lambda\right\|^2\right] + \mathbb{E}\left[\left\|\hat{f}_\lambda - \Pi_{S_\lambda}\hat{f}_\lambda\right\|^2\right]\right. \\ & \left. +(a\vee 1)\left[\dim(S_\lambda)\sigma^2 + \frac{\kappa}{1-\kappa}\|f - \Pi_{S_\lambda} f\|^2\right]\right\} + \sigma^2\Sigma. \end{aligned}$$

We shall bound both $\mathbb{E}\left[\|\hat{f}_\lambda - \Pi_{S_\lambda}\hat{f}_\lambda\|^2\right]$ and $\|f - \Pi_{S_\lambda} f\|^2 + \dim(S_\lambda)\sigma^2$ in terms of $\mathbb{E}\left[\|f - \hat{f}_\lambda\|^2\right]$. Since for all $\lambda \in \Lambda$ we have

$$\mathbb{E}\left[\|f - \hat{f}_\lambda\|^2\right] = \|f - A_\lambda f\|^2 + \mathbb{E}\left[\|A_\lambda\varepsilon\|^2\right] = \|f - A_\lambda f\|^2 + \mathrm{Tr}(A^*_\lambda A_\lambda)\sigma^2$$

and

$$\mathbb{E}\left[\left\|\hat{f}_\lambda - \Pi_{S_\lambda}\hat{f}_\lambda\right\|^2\right] = \|(I - \Pi_{S_\lambda})A_\lambda f\|^2 + \mathbb{E}\left[\|(I - \Pi_{S_\lambda})A_\lambda\varepsilon\|^2\right],$$

Corollary 2 follows from $\|(I - \Pi_{S_\lambda})A_\lambda\varepsilon\| \leq \|A_\lambda\varepsilon\|$ and the next lemma.

LEMMA 1. *For all $\lambda \in \Lambda$ we have*

$$\begin{aligned} &(i) \quad \|A_\lambda f - \Pi_{S_\lambda} A_\lambda f\| \leq \|f - A_\lambda f\|, \\ &(ii) \quad \|f - \Pi_{S_\lambda} f\| \leq 2\|f - A_\lambda f\|, \\ &(iii) \quad \dim(S_\lambda) \leq 4\,\mathrm{Tr}(A^*_\lambda A_\lambda). \end{aligned}$$

Proof of Lemma 1: Writing $f = f_0 + f_1 \in \ker(A_\lambda) \oplus \mathrm{rg}(A^*_\lambda)$ and using the fact that $\mathrm{rg}(A^*_\lambda) = \ker(A_\lambda)^\perp$ and the definition of $\overline{\Pi}_\lambda$, we obtain

$$\begin{aligned} \|f - A_\lambda f\|^2 &= \|f_0 + f_1 - A_\lambda f_1\|^2 \\ &= \left\|f_0 - \Pi_{\ker(A_\lambda)} A_\lambda f_1\right\|^2 + \left\|(I - \overline{\Pi}_\lambda A_\lambda) f_1\right\|^2 \\ &\geq \left\|(A^+_\lambda - \overline{\Pi}_\lambda) A_\lambda f_1\right\|^2 \\ &\geq \sum_{k=1}^{m_\lambda} s_k^2 < A_\lambda f, v_k >^2, \end{aligned}$$

where $s_1 \geq \ldots \geq s_{m_\lambda}$ are the singular values of $A^+_\lambda - \overline{\Pi}_\lambda$ counted with their multiplicity and $(v_1, \ldots, v_{m_\lambda})$ is an orthonormal family of right-singular

vectors associated to $(s_1, \ldots, s_{m_\lambda})$. We write k_λ for the largest k such that $s_k \geq 1$ and derive that

$$\begin{aligned} \|f - A_\lambda f\|^2 &\geq \sum_{k=1}^{k_\lambda} s_k^2 < A_\lambda f, v_k >^2 \\ &\geq \sum_{k=1}^{k_\lambda} < A_\lambda f, v_k >^2 = \|(I - \Pi_{S_\lambda}) A_\lambda f\|^2 , \end{aligned}$$

which proves the assertion (i).

For the second part (ii), we note that

$$\begin{aligned} \|f - \Pi_{S_\lambda} f\| &\leq \|f - \Pi_{S_\lambda} A_\lambda f\| \\ &\leq \|f - A_\lambda f\| + \|A_\lambda f - \Pi_{S_\lambda} A_\lambda f\| . \end{aligned}$$

The bound (ii) then follows from (i).

For the last bound (iii), we set $M_\lambda = A_\lambda^+ - \overline{\Pi}_\lambda$ and note that

$$(M_\lambda - \overline{\Pi}_\lambda)(M_\lambda - \overline{\Pi}_\lambda)^* = M_\lambda M_\lambda^* + \overline{\Pi}_\lambda \overline{\Pi}_\lambda^* - M_\lambda \overline{\Pi}_\lambda^* - \overline{\Pi}_\lambda M_\lambda^*$$

induces a semi-positive quadratic form on $\mathrm{rg}(A_\lambda^*)$. As a consequence the quadratic form $(M_\lambda + \overline{\Pi}_\lambda)(M_\lambda + \overline{\Pi}_\lambda)^*$ is dominated by the quadratic form $2(M_\lambda M_\lambda^* + \overline{\Pi}_\lambda \overline{\Pi}_\lambda^*)$ on $\mathrm{rg}(A_\lambda^*)$. Furthermore

$$(M_\lambda + \overline{\Pi}_\lambda)(M_\lambda + \overline{\Pi}_\lambda)^* = (A_\lambda^+)(A_\lambda^+)^* = (A_\lambda^* A_\lambda)^+$$

where $(A_\lambda^* A_\lambda)^+$ is the inverse of the linear operator $L_\lambda : \mathrm{rg}(A_\lambda^*) \to \mathrm{rg}(A_\lambda^*)$ induced by $A_\lambda^* A_\lambda$ restricted on $\mathrm{rg}(A_\lambda^*)$. We then have that the quadratic form induced by $(A_\lambda^* A_\lambda)^+$ is dominated by the quadratic form

$$2(A_\lambda^+ - \overline{\Pi}_\lambda)(A_\lambda^+ - \overline{\Pi}_\lambda)^* + 2\overline{\Pi}_\lambda \overline{\Pi}_\lambda^*$$

on $\mathrm{rg}(A_\lambda^*)$. In particular the sequence of the eigenvalues of $(A_\lambda^* A_\lambda)^+$ is dominated by the sequence $(2s_k^2 + 2)_{k=1,m_\lambda}$ so

$$\begin{aligned} \mathrm{Tr}(A_\lambda^* A_\lambda) = \mathrm{Tr}(L_\lambda) &\geq \sum_{k=1}^{m_\lambda} \frac{1}{2(1+s_k^2)} \\ &\geq \sum_{k=k_\lambda+1}^{m_\lambda} \frac{1}{2(1+s_k^2)} \geq \dim(S_\lambda)/4, \end{aligned}$$

which conclude the proof of Lemma 1.

References.

Arlot, S. (2007). *Rééchantillonnage et Sélection de modèles*. PhD thesis, University Paris XI.

Arlot, S. (2009). Model selection by resampling penalization. *Electron. J. Stat.*, 3:557–624.

Arlot, S. and Bach, F. (2009). Data-driven calibration of linear estimators with minimal penalties. *Advances in Neural Information Processing Systems (NIPS)*, 22:46–54.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79.

Baraud, Y. (2000). Model selection for regression on a fixed design. *Probab. Theory Related Fields*, 117(4):467–493.

Baraud, Y. (2010). Estimator selection with respect to Hellinger-type risks. *Probab. Theory Relat. Fields*.

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

Birgé, L. (2006). Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *Ann. Inst. H. Poincaré Probab. Statist.*, 42(3):273–325.

Birgé, L. and Massart, P. (2001). Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268.

Boulesteix, A. and Strimmer, K. (2006). Partail least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics*, 8(1):32–44.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Bunea, F., Tsybakov, A. B., and Wegkamp, M. H. (2007). Aggregation for Gaussian regression. *Ann. Statist.*, 35(4):1674–1697.

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

Cao, Y. and Golubev, Y. (2006). On oracle inequalities related to smoothing splines. *Math. Methods Statist.*, 15(4):398–414 (2007).

Catoni, O. (1997). Mixture approach to universal model selection. Technical report, Ecole Normale Supérieure, France.

Catoni, O. (2004). Statistical learning theory and stochastic optimization. In *Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001*. Springer-Verlag, Berlin.

Celisse, A. (2008). *Model selection via cross-validation in density estimation, regression, and change-points detection*. PhD thesis, University Paris XI.

Chen, S. S., Donoho, D. L., and Saunders, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61 (electronic).

Díaz-Uriarte, R. and Alvares de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(3).

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.

Genuer, R., Poggi, J.-M., and Tuleau-Malot, C. (2010). Variable selection using random forests. *Patter Recognition Lett.*, to appear.

Giraud, C. (2008). Mixing least-squares estimators when the variance is unknown. *Bernoulli*, 14(4):1089–1107.

Goldenshluger, A. (2009). A universal procedure for aggregating estimators. *Ann. Statist.*, 37(1):542–568.

Goldenshluger, A. and Lepski, O. (2009). Structural adaptation via $\mathbb{L}_p$-norm oracle inequalities. *Probab. Theory Related Fields*, 143(1-2):41–71.

Helland, I. (2006). Partial least squares regression. In Kotz, S., Balakrishnan, N., Read,

C., Vidakovic, B., and Johnston, N., editors, *Encyclopedia of statistical sciences (2nd ed.)*, volume 9, pages 5957–5962, New York. Wiley.

Hoerl, A. and Kennard, R. (2006). Ridge regression. In Kotz, S., Balakrishnan, N., Read, C., Vidakovic, B., and Johnston, N., editors, *Encyclopedia of statistical sciences (2nd ed.)*, volume 11, pages 7273–7280, New York. Wiley.

Huang, J., Ma, S., and Zhang, C.-H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica*, 4(1603-1618).

Juditsky, A. and Nemirovski, A. (2000). Functional aggregation for nonparametric regression. *Ann. Statist.*, 28(3):681–712.

Lepskiĭ, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. *Teor. Veroyatnost. i Primenen.*, 35(3):459–470.

Lepskiĭ, O. V. (1991). Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Teor. Veroyatnost. i Primenen.*, 36(4):645–659.

Lepskiĭ, O. V. (1992a). Asymptotically minimax adaptive estimation. II. Schemes without optimal adaptation. Adaptive estimates. *Teor. Veroyatnost. i Primenen.*, 37(3):468–481.

Lepskiĭ, O. V. (1992b). On problems of adaptive estimation in white Gaussian noise. In *Topics in nonparametric estimation*, volume 12 of *Adv. Soviet Math.*, pages 87–106. Amer. Math. Soc., Providence, RI.

Leung, G. and Barron, A. R. (2006). Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory*, 52(8):3396–3410.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications*, 9(1):141—142.

Nemirovski, A. (2000). Topics in non-parametric statistics. In *Lectures on probability theory and statistics (Saint-Flour, 1998)*, volume 1738 of *Lecture Notes in Math.*, pages 85–277. Springer, Berlin.

Rigollet, P. and Tsybakov, A. B. (2007). Linear and convex aggregation of density estimators. *Math. Methods Statist.*, 16(3):260–280.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, 9(307).

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(25).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288.

Tsybakov, A. B. (2003). Optimal rates of aggregation. In *Proceedings of the 16th Annual Conference on Learning Theory (COLT) and 7th Annual Workshop on Kernel Machines*, pages 303–313. Lecture Notes in Artificial Intelligence 2777, Springer-Verlag, Berlin.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā Ser. A*, 26:359–372.

Wegkamp, M. (2003). Model selection in nonparametric regression. *Ann. Statist.*, 31:252–273.

Yang, Y. (1999). Model selection for nonparametric regression. *Statist. Sinica*, 9:475–499.

Yang, Y. (2000a). Combining different procedures for adaptive regression. *J. Multivariate Anal.*, 74(1):135–161.

Yang, Y. (2000b). Mixing strategies for density estimation. *Ann. Statist.*, 28(1):75–87.

Yang, Y. (2001). Adaptive regression by mixing. *J. Amer. Statist. Assoc.*, 96(454):574–588.

Zhang, T. (2005). Learning bounds for kernel regression using effective data dimensionality. *Neural Comput.*, 17(9):2077–2098.

Zhang, T. (2008). Adaptive forward-backward greedy algorithm for learning sparse representations. Technical report, Rutgers University, NJ.
Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476):1418–1429.
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.

7. Appendix.

7.1. *Computation of* $\mathrm{pen}_\Delta(S)$. The penalty $\mathrm{pen}_\Delta(S)$, defined at equation (3), is linked to the EDKhi function introduced in Baraud *al* (2009) (see Definition 3), via the following formula:

$$\mathrm{pen}_\Delta(S) = \frac{n-\dim(S)}{n-\dim(S)-1}\mathrm{EDKhi}\left(\dim(S)+1, n-\dim(S)-1, \frac{e^{-\Delta(S)}}{\dim(S)+1}\right).$$

Therefore, according to the result given in Section 6.1 in Baraud *et al* (2009), $\mathrm{pen}_\Delta(S)$ is the solution in x of the equation

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

7.2. *Simulated examples.* The collection $\mathcal{E}$ is composed of several collections $\mathcal{E}_1, \ldots, \mathcal{E}_{11}$ that are detailed below. The collections $\mathcal{E}_1$ to $\mathcal{E}_{10}$ are composed of examples where X is generated as n independent centered Gaussian vectors with covariance matrix C. For each $e \in \{1, \ldots, 10\}$, we define a $p \times p$ matrix C_e and a p-vector of parameters β_e. We denote by $\mathcal{X}_e$ the set of 5 matrices X simulated as n-i.i.d $\mathcal{N}_p(0, C_e)$. The collection $\mathcal{E}_e$ is then defined as follows:

$$\mathcal{E}_e = \{\mathrm{ex}(n,p,X,\beta,\rho), (n,p) \in \mathcal{I}, X \in \mathcal{X}_e, \beta = \beta_e, \rho \in \mathcal{R}\}$$

where $\mathcal{R} = \{5, 10, 20\}$ and

$$\mathcal{I} = \{(100,50), (100,100), (100,1000), (200,100), (200,200)\} \tag{19}$$

in Section 5.2, and

$$\mathcal{I} = \{(100,50), (100,100), (200,100), (200,200)\} \tag{20}$$

in Section 5.3.

Let us now describe the collections $\mathcal{E}_1$ to $\mathcal{E}_{10}$.

Collection $\mathcal{E}_1$. The matrix C equals the $p \times p$ identity matrix denoted I_p. The parameters β satisfy $\beta_j = 0$ for $j \geq 16$, $\beta_j = 2.5$ for $1 \leq j \leq 5$, $\beta_j = 1.5$ for $6 \leq j \leq 10$, $\beta_j = 0.5$ for $11 \leq j \leq 15$.

Collection $\mathcal{E}_2$. the matrix C is such that $C_{jk} = r^{|j-k|}$, for $1 \leq j, k \leq 15$ and $16 \leq j, k \leq p$ with $r = 0.5$. Otherwise $C_{j,k} = 0$. The parameters β are as in Collection $\mathcal{E}_1$.

Collection $\mathcal{E}_3$. The matrix C is as in Collection $\mathcal{E}_2$ with $r = 0.95$, the parameters β are as in Collection $\mathcal{E}_1$.

Collection $\mathcal{E}_4$. The matrix C is such that $C_{jk} = r^{|j-k|}$, for $1 \leq j, k \leq p$, with $r = 0.5$, the parameters β are as in Collection $\mathcal{E}_1$.

Collection $\mathcal{E}_5$. the matrix C is as in Collection $\mathcal{E}_4$ with $r = 0.95$, the parameters β are as in Collection $\mathcal{E}_1$.

Collection $\mathcal{E}_6$. The matrix C equals I_p. The parameters β satisfy $\beta_j = 0$ for $j \geq 16$, $\beta_j = 1.5$ for $j \leq 15$.

Collection $\mathcal{E}_7$. The matrix C satisfies $C_{j,k} = (1 - \rho_1)\mathbb{1}_{j=k} + \rho_1$ for $1 \leq , j, k \leq 3$, $C_{j,k} = C_{k,j} = \rho_2$ for $j = 4, k = 1, 2, 3$, $C_{j,k} = \mathbb{1}_{j=k}$ for $j, k \geq 5$, with $\rho_1 = .39$ and $\rho_2 = .23$. The parameters β satisfy $\beta_j = 0$ for $j \geq 4$, $\beta_j = 5.6$ for $j \leq 3$.

Collection $\mathcal{E}_8$. The matrix C satisfies $C_{j,k} = 0.5^{|j-k|}$ for $j, k \leq 8$, $C_{j,k} = \mathbb{1}_{j=k}$ for $j, k \geq 9$. The parameters β satisfy $\beta_j = 0$ for $j \notin \{1, 2, 5\}$, $\beta_1 = 3$, $\beta_2 = 1.5$, $\beta_5 = 2$.

Collection $\mathcal{E}_9$. The matrix C is defined as in Example $\mathcal{E}_8$. The parameters β satisfy $\beta_j = 0$ for $j \geq 9$, $\beta_j = 0.85$ for $j \leq 8$.

Collection $\mathcal{E}_{10}$. The matrix C satisfies $C_{j,k} = 0.5\mathbb{1}_{j \neq k} + \mathbb{1}_{j=k}$ for $j, k \leq 40$, $C_{j,k} = \mathbb{1}_{j=k}$ for $j, k \geq 41$. The parameters β satisfy $\beta_j = 2$ for $11 \leq j \leq 20$ and $31 \leq j \leq 40$, $\beta_j = 0$ otherwise.

Collection $\mathcal{E}_{11}$. In this last example, we denote by $\mathcal{X}_{11}$ the set of 5 matrices X simulated as follows. For $1 \leq j \leq p$, we denote by X_j the column j of X. Let E be generated as n i.i.d. $\mathcal{N}_p(0, 0.01 I_p)$ and let Z_1, Z_2, Z_3 be generated as n i.i.d. $\mathcal{N}_3(0, I_3)$. Then for $j = 1, \ldots, 5$, $X_j = Z_1 + E_j$, for $j = 6, \ldots, 10$, $X_j = Z_2 + E_j$, for $j = 11, \ldots, 15$, $X_j = Z_3 + E_j$, for $j \geq 16$, $X_j = E_j$. The parameters β are as in Collection $\mathcal{E}_6$. The collection $\mathcal{E}_{11}$ is defined as the set of examples $\mathrm{ex}(n, p, X, \beta, \rho)$ for $(n, p) \in \mathcal{I}$, $X \in \mathcal{X}_{11}$, and $\rho \in \mathcal{R}$.

The collection $\mathcal{E}$ is thus composed of 660 examples for $\mathcal{I}$ chosen as in (20), and 825 for $\mathcal{I}$ chosen as in (19). For some of the examples, the Lasso esti-

mators were highly biased leading to high values of the ratio $O_{\mathrm{ex}}/n\sigma^2$, see Equation (14). We only keep the examples for which the Lasso estimator improves the risk of the naive estimator Y by a factor at least $1/3$. This convention leads us to remove 171 examples over 825. These pathological examples are coming from the collections $\mathcal{E}_1$, $\mathcal{E}_6$ and $\mathcal{E}_7$ for $n = 100$ and $p \geq 100$, and from collections $\mathcal{E}_2$ and $\mathcal{E}_4$ when $p = 1000$. The examples of collection $\mathcal{E}_7$ were chosen by Zou to illustrate that the Lasso estimators may be highly biased. All the other examples, correspond to matrices X that are nearly orthogonal.

7.3. *Procedures for calculating sets of predictors.* Let $\widehat{\mathcal{M}} = \bigcup_{\ell \in \mathcal{L}} \widehat{\mathcal{M}}_\ell$ where we recall that for $\ell \in \mathcal{L}$, $\widehat{\mathcal{M}}_\ell = \{\widehat{m}(\ell, h) |\ h \in H_\ell\}$.

The Lasso procedure is described in Section 5.2. The collection $\widehat{\mathcal{M}}_{\mathrm{Lasso}} = \{\widehat{m}(1), \ldots, \widehat{m}(D_{\max})\}$ where $\widehat{m}(h)$ is the set of indices corresponding to the predictors retuned by the LARS-Lasso algorithm at step $h \in \{1, \ldots, D_{\max}\}$ (see Section 5.2).

The ridge procedure is based on the minimization of $\|Y - X\beta\|^2 + h\|\beta\|^2$ with respect to β, for some positive h, see for example Hoerl and Kennard (2006). Tibshirani (1996) noted that in the case of a large number of small effects, ridge regression gives better results than the lasso for variable selection. For each $h \in H_{\mathrm{ridge}}$, the regression coefficients $\widehat{\beta}(h)$ are calculated and a collection of predictors sets is built as follows. Let $j_1, \ldots j_p$ be such that $|\widehat{\beta}_{j_1}(h)| > \ldots > |\widehat{\beta}_{j_p}(h)|$ and set

$$M_h = \{\{j_1, \ldots, j_k\},\ k = 1, \ldots, D_{\max}\}\,.$$

Then, the collection $\widehat{\mathcal{M}}_{\mathrm{ridge}}$ is defined as $\widehat{\mathcal{M}}_{\mathrm{ridge}} = \{M_h, h \in H_{\mathrm{ridge}}\}$.

The elastic net procedure proposed by Zou and Hastie (2005) mixes the ℓ_1 and ℓ_2 penalties of the Lasso and the ridge procedures. Let H_{ridge} be a grid a values for the tuning parameter h of the ℓ_2 penalty. We choose $\widehat{\mathcal{M}}_{\mathrm{en}} = \{M_{(\mathrm{en},h)} : h \in H_{\mathrm{ridge}}\}$ where $M_{(\mathrm{en},h)}$ denotes the collection of the active sets of cardinality less than $D_{\max}$, selected by the elastic net procedure when the ℓ_2-smoothing parameter equals h. For each $h \in H_{\mathrm{ridge}}$ the collection $M_{(\mathrm{en},h)}$ can be conveniently computed by first calculating the ridge regression coefficients and then applying the LARS-lasso algorithm, see Zou and Hastie (2005).

The partial least squares regression (PLSR1) aims to reduce the dimensionality of the regression problem by calculating a small number of components that are usefull for predicting Y. Several applications of this procedure for

analysing high-dimensional genomic data have been reviewed by Boulesteix and Strimmer (2006). In particular, it can be used for calculating subsets of covariates as we did for the ridge procedure. The PLSR1 procedure constructs, for a given h, uncorrelated latent components $t_1, \ldots, t_h$ that are highly correlated with the response Y, see Helland (2006). Let H_{pls} be a grid a values for the tuning parameter h. For each $h \in H_{\text{pls}}$, we write $\widehat{\beta}(h)$ for the PLS regression coefficients calculated with the first h components. We then set $\widehat{\mathcal{M}}_{\text{PLS}} = \{M_h : h \in H_{\text{pls}}\}$, where M_h is build from $\widehat{\beta}(h)$ as for the ridge procedure.

The adaptive lasso procedure proposed by Zou (2006) starts with a preliminary estimator $\widetilde{\beta}$. Then one applies the lasso procedure replacing the parameters $|\beta_j|, j = 1, \ldots, p$ in the ℓ_1 penalty by the weighted parameters $|\beta_j|/|\tilde{\beta}_j|^\gamma, j = 1, \ldots, p$ for some positive γ. The idea is to increase the penalty for coefficients that are close to zero, reducing thus the bias in the estimation of f and improving the variable selection accuracy. Zou showed that, if $\widetilde{\beta}$ is a $\sqrt{n}$-consistent estimator of β, then the adaptive lasso procedure is consistent in situations where the lasso is not. A lot of work has been done around this subject, see Huang et al. (2008) for example.

We apply the procedure with $\gamma = 1$, and considering two different preliminary estimators:

- using the ridge estimator, $\widetilde{\beta}(h)$ as preliminary estimator. For each $h \in H_{\text{ridge}}$, the adaptive lasso procedure is applied for calculating the active sets, $M_{\text{ALridge},h}$, of cardinality less than $D_{\max}$. The collection $\widehat{\mathcal{M}}_{\text{ALridge}}$ is thus defined as $\widehat{\mathcal{M}}_{\text{ALridge}} = \{M_{\text{ALridge},h}, h \in H_{\text{ridge}}\}$.

- using the PLSR1 estimator, $\widetilde{\beta}(h)$, as preliminary estimator. The procedure is the same as described just above. The collection M_{ALpls} is defined as $M_{\text{ALpls}} = \{M_{\text{ALpls},h}, h \in H_{\text{pls}}\}$.

The random forest algorithm was proposed by Breiman (2001) for classification and regression problems. The procedure averages several regression trees calculated on bootstrap samples. The algorithm returns measures of variable importance that may be used for variable selection, see for example Díaz-Uriarte and Alvares de Andrés (2006), Genuer et al. (2010), Strobl et al. (2007; 2008).

Let us denote by h the number of variables randomly chosen at each split when constructing the trees and

$$H_{rF} = \{p/j \mid j \in \{3, 2, 1.5, 1\}\}.$$

For each $h \in H_{rF}$, we consider the set of indices

$$M_h = \{\{j_1, \ldots, j_k\}, k = 1, \ldots, D_{\max}\},$$

where $\{j_1, \ldots, j_k\}$ are the ranks of the variable importance measures. Two importance measures are proposed. The first one is based on the decrease in the mean square error of prediction after permutation of each of the variables. It leads to the collection $\widehat{\mathcal{M}}_{\text{rFmse}} = \{M_h, h \in H_{rF}\}$. The second one is based on the decrease in node impurities, and leads similarly to the collection $\widehat{\mathcal{M}}_{\text{purity}}$.

The exhaustive procedure considers the collection of all subsets of $\{1, \ldots p\}$ with dimension smaller than $D_{\max}$. We denote this collection $\mathcal{M}_{\text{exhaustive}}$.

Choice of tuning parameters. We have to choose $D_{\max}$, the largest number of predictors considered in the collection $\widehat{\mathcal{M}}$. For all methods, except the exhaustive method, $D_{\max}$ may be large, say $D_{\max} \leq \min(n-2, p)$. Nevertheless, for saving computing time, we chose $D_{\max}$ large enough such that the dimension of the estimated subset is always smaller than $D_{\max}$. For the exhaustive method, $D_{\max}$ must be chosen in order to make the calculation feasible: $D_{\max} = 4$ for $p = 50$, $D_{\max} = 3$ for $p = 100$ and $D_{\max} = 2$ for $p = 200$.

For the ridge method we choose $H_{\text{ridge}} = \{10^{-3}, 10^{-2}, 10^{-1}, 1, 5\}$, and for the PLSR1 method, $H_{\text{pls}} = 1, \ldots, 5$.

Université de Nice Sophia-Antipolis,
Laboratoire J-A Dieudonné, UMR CNRS 6621
Parc Valrose
06108, Nice cedex 02
France
E-mail: baraud@unice.fr

Ecole Polytechnique,
CMAP, UMR CNRS 7641
route de Saclay
91128 Palaiseau Cedex
France
E-mail: christophe.giraud@polytechnique.edu

INRA MIAJ
78352, Jouy en Josas cedex
France
E-mail: sylvie.huet@jouy.inra.fr