Annales de l'Institut Henri Poincaré - Probabilités et Statistiques
2014, Vol. 50, No. 3, 1092–1119
DOI: 10.1214/13-AIHP539

ANNALES
DE L'INSTITUT
HENRI
POINCARÉ
PROBABILITÉS
ET STATISTIQUES
www.imstat.org/aihp

Estimator selection in the Gaussian setting

Yannick Baraud[a], Christophe Giraud[b] and Sylvie Huet[c]

[a] *Laboratoire J. A. Dieudonné UMR CNRS 7351, Université de Nice Sophia-Antipolis, Parc Valrose, 06108 Nice cedex 02, France. E-mail: baraud@unice.fr*
[b] *CMAP, UMR CNRS 7641, Ecole Polytechnique, Route de Saclay, 91128 Palaiseau Cedex, France. E-mail: christophe.giraud@polytechnique.edu*
[c] *UR341 MIA, INRA, F-78350 Jouy-en-Josas, France. E-mail: sylvie.huet@jouy.inra.fr*

Received 10 November 2011; revised 21 December 2012; accepted 8 January 2013

Abstract. We consider the problem of estimating the mean f of a Gaussian vector Y with independent components of common unknown variance σ^2. Our estimation procedure is based on estimator selection. More precisely, we start with an arbitrary and possibly infinite collection $\mathbb{F}$ of estimators of f based on Y and, with the same data Y, aim at selecting an estimator among $\mathbb{F}$ with the smallest Euclidean risk. No assumptions on the estimators are made and their dependencies with respect to Y may be unknown. We establish a non-asymptotic risk bound for the selected estimator and derive oracle-type inequalities when $\mathbb{F}$ consists of linear estimators. As particular cases, our approach allows to handle the problems of aggregation, model selection as well as those of choosing a window and a kernel for estimating a regression function, or tuning the parameter involved in a penalized criterion. In all theses cases but aggregation, the method can be easily implemented. 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.

Résumé. Nous présentons une nouvelle procédure de sélection d'estimateurs pour estimer l'espérance f d'un vecteur Y de n variables gaussiennes indépendantes dont la variance est inconnue. Nous proposons de choisir un estimateur de f, dont l'objectif est de minimiser le risque l_2, dans une collection arbitraire et éventuellement infinie $\mathbb{F}$ d'estimateurs. La procédure de choix ainsi que la collection $\mathbb{F}$ ne dépendent que des seules observations Y. Nous calculons une borne de risque, non asymptotique, ne nécessitant aucune hypothèse sur les estimateurs dans $\mathbb{F}$, ni la connaissance de leur dépendance en Y. Nous calculons des inégalités de type "oracle" quand $\mathbb{F}$ est une collection d'estimateurs linéaires. Nous considérons plusieurs cas particuliers : estimation par aggrégation, estimation par sélection de modèles, choix d'une fenêtre et du paramètre de lissage en régression fonctionnelle, choix du paramètre de régularisation dans un critère pénalisé. Pour tous ces cas particuliers, sauf pour les méthodes d'aggrégation, la méthode est très facile à programmer. A titre d'illustration nous montrons des résultats de simulations avec deux objectifs : comparer notre méthode à la procédure de cross-validation, montrer comment la mettre en œuvre dans le cadre de la sélection de variables.

MSC: 62J05; 62J07; 62G08

Keywords: Estimator selection; Model selection; Variable selection; Linear estimator; Kernel estimator; Ridge regression; Lasso; Elastic net; Random forest; PLS1 regression

1. Introduction

1.1. *The setting and the approach*

We consider the Gaussian regression framework

$$Y_i = f_i + \varepsilon_i, \quad i = 1, \dots, n,$$

where $f = (f_1, \dots, f_n)$ 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 which corresponds to the practical case. Our aim is to estimate f from the observation of Y and we shall use the squared Euclidean norm of $\mathbb{R}^n$ denoted $\|\cdot\|^2$ as a loss function. 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*

$$f = \bigl(F(x_1), \dots, F(x_n)\bigr), \tag{1.1}$$

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

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

$$f = X\beta, \tag{1.2}$$

where X is a non-random $n \times p$ 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* (1.2) *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 classical issues. The latter is called variable selection.*

Our estimation strategy is based on estimator selection. More precisely, we start with an arbitrary collection $\mathbb{F} = \{\widehat{f_\lambda}, \lambda \in \Lambda\}$ 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 $\widehat{f_\lambda}$ depend on Y may be arbitrary and possibly unknown. For example, the $\widehat{f_\lambda}$ may be obtained from the minimization of a criterion, a Bayesian procedure or the guess of some experts.

1.2. *The motivation*

The problem of choosing some best estimator among a family of candidate ones is central in Statistics. 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* [46] *and Chen et al.* [19]) *defined by $\widehat{f_\lambda} = X\widehat{\beta_\lambda}$ with*

$$\widehat{\beta_\lambda} = \arg\min_{\beta \in \mathbb{R}^p} \left[\|Y - X\beta\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right] \tag{1.3}$$

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

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 a procedure based on ridge regression (see Hoerl and Kennard [29]), random forest (see Breiman [12]) or PLS (see Tenenhaus [45], Helland [28] and Helland [27]). 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 kernel may be 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* [39] *and Watson* [48]) *$\widehat{f}_{K,h} \in \mathbb{R}^n$ is defined as*

$$\widehat{f}_{K,h} = \bigl(\widehat{F}_{K,h}(x_1), \dots, \widehat{F}_{K,h}(x_n)\bigr),$$

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

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

There exist many possible choices for the kernel K, *such as the Gaussian kernel* $K(x) = \mathrm{e}^{-x^2/2}$, *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* $\mathbb{F} = \{\widehat{f}_\lambda, \lambda = (K,h) \in \mathcal{K} \times \mathcal{H}\}$.

1.3. *A look at the literature*

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 their theoretical performances. For more details, we refer to Arlot and Celisse [4] for a survey on cross-validation techniques applied to model selection. Compared to these approaches, as we shall see, the procedure we propose may be less time consuming (in the context of Example 3, a numerical comparison can be found at the end of Section A.2). Moreover, it does not require to know how the estimators depend on the data Y and we can therefore handle the following problem.

Example 5 (Selecting among mute experts). *A statistician is given a collection* $\mathbb{F} = \{\widehat{f}_\lambda, \lambda \in \Lambda\}$ *of estimators from a family* Λ *of experts* λ, *each of which keeping secret the way his/her estimator* $\widehat{f}_\lambda$ *depends on the observation* Y. *The problem is 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 where the estimators are indexed by a non-random collection of models, and 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 subspaces of $\mathbb{R}^n$ chosen accordingly to the problem at hand and one associates to each model $S \in \mathbb{S}$ a suitable estimator $\widehat{f}_S$ with values in S. Selecting a model then amounts to selecting an estimator among the collection $\mathbb{F} = \{\widehat{f}_S, S \in \mathbb{S}\}$. For this problem, selection rules based on the minimization of a penalized criterion have been proposed in the regression setting by Yang [50], Baraud [5], Birgé and Massart [10] and Baraud et al. [7]. Another way, usually called Lepski's method, appears in a series of papers by Lepski [33–36] and was originally designed to perform model selection among collections of nested models. Finally, other procedures based on resampling have interestingly emerged from the work of Arlot [1,2] and Célisse [18]. An unattractive 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 convex, linear or sparse 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 [16,17], Juditsky and Nemirovski [32], Nemirovski [40], Yang [51–53], Tsybakov [47], Wegkamp [49], Birgé [9], Rigollet and Tsybakov [41], Bunea, Tsybakov and Wegkamp [13] and Goldenshluger [25] 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 procedure differs from classical *aggregation* procedures since we use the whole data Y to build and select. In the Gaussian regression setting that is considered here, we mention the results of Leung and Barron [37] for the problem of mixing least-squares estimators, and of Salmon and Dalalyan [42] for the case of affine estimators. Their procedures use the same data Y to estimate and to aggregate but require the variance to be known. Giraud [23] extends the results of [37] to the case where it is unknown.

1.4. *What is new here?*

Our approach for solving the problem of estimator selection is new. We introduce a collection $\mathbb{S}$ of linear subspaces of $\mathbb{R}^n$ for approximating the estimators in $\mathbb{F}$ and use a penalized criterion to compare them. As already mentioned and as we shall see, this approach requires no assumption on the family of estimators at hand. The general way of comparing estimators described in Baraud [6] has influenced the present paper and the flavor of our results are akin to those presented there. However, the procedure proposed in Baraud [6] was mainly abstract and inadequate in the Gaussian framework we consider.

We prove a non-asymptotic risk bound for the estimator we select and show that this bound is optimal in the sense that it essentially cannot be improved (except for numerical constants maybe) by any other selection rule.

For the sake of illustration and comparison, we apply our procedure to various problems among which model selection, variable selection and selection among linear estimators. In each of these cases, our approach allows to recover classical results in the area as well as to establish new ones. Let us give an account of those results. In the context of selecting some best estimator among a family of linear ones, we propose a new procedure and show that the selected estimator satisfies an oracle-type inequality. This result requires very few assumptions on the family at hand. In the context of variable selection, our approach provides a way of selecting a suitable variable selection procedure among a family of candidate ones. For practical issues, our method is easy to implement, an R-package being available on http://w3.jouy.inra.fr/unites/miaj/public/perso/SylvieHuet_en.html. We propose thus an alternative to the well-known cross-validation scheme that is largely used but for which little is known from a theoretical point of view.

We also consider the aggregation method, focusing on linear, convex and model selection aggregation problems. For each of these, we propose a procedure which does not assume that the variance is known. We prove that the resulting estimator satisfies risk bounds which are similar (up to constants) to those obtained in Bunea, Tsybakov and Wegkamp [13] when the variance is known. Besides, our approach allows to relax the assumption that the components of the vector f as well as those of the preliminary estimators are uniformly bounded.

Finally, since the first version of this paper [8] a few papers have also addressed the specific problem of selecting the parameter λ of the Lasso estimator (1.3) when the variance σ^2 is unknown. We refer to Giraud, Huet and Verzelen [24] for a review of these procedures.

The paper is organized as follows. In Section 2 we present our selection rule and the theoretical properties of the resulting estimator. We show in Section 3 how the procedure can be used to select a linear estimator among a collection of candidate ones. In particular, we provide an oracle risk bound for the problem of selecting among a continuous family of kernel ridge estimators. In Section 4, we show how to solve the problem of variable selection, and illustrate in Section 4.3 how our procedure performs on the basis of two simulation studies. One aims at comparing the performance of our procedure to the classical V-fold in view of selecting a tuning parameter among a grid. The other aims at comparing the performance of the variable selection procedure we propose to some classical ones such as the Lasso, random forest, and others based on ridge and PLS regression.

Finally, Section 5 shows how the procedure can be used to aggregate preliminary estimators and Section 6 is devoted to the proofs.

Throughout the paper, $|A|$ denotes the cardinality of a finite set A and C, C', C'' are constants that may vary from line to line.

2. The procedure and the main result

2.1. *The procedure*

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 data-driven) 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 for illustration describe them for the particular case of tuning the parameter in the Lasso procedure as described in Example 3. More examples are given in Sections 3, 4 and 5 in view of handling other statistical problems.

2.1.1. *The collection of estimators* $\mathbb{F}$

The collection $\mathbb{F} = \{\widehat{f}_\lambda, \lambda \in \Lambda\}$ can be arbitrary. In particular, $\mathbb{F}$ need not be finite nor countable and it may consist of a mix of estimators based on the minimization of a criterion, a Bayes procedure or the guess of some experts. The dependency of these estimators with respect to Y need not be known. Nevertheless, we shall see on examples how we can use this information, when available, to improve the performance of our selection rule.

2.1.2. *The families $\mathbb{S}$ and $\mathbb{S}_\lambda$*
Let $\mathbb{S}$ be a family of linear subspaces of $\mathbb{R}^n$ satisfying the following.

Assumption 1. *The family $\mathbb{S}$ is finite or countable and for all $S \in \mathbb{S}$, $\dim(S) \le n - 2$.*

The restriction on the dimensions of the linear subspaces S is only due to the fact the we do not assume that the variance σ^2 is known.

To each estimator $\widehat{f_\lambda} \in \mathbb{F}$, we associate a (possibly data-driven) subset $\mathbb{S}_\lambda \subset \mathbb{S}$.

There is no universal choice for the collection $\mathbb{S}$. It should depend on the statistical context (signal estimation, change point problem, variable selection, etc.) and, when available, on the structure of the estimators lying in $\mathbb{F}$. Typically, the family $\mathbb{S}$ should be chosen to possess good approximation properties with respect to the elements of $\mathbb{F}$. For each λ, $\mathbb{S}_\lambda$ should approximate $\widehat{f_\lambda}$ more specifically. One may take $\mathbb{S}_\lambda = \mathbb{S}$ but for computational reasons it will be convenient to allow $\mathbb{S}_\lambda$ to be smaller.

We provide examples of $\mathbb{S}$ and $\mathbb{S}_\lambda$ in various statistical settings described in Sections 4 to 5.

Example 3 (continued). *Let $\mathbb{F}$ be the family of Lasso estimators $\widehat{f_\lambda} = X\widehat{\beta_\lambda}$ corresponding to the values of λ for which $|\widehat{\beta_\lambda}|_0 = |\{i = \{1, \dots, p\}, (\widehat{\beta_\lambda})_i \neq 0\}|$ is not larger than some $D_{\max} \le n - 2$. This amounts to considering the family of $\widehat{f_\lambda}$ associated to λ that are large enough or equivalently to dealing with the family of (modified) Lasso estimators indexed by $\Lambda = \mathbb{R}_+$ and defined by $\widehat{f_\lambda} = X\widehat{\beta_\lambda}$ when $|\widehat{\beta_\lambda}|_0 \le D_{\max}$ and $\widehat{f_\lambda} = 0$ otherwise. Denoting by X_j the jth column of X, we choose $\mathbb{S}$ as the family gathering all the linear spans of $\{X_j, j \in m\}$ when m varies among all the subsets of $\{1, \dots, p\}$ satisfying $|m| \le D_{\max}$ (with the convention that the linear span generated by the empty set is $\{0\}$). We more specifically associate to each estimator $\widehat{f_\lambda} \in \mathbb{F}$, the subfamily $\mathbb{S}_\lambda$ reduced to S_λ where S_λ is the (random) linear span of the columns j of X for which $(\widehat{\beta_\lambda})_j \neq 0$. With such choices, $\widehat{f_\lambda} \in S_\lambda$ for all $\lambda \in \Lambda$ and the approximations of the $\widehat{f_\lambda}$ by the S_λ are therefore perfect.*

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.

$$\Sigma = \sum_{S \in \mathbb{S}} \mathrm{e}^{-\Delta(S)} < +\infty. \tag{2.1}$$

Whenever $\mathbb{S}$ is finite, inequality (2.1) automatically holds true. However, in practice Σ should be kept to a reasonable size. When $\Sigma = 1$, $\mathrm{e}^{-\Delta(\cdot)}$ can be interpreted as a prior distribution on $\mathbb{S}$ and gives thus a Bayesian flavor to the procedure we propose. Following the work of Baraud et al. [7], we associate to the weight function Δ, the function pen_Δ mapping $\mathbb{S}$ into $\mathbb{R}_+$ and defined by

$$\mathbb{E}\left[\left(U - \frac{\mathrm{pen}_\Delta(S)}{n - \dim(S)} V\right)_+\right] = \mathrm{e}^{-\Delta(S)}, \tag{2.2}$$

where x_+ denotes the positive part of $x \in \mathbb{R}$ and U, V are two independent χ^2 random variables with respectively $\dim(S) + 1$ and $n - \dim(S) - 1$ degrees of freedom. This function can be easily computed from the quantiles of the Fisher distribution as we shall see in Section A.1. From a more theoretical point of view, it is shown in Baraud et al. [7] that under Assumption 3 below, there exists a positive constant C (depending on κ only) such that

$$\mathrm{pen}_\Delta(S) \le C\bigl(\dim(S) \vee \Delta(S)\bigr). \tag{2.3}$$

This upper bound is sharp (up to numerical constants). A lower bound of the same order is established in Giraud et al. [24] (Lemma D.3).

Assumption 3. *The collection $\mathbb{S}$ is finite and there exists $\kappa \in (0, 1)$ such that for all $S \in \mathbb{S}$,*

$$1 \le \dim(S) \vee \Delta(S) \le \kappa n.$$

2.1.4. *The selection criterion*

The selection procedure we propose involves a penalty function pen from $\mathbb{S}$ into $\mathbb{R}_+$ with the following property.

Assumption 4. *The penalty function* pen *satisfies for some* $K > 1$,

$$\operatorname{pen}(S) \ge K \operatorname{pen}_{\Delta}(S) \quad \textit{for all } S \in \mathbb{S}. \tag{2.4}$$

Whenever equality holds in (2.4), it follows from (2.3) that $\operatorname{pen}(S)$ measures the complexity of the model S in terms of dimension and weight.

Denoting Π_S the projection operator onto a linear subspace $S \subset \mathbb{R}^n$, given the families $\mathbb{S}_\lambda$, the penalty function pen and some positive number α, we define

$$\operatorname{crit}_\alpha(\widehat{f}_\lambda) = \inf_{S \in \mathbb{S}_\lambda} \bigl[\|Y - \Pi_S \widehat{f}_\lambda\|^2 + \alpha \|\widehat{f}_\lambda - \Pi_S \widehat{f}_\lambda\|^2 + \operatorname{pen}(S)\widehat{\sigma}_S^2\bigr], \tag{2.5}$$

where

$$\widehat{\sigma}_S^2 = \frac{\|Y - \Pi_S Y\|^2}{n - \dim(S)}. \tag{2.6}$$

For each estimator $\widehat{f}_\lambda$, the criterion (2.5) seeks among the collection $\mathbb{S}_\lambda$ the space S achieving the best trade-off between three terms: the first term evaluates the fit of the projected estimator to the data, the second term quantifies the approximation quality of the space S regarding to the estimator $\widehat{f}_\lambda$ and the last term penalizes S according to its complexity.

From a computational point of view, minimizing (2.5) over Λ requires at most $\sum_{\lambda \in \Lambda} |\mathbb{S}_\lambda|$ steps. In many cases the criterion (2.5) can be minimized much more efficiently, see e.g. Section 3.4 for the case of kernel ridge estimators with $\Lambda = \mathbb{R}_+$.

2.2. *The main result*

For all $\lambda \in \Lambda$ let us set

$$A(\widehat{f}_\lambda, \mathbb{S}_\lambda) = \inf_{S \in \mathbb{S}_\lambda} \bigl[\|\widehat{f}_\lambda - \Pi_S \widehat{f}_\lambda\|^2 + \operatorname{pen}(S)\widehat{\sigma}_S^2\bigr]. \tag{2.7}$$

This quantity corresponds to an accuracy index for the estimator $\widehat{f}_\lambda$ with respect to the family $\mathbb{S}_\lambda$. It is small when the estimator $\widehat{f}_\lambda$ is well approximated by a low dimensional subspace in $\mathbb{S}_\lambda$. The following result holds.

Theorem 2.1. *Let* $K > 1, \alpha > 0, \delta \ge 0$. *Assume that Assumptions* 1, 2 *and* 4 *hold. There exists a constant* $C > 0$ (*given by* (6.4)) *depending on K and α only such that for any* $\widehat{f}_{\widehat{\lambda}}$ *in* $\mathbb{F}$ *satisfying*

$$\operatorname{crit}_\alpha(\widehat{f}_{\widehat{\lambda}}) \le \inf_{\lambda \in \Lambda} \operatorname{crit}_\alpha(\widehat{f}_\lambda) + \delta, \tag{2.8}$$

we have the following bounds

$$C\mathbb{E}\bigl(\|f - \widehat{f}_{\widehat{\lambda}}\|^2\bigr) \le \mathbb{E}\Bigl(\inf_{\lambda \in \Lambda}\bigl\{\|f - \widehat{f}_\lambda\|^2 + A(\widehat{f}_\lambda, \mathbb{S}_\lambda)\bigr\}\Bigr) + \Sigma\sigma^2 + \delta \tag{2.9}$$

$$\le \inf_{\lambda \in \Lambda}\bigl\{\mathbb{E}\bigl(\|f - \widehat{f}_\lambda\|^2\bigr) + \mathbb{E}\bigl(A(\widehat{f}_\lambda, \mathbb{S}_\lambda)\bigr)\bigr\} + \Sigma\sigma^2 + \delta \tag{2.10}$$

(*provided that the quantities involved in the expectations are measurable*).

Furthermore, *if equality holds in* (2.4) *and Assumption* 3 *is satisfied*,

$$C'\mathbb{E}\bigl(\|f - \widehat{f}_{\widehat{\lambda}}\|^2\bigr) \le \mathbb{E}\Bigl(\inf_{\lambda \in \Lambda}\Bigl\{\|f - \widehat{f}_\lambda\|^2 + \inf_{S \in \mathbb{S}_\lambda}\bigl[\|\widehat{f}_\lambda - \Pi_S \widehat{f}_\lambda\|^2 + \bigl[\Delta(S) \vee \dim(S)\bigr]\sigma^2\bigr]\Bigr\}\Bigr) + \Sigma\sigma^2 + \delta, \tag{2.11}$$

where C' *is a positive constant only depending on* κ *and* K.

Let us now comment Theorem 2.1.

It turns out that inequality (2.9) leaves no place for a substantial improvement in the sense that the bound we get is essentially optimal and cannot be uniformly improved (apart from constants) by any other selection rule among $\mathbb{F}$. To see this, let us assume for simplicity that $\mathbb{F}$ is finite so that a measurable minimizer of crit_α always exists and δ can be chosen as 0. Let $K>1$, $\alpha>0$, $\mathbb{S}$ a family of linear subspaces satisfying the assumptions of Theorem 2.1 and pen, the penalty function achieving equality in (2.4). Besides, assume that $\mathbb{S}$ contains a linear subspace S such that $1\le \dim(S)\le n/2$ and associate to S the weight $\Delta(S)=\dim(S)$. If $\mathbb{S}_\lambda=\mathbb{S}$ for all λ, we deduce from (2.11) that for some universal constant C', whatever $\mathbb{F}$ and $f\in\mathbb{R}^n$

$$C'\mathbb{E}\big(\|f-\widehat{f}_{\widetilde{\lambda}}\|^2\big)\le \mathbb{E}\Big(\inf_{\lambda\in\Lambda}\big\{\|f-\widehat{f}_\lambda\|^2+\|\widehat{f}_\lambda-\Pi_S\widehat{f}_\lambda\|^2+\dim(S)\sigma^2\big\}\Big). \tag{2.12}$$

In the opposite direction, the following result holds.

Proposition 1. *Let S be a linear subspace of $\mathbb{R}^n$. There exists a universal constant $C''>0$, such that for any finite family $\mathbb{F}=\{\widehat{f}_\lambda,\lambda\in\Lambda\}$ of estimators and any selection rule $\widetilde{\lambda}$ based on Y among Λ, there exists $f\in S$ such that*

$$C''\mathbb{E}\big[\|f-\widehat{f}_{\widetilde{\lambda}}\|^2\big]\ge \mathbb{E}\Big[\inf_{\lambda\in\Lambda}\big[\|f-\widehat{f}_\lambda\|^2+\|\widehat{f}_\lambda-\Pi_S\widehat{f}_\lambda\|^2+\dim(S)\sigma^2\big]\Big]. \tag{2.13}$$

We see that, up to a numerical constant, the right-hand sides of (2.12) and (2.13) coincide.

In view of commenting (2.10) further, we continue assuming that $\mathbb{F}$ is finite so that we can keep $\delta=0$ in (2.10). A particular feature of (2.10) lies in the fact that the risk bound pays no price for considering a large collection $\mathbb{F}$ of estimators. In fact, it is actually decreasing with respect to $\mathbb{F}$ (or equivalently Λ) for the inclusion. 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 $\widehat{f}_{\widetilde{\lambda}}$ can only be improved. In contrast, the computation of the estimator $\widehat{f}_{\widetilde{\lambda}}$ is all the more difficult that $|\mathbb{F}|$ is large. More precisely, if the cardinalities of the families $\mathbb{S}_\lambda$ are not too large, the computation of $\widehat{f}_{\widetilde{\lambda}}$ requires around $|\mathbb{F}|$ steps.

The selection rule we use does not require to know how the estimators depend on Y. In fact, as we shall see, a more important piece of information is the ranges of the estimators $\widehat{f}_\lambda=\widehat{f}_\lambda(Y)$ as Y varies in $\mathbb{R}^n$. A situation of special interest occurs when each $\widehat{f}_\lambda$ belongs to some (possibly data-driven) linear subspace $\widehat{S}_\lambda$ in $\mathbb{S}$ with probability one. This is the case if one considers Lasso types estimators for example. By taking $\mathbb{S}_\lambda$ such that $\widehat{S}_\lambda\in\mathbb{S}_\lambda$ for all λ, we deduce from bound (2.11) in Theorem 2.1 the following corollary.

Corollary 1. *Assume that the assumptions of Theorem 2.1 are satisfied, that Assumption 3 holds and that equality holds in (2.4). If for all $\lambda\in\Lambda$ there exists a (possibly data-driven) linear subspace $\widehat{S}_\lambda\in\mathbb{S}_\lambda$ such that $\widehat{f}_\lambda\in\widehat{S}_\lambda$ with probability 1, then $\widehat{f}_{\widetilde{\lambda}}$ satisfies*

$$C\mathbb{E}\big[\|f-\widehat{f}_{\widetilde{\lambda}}\|^2\big]\le \inf_{\lambda\in\Lambda}\big[\mathbb{E}\big[\|f-\widehat{f}_\lambda\|^2\big]+\mathbb{E}\big[\dim(\widehat{S}_\lambda)\vee\Delta(\widehat{S}_\lambda)\big]\sigma^2\big]+\delta, \tag{2.14}$$

for some C depending on K and κ only.

One may apply this result in the context of model selection. One starts with a collection of models $\mathbb{S}=\{S_m,m\in\mathcal{M}\}$ and associates to each S_m an estimator $\widehat{f}_m$ with values in S_m. By taking $\mathbb{F}=\{\widehat{f}_m,m\in\mathcal{M}\}$ (here $\Lambda=\mathcal{M}$) and $\mathbb{S}_m=\{S_m\}$ for all $m\in\mathcal{M}$, our selection procedure leads to an estimator $\widehat{f}_{\widehat{m}}$ which satisfies

$$C\mathbb{E}\big[\|f-\widehat{f}_{\widehat{m}}\|^2\big]\le \inf_{m\in\mathcal{M}}\big[\mathbb{E}\big[\|f-\widehat{f}_m\|^2\big]+\big(\dim(S_m)\vee\Delta(S_m)\big)\sigma^2\big]. \tag{2.15}$$

When $\widehat{f}_m=\Pi_{S_m}Y$ for all $m\in\mathcal{M}$, our selection rule becomes

$$\widehat{m}=\arg\min_{m\in\mathcal{M}}\big[\|Y-\widehat{f}_m\|^2+\mathrm{pen}(S_m)\widehat{\sigma}^2_{S_m}\big] \tag{2.16}$$

and coincides with the one described in Baraud et al. [7]. Interestingly, Corollary 1 shows that this selection rule can still be used for families $\mathbb{F}$ of (non-linear) estimators of the form $\Pi_{S_{\widetilde{m}}}Y$ where the $S_{\widetilde{m}}$ are chosen randomly among $\mathbb{S}$

on the basis of Y, doing thus as if the linear subspaces $S_{\widetilde{m}}$ were non-random. An estimator of the form $\Pi_{S_{\widetilde{m}}} Y$ can be seen as resulting from a model selection procedures among the family of projection estimators $\{\Pi_m Y, m \in \mathcal{M}\}$ and our selection rule as a way to select among such candidates procedures.

3. Selecting among linear estimators

In this section, we consider the situation where the estimators $\widehat{f_\lambda}$ are linear, that is, are of the form $\widehat{f_\lambda} = A_\lambda Y$ for some known and deterministic $n \times n$ matrix A_λ. 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). A common feature of those procedures lies in the fact that they depend on a tuning parameter (possibly multidimensional) and their practical performances can be quite poor if this parameter is not suitably calibrated. A series of papers have investigated the calibration of some of these procedures. To mention a few of them, Cao and Golubev [15] focus on spline smoothing, Zhang [54] on kernel ridge regression, Goldenshluger and Lepski [26] on kernel estimators and Arlot and Bach [3] propose a procedure to select linear estimators which are, roughly speaking, "shrinkage" or "averaging" estimators. The procedure we present can handle all these cases in an unified framework. Throughout the section, we assume that Λ is finite, except in Section 3.4.

3.1. *The families* $\mathbb{S}_\lambda$

To apply our selection procedure, we need to associate to each A_λ a suitable collection of approximation subspaces $\mathbb{S}_\lambda$. To do so, we introduce below a linear subspace S_λ which plays a key role in our analysis.

For the sake of simplicity, let us consider first the case where A_λ is non-singular. Then S_λ is 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. If none of the singular values are smaller than 1, then $S_\lambda = \{0\}$.

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 $\overline{\Pi}_\lambda$. Then 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. Again if this set is empty, $S_\lambda = \{0\}$. When A_λ is non-singular or symmetric, we recover the definition of S_λ given above.

For each $\lambda \in \Lambda$, take $\mathbb{S}_\lambda$ such that $\mathbb{S}_\lambda \supset \{S_\lambda\}$. From a theoretical point of view, it is enough to take $\mathbb{S}_\lambda = \{S_\lambda\}$ but practically it may be wise to use a larger set and by doing so, to possibly improve the approximation of $\widehat{f_\lambda}$ by elements of $\mathbb{S}_\lambda$. One may for example take $\mathbb{S}_\lambda = \{S_\lambda^1, \dots, S_\lambda^{n-2}\}$ where S_λ^k is the linear span of the right-singular vectors associated to the k smallest singular values of $A_\lambda^+ - \overline{\Pi}_\lambda$.

3.2. *Choices of* $\mathbb{S}$, Δ *and* pen

Take $\mathbb{S} = \bigcup_{\lambda \in \Lambda} \mathbb{S}_\lambda$ and Δ of the form

$$\Delta(S) = a\bigl(1 \vee \dim(S)\bigr) \quad \text{for all } S \in \mathbb{S},$$

where $a \geq 1$ satisfies Assumption 2 with $\Sigma \leq 1$. One may take $a = (\log|\Lambda|) \vee 1$ even though this choice is not necessarily the best. Finally, for some $K > 1$, take $\mathrm{pen}(S) = K\mathrm{pen}_\Delta(S)$ for all $S \in \mathbb{S}$ and select $\widehat{f_{\widehat{\lambda}}}$ by minimizing the criterion given by (2.5), taking thus $\delta = 0$ in (2.8).

3.3. *An oracle-type inequality for linear estimators*

The following holds.

Corollary 2. *Let $K > 1$, $\kappa \in (0,1)$ and $\alpha > 0$. If Assumption 1 holds and $\Delta(S) \le \kappa n$ for all $S \in \mathbb{S}$, the estimator $\widehat{f}_{\widehat{\lambda}}$ satisfies*

$$Ca^{-1}\mathbb{E}\big[\|f - \widehat{f}_{\widehat{\lambda}}\|^2\big] \le \inf_{\lambda} \mathbb{E}\big[\|f - \widehat{f}_{\lambda}\|^2\big] + \sigma^2,$$

for some $C > 0$ depending on K, α and κ only.

The problem of selecting some best linear estimator among a family of those have also been considered in Arlot and Bach [3] in the Gaussian regression framework, and in Goldenshluger and Lepski [26] in the multidimensional Gaussian white noise model. Arlot and Bach proposed a penalized procedure based on random penalties. Their approach requires that the operators have some "shrinkage" or "averaging" properties (which is the case for all classical procedures) and that the cardinality of Λ is at most polynomial with respect to n, except for families of Kernel-ridge estimators discussed in the next paragraph. Goldenshluger and Lepski proposed 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 nothing. Nevertheless, we restrict to the case of the Euclidean loss whereas Goldenshluger and Lepski considered more general $\mathbb{L}_p$ ones.

3.4. *Case of kernel-ridge estimators*

We can give a more precise result for the case where the family $\{A_\lambda : \lambda \in \Lambda\}$ has a singular value decomposition of the form $A_\lambda = \sum_{k=1}^n \sigma_k(\lambda) u_k v_k^T$, for all $\lambda \in \Lambda$ with $\sigma_1(\lambda) \ge \cdots \ge \sigma_n(\lambda)$ for all $\lambda \in \Lambda$. Such a situation occurs for example for low-pass filters and kernel ridge regression (including spline smoothing). For simplicity, we restrict henceforth to kernel-ridge regression.

Kernel ridge regression arises in the signal denoising setting (1.1). Let $\mathcal{H}$ be a Reproducing Kernel Hilbert Space on $\mathcal{X}$ with kernel $\mathbf{k}$ and norm $\|\cdot\|_{\mathcal{H}}$. For $\lambda > 0$, the kernel ridge regression estimator is the estimator $\widehat{f}_\lambda = (\widehat{F}_\lambda(x_1), \ldots, \widehat{F}_\lambda(x_n))$ where $\widehat{F}_\lambda$ is the solution of the minimization problem

$$\widehat{F}_\lambda \in \operatorname*{arg\,min}_{F \in \mathcal{H}} \left\{ \sum_{i=1}^n \big(y_i - F(x_i)\big)^2 + \lambda \|F\|_{\mathcal{H}}^2 \right\}.$$

It is a linear estimator given by $\widehat{f}_\lambda = K(K + \lambda I_n)^{-1} Y$ where I_n denotes the identity matrix on $\mathbb{R}^n$ and K the positive semi-definite matrix $K = [\mathbf{k}(x_i, x_j)]_{i,j=1,\ldots,n}$. Hence, by writing $K = \sum_k s_k v_k v_k^T$ (with $s_1 \ge \cdots \ge s_n \ge 0$) for the singular value decomposition of the kernel matrix K, the associated kernel ridge operator A_λ is given by

$$A_\lambda = \sum_{k=1}^n \frac{s_k}{s_k + \lambda} v_k v_k^T \quad \text{for all } \lambda > 0.$$

For a given $\kappa \in (0,1)$, we set $k_n = \lfloor \kappa n \rfloor$, $\Lambda = (s_{k_n}, +\infty)$ and $\mathbb{S} = \{S_1, \ldots, S_{k_n}\}$ where $S_d = \operatorname{span}\{v_1, \ldots, v_d\}$. Writing $c_j = \langle f, v_j \rangle$ for all $j = 1, \ldots, n$, the selection criterion (2.5) is given by

$$\operatorname{crit}_\alpha(\widehat{f}_\lambda) = \inf_{1 \le d \le k_n} G(\lambda, d), \quad \text{where } G(\lambda, d) = \sum_{j \le d} c_j^2 \left(\frac{\lambda}{\lambda + s_j}\right)^2 + \sum_{j > d} c_j^2 \left[1 + \frac{\operatorname{pen}(S_d)}{n - d} + \alpha \left(\frac{s_j}{\lambda + s_j}\right)^2\right].$$

This criterion can be efficiently minimized by computing for each value of d the parameter λ_d minimizing $\lambda \to G(\lambda, d)$ and then by taking $\widehat{\lambda} = \lambda_{\hat{d}}$, where $\hat{d}$ minimizes $d \to G(\lambda_d, d)$ over $\{1, \ldots, k_n\}$. For the choice $\operatorname{pen}(S) = K \operatorname{pen}_\Delta(S)$ with $K > 1$ and $\Delta(S) = \dim(S)$, the resulting estimator $\widehat{f}_{\widehat{\lambda}}$ fulfills the risk bound

$$C\mathbb{E}\big[\|f - \widehat{f}_{\widehat{\lambda}}\|^2\big] \le \inf_{\lambda > s_{k_n}} \mathbb{E}\big[\|f - \widehat{f}_\lambda\|^2\big] + \sigma^2,$$

for some $C > 0$ depending on K, α and κ only. Our procedure therefore achieves an oracle risk bound on the continuous family of kernel ridge estimators $\{\widehat{f_\lambda} : \lambda > s_{k_n}\}$. The problem of selecting among the collection $\{\widehat{f_\lambda} : \lambda > 0\}$ of kernel ridge estimators has also been tackled recently by Arlot and Bach [3]. They provide an oracle risk bound for this problem but their approach requires the assumptions that for some $\lambda > 0$, $\mathrm{Tr}(A_\lambda) \le \sqrt{n}$ and $\|(I - A_\lambda) f\|^2 \le \sigma^2\sqrt{n \log(n)}$.

4. Variable selection

Throughout this section, we consider the problem of variable selection introduced in Example 2. There exist various ways of evaluating the theoretical performance of a variable selection procedure. One is to look at the difference between the selected set of predictors and the true one. This will not be the point of view developed in this section which, as we shall see, will rather be oriented towards the minimization of the risk.

Throughout this section, the vector of observation Y is assumed to be of the form $Y = X\beta + \varepsilon$, with a $n \times p$ fixed design matrix X. We assume that $p \ge 2$ in order to avoid trivialities. When p is small enough (say smaller than 20), this problem can be solved by using a suitable variable selection procedure that explores all the subsets of $\{1, \dots, p\}$. For example, one may use the penalized criterion introduced in Birgé and Massart [10] when the variance is known, and the one in Baraud et al. [7] 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, many algorithms based on the minimization of convex criteria have been proposed: the Lasso, the Dantzig selector of Candès and Tao [14], the elastic net of Zou and Hastie [58], to mention a few. An alternative to those criteria is the forward-backward algorithm described in Zhang [55], 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.

4.1. *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. 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 $S_\emptyset = \{0\}$). For $\ell \in \mathcal{L}$ and $h \in H_\ell$, associate to the subset $\widehat{m}(\ell, h)$ an estimator $\widehat{f}_{(\ell,h)}$ of f with values in $S_{\widehat{m}(\ell,h)}$ (one may for example take the projection of Y onto the random linear subspace $S_{\widehat{m}(\ell,h)}$ but any other choice would suit as well). Finally, consider the family $\mathbb{F} = \{\widehat{f_\lambda}, \lambda \in \Lambda\}$ of these estimators by taking $\Lambda = \bigcup_{\ell \in \mathcal{L}}(\{\ell\} \times H_\ell)$ and set $\widehat{\mathcal{M}} = \{\widehat{m}(\lambda), \lambda \in \Lambda\}$. All along we assume that Λ is finite (so that we take $\delta = 0$ in (2.8)).

The approximation spaces and the weight function

Throughout, we shall restrict ourselves to subsets of predictors with cardinality not larger than some $D_{\max} \le n - 2$. In view of approximating the estimators $\widehat{f_\lambda}$, we suggest the collection $\mathbb{S}$ given by

$$\mathbb{S} = \bigcup \bigl\{ S_m | m \subset \{1, \dots, p\}, \mathrm{card}(m) \le D_{\max} \bigr\}. \tag{4.1}$$

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

$$\Delta(S) = \log\left[\binom{p}{D}\right] + \log(1 + D) \quad \text{with } D = \dim(S) \vee 1. \tag{4.2}$$

Since

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

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 (2.5) 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 $\lambda = (\ell, h) \in \Lambda$

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

(where the S_m are defined above), or preferably

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

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

4.2. *The results*

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

$$1 \le \dim(S_{\widehat{m}(\lambda)}) \vee \Delta(S_{\widehat{m}(\lambda)}) \le 2\big(\dim(S_{\widehat{m}(\lambda)}) \vee 1\big)\log p.$$

Hence, by applying Corollary 1 with $\widehat{S}_\lambda = S_{\widehat{m}(\lambda)}$, we get the following result.

Corollary 3. *Let $K > 1$, $\kappa \in (0,1)$ and $D_{\max}$ be some positive integer satisfying $D_{\max} \le \kappa n/(2\log p)$. Let $\widehat{\mathcal{M}} = \{\widehat{m}(\lambda), \lambda \in \Lambda\}$ be a (finite) collection of random subsets of $\{1,\dots,p\}$ with cardinality not larger than $D_{\max}$ based on the observation Y and $\{\widehat{f}_\lambda, \lambda \in \Lambda\}$ a family of estimators f, also based on Y, such that $\widehat{f}_\lambda \in S_{\widehat{m}(\lambda)}$. By applying our selection procedure, the resulting estimator $\widehat{f}_{\widehat{\lambda}}$ satisfies*

$$C\mathbb{E}\big[\|f-\widehat{f}_{\widehat{\lambda}}\|^2\big] \le \inf_{\lambda\in\Lambda}\big[\mathbb{E}\big[\|f-\widehat{f}_\lambda\|^2\big] + \mathbb{E}\big[\dim(S_{\widehat{m}(\lambda)}) \vee 1\big]\log(p)\sigma^2\big], \tag{4.3}$$

where C is a constant depending on the choices of K and κ only.

Again, note that the risk bound we get is non-increasing with respect to Λ. This means that if one adds a new variable selection procedure or considers more tuning parameters to increase Λ, the risk bound we get can only be improved. It is also worth mentioning that our selection procedure does not require to know how the estimators $\widehat{f}_\lambda$ of the family $\mathbb{F}$ depend on the data Y. In particular, these estimators could be obtained from the computation of some software for which the detailed program is unknown to the user or from the computation of some expert keeping his art secret.

As already mentioned, our selection procedure can be used in view of tuning the parameter $\lambda > 0$ involved in the Lasso criterion as presented in Example 3. For the family of estimators $\{\widehat{f}_\lambda, \lambda > 0\}$ given by (1.3), our selection rule (2.5) with $\mathbb{S}_\lambda$ restricted to $\{S_{\widehat{m}(\lambda)}\}$ for all $\lambda > 0$ is very similar to that proposed by Zou et al. [59] and amounts to replacing $w_n|\widehat{m}(\lambda)|$ by $\mathrm{pen}(S_{\widehat{m}(\lambda)})\widehat{\sigma}^2_{\widehat{m}(\lambda)}/\sigma^2$ in their formula (2.17) on p. 2182. Since $\mathrm{pen}(S_m)$ is of order $|m|\log p$ when $|m|\log p$ is small compared to n, the two selection procedures are essentially the same for w_n of order $\log p$ and these particular choices of $\mathbb{S}_\lambda$ (up to a model-dependent estimator of σ^2). Nevertheless, as we shall see in Section 4.3.2, in practice we rather suggest to use (2.5) with a family $\mathbb{S}_\lambda$ which is not restricted to $\{S_{\widehat{m}(\lambda)}\}$ in order to improve the performance of the selection rule.

On the basis of the present paper, further developments have been done in Giraud et al. [24] for the problem of tuning the parameters involved in the Lasso and the group-Lasso type procedures. In particular, the reader will find there that (4.3) turns to an oracle inequality under a suitable assumption on design matrix X.

Finally, under the assumption that $f \in S_{m^*}$ and that m^* belongs to $\widehat{\mathcal{M}}$ with probability close enough to 1, we can compare the risk of the estimator $\widehat{f}_{\widehat{\lambda}}$ to the cardinality of m^*.

Corollary 4. *Assume that the assumptions of Corollary 3 hold and that $\widehat{f}_\lambda = \Pi_{S_{\widehat{m}(\lambda)}} Y$ for all $\lambda \in \Lambda$. If $f \in S_{m^*}$ for some non-void subset $m^* \subset \{1, \dots, p\}$ with cardinality not larger than $D_{\max}$, then*

$$C\mathbb{E}\big[\|f - \widehat{f}_{\widehat{\lambda}}\|^2\big] \le \log(p)\big|m^*\big|\sigma^2 + R_n\big(m^*\big),$$

where C is a constant depending on K and κ only, and

$$R_n\big(m^*\big) = \big(\|f\|^2 + n\sigma^2\big)\big(\mathbb{P}\big[m^* \notin \widehat{\mathcal{M}}\big]\big)^{1/2}.$$

Zhao and Yu [56] give sufficient conditions on the design X to ensure that $\mathbb{P}[m^* \notin \widehat{\mathcal{M}}]$ is exponentially small with respect to n when the family $\widehat{\mathcal{M}}$ is obtained by using the LARS-Lasso algorithm with different values of the tuning parameter.

4.3. *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.

4.3.1. *Simulation design*

An 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 $e = 1, \dots, E$ where each collection $\mathcal{E}_e$ is characterized by a vector of parameters β_e, and a set $\mathcal{X}_e$ of matrices X:

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

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 A.2. They are inspired by examples found in Tibshirani [46], Zou and Hastie [58], Zou [57], and Huang et al. [31] 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.

Table 1
Mean, standard-error and quantiles of the ratios $R_{\mathrm{ex}}/O_{\mathrm{ex}}$ calculated over all ex $\in \mathcal{E}$ such that $O_{\mathrm{ex}} < n\sigma^2/3$. The number of such examples equals 654, see Section A.2

			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

4.3.2. *Tuning a smoothing parameter*

In this section, we consider 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 $\{\widehat{f}_1, \dots, \widehat{f}_{D_{\max}}\}$ of estimators given by the $D_{\max}$ first steps of the LARS-Lasso algorithm proposed by Efron et al. [21]. Hence, the tuning parameter is here the number $h \in H = \{1, \dots, D_{\max}\}$ 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 $\mathbb{F} = \{\widehat{f}_1, \dots, \widehat{f}_{D_{\max}}\}$.

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 (4.1). We associate to $\widehat{f}_h$ the family $\mathbb{S}_h = \{S_{\widehat{m}(h')} | h' \in H\} \subset \mathbb{S}$ where the S_m are defined in Section 4.1 and $\widehat{m}(h') \subset \{1, \dots, p\}$ is the set of indices corresponding to the predictors returned 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 (4.2) and $K = 1.1$. This value of K is consistent with what is suggested in Baraud et al. [7]. 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 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 (2.5). Second, simple calculation shows that the constant $C^{-1} = C^{-1}(1.1, \alpha)$ involved in Theorem 2.1 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 $\widehat{f}_{\mathrm{pen}_\Delta}$ 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 ex $\in \mathcal{E}$, we estimate on the basis of 400 simulations the oracle risk

$$O_{\mathrm{ex}} = \mathbb{E}\Big(\min_{h \in H} \|f - \widehat{f}_h\|^2\Big), \tag{4.4}$$

and the Euclidean risks $R_{\mathrm{ex}}(\widehat{f}_{\mathrm{pen}_\Delta})$ and $R_{\mathrm{ex}}(\widehat{f}_{CV})$ of $\widehat{f}_{\mathrm{pen}_\Delta}$ 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 $R_{\mathrm{ex}}(\widehat{f}_{\mathrm{pen}_\Delta})/O_{\mathrm{ex}}$ are closer to one than $R_{\mathrm{ex}}(\widehat{f}_{CV})/O_{\mathrm{ex}}$.

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.

4.3.3. *Variable selection*

In this section, we consider the problem of variable selection and use the procedure and notations introduced in Section 4.1. To solve this problem, we consider estimators of the form $\widehat{f}_{\widehat{m}} = \Pi_{S_{\widehat{m}}} Y$ where $\widehat{m}$ is a random subset of $\{1, \dots, p\}$ depending on Y. Given a family $\widehat{\mathcal{M}} = \{\widehat{m}(\ell, h), \widehat{m}(\ell, h) \in \mathcal{L} \times H_\ell\}$ of such random sets, we consider the

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

			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

family $\mathbb{F} = \{\widehat{f}_{\widehat{m}(\ell,h)} | (\ell, h) \in \mathcal{L} \times H_\ell\}$. The descriptions of $\mathcal{L}$ and H_ℓ are postponed to Section A.3. Let us merely mention that we choose $\mathcal{L}$ which gathers variable selection procedures based on the Lasso, ridge regression, Elastic net, PLS1 regression, Adaptive Lasso, Random Forest, and on an exhaustive research among the subsets of $\{1, \dots, p\}$ with small cardinality. For each procedure ℓ, the parameter set H_ℓ corresponds to different choices of tuning parameters. For each $\lambda = (\ell, h) \in \mathcal{L} \times H_\ell$, we take $\mathbb{S}_\lambda = \{S_{\widehat{m}(\ell,h)}\}$ so that our selection rule among $\mathbb{F}$ amounts to minimizing over $\widehat{\mathcal{M}}$

$$\text{crit}(m) = \|Y - \Pi_{S_m} Y\|^2 + K \text{pen}_\Delta(S_m)\widehat{\sigma}^2_{S_m}, \tag{4.5}$$

where pen_Δ is given by (2.2).

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 $\mathbb{F}_\ell = \{\widehat{f}_{\widehat{m}(\ell,h)} | h \in H_\ell\}$ by minimizing criterion (4.5) over $\widehat{\mathcal{M}}_\ell = \{\widehat{m}(\ell, h) | h \in H_\ell\}$. We denote by $\widehat{m}(\ell)$ the selected set and by $\widehat{f}_{\widehat{m}(\ell)}$ the corresponding projection estimator. For each example ex $\in \mathcal{E}$ and each method $\ell \in \mathcal{L}$, we estimate the risk

$$R_{\text{ex},\ell} = \mathbb{E}\big(\|f - \widehat{f}_{\widehat{m}(\ell)}\|^2\big)$$

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

$$R_{\text{ex,all}} = \mathbb{E}\big(\|f - \widehat{f}_{\widehat{m}}\|^2\big).$$

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

$$R_{\text{ex},\min} = \min\{R_{\text{ex,all}}, R_{\text{ex},\ell}, \ell \in \mathcal{L}\}.$$

We compare the ratios $R_{\text{ex},\ell}/R_{\text{ex},\min}$ for $\ell \in \mathcal{L} \cup \{\text{all}\}$ to judge the performances of the candidate procedures on each example ex $\in \mathcal{E}$. The mean, standard deviations and quantiles of the sequence $\{R_{\text{ex},\ell}/R_{\text{ex},\min}, \text{ex} \in \mathcal{E}\}$ 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.
- The exhaustive procedure gives very bad results, because the research in subsets of $\{1, \dots, p\}$ is limited to subsets of very small cardinality, see Section A.3. Nevertheless in some examples with $p = 50$, the exhaustive method may give better results than all the others.
- Our procedure, corresponding to $\ell = \text{all}$, achieves the smallest mean value of the risk ratio. Besides, this value is very close to one.

Table 3
False dicovery rate (FDR) and true discovery rate (TDR) using our method, for each example with $\rho = 10$ and $n = p = 100$

	$\mathcal{E}_1$	$\mathcal{E}_2$	$\mathcal{E}_3$	$\mathcal{E}_4$	$\mathcal{E}_5$	$\mathcal{E}_6$	$\mathcal{E}_7$	$\mathcal{E}_8$	$\mathcal{E}_9$	$\mathcal{E}_{10}$	$\mathcal{E}_{11}$
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

- The variability of our procedure is small compared to the others. In particular, it is smaller than the variability of ALridge which behaves similarly in expectation. Even for the worst examples considered in this collection the risk is under control.
- For all examples, our procedure selects an estimator the risk of which does not exceed twice that of the oracle.

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.

5. Aggregation

In this section, we assume that we have at hand $M \geq 2$ preliminary estimators of f, denoted $\{\phi_k, k = 1, \dots, M\}$ that do not depend on Y. One may either think of the situation where there exists an independent copy Y' of Y and that the estimators ϕ_k are obtained from Y', or that the ϕ_k are deterministic vectors. Let us mention that when the variance is known, it is always possible to duplicate the Gaussian vector Y in order to have an independent copy of it (by following some trick given by Nemirovskii in his course of Saint-Flour [40]). Unfortunately, it is no longer possible when the variance is unknown. In this specific context of an unknown variance, we address here the problems of *Model Selection Aggregation* (MS), *Convex Aggregation* (Cv) and *Linear Aggregation* (L) defined below. Our aim is to build an estimator $\widehat{f}$ based on Y whose risk is as close as possible to $\inf_{g \in \mathbb{F}_\Lambda} \|f - g\|^2$ where

$$\mathbb{F}_\Lambda = \left\{ f_\lambda = \sum_{j=1}^{M} \lambda_j \phi_j, \lambda \in \Lambda \right\}$$

and, according to the aggregation problem at hand, Λ is one of the three sets

$$\Lambda_{\mathrm{MS}} = \left\{ \lambda \in \{0,1\}^M, \sum_{j=1}^{M} \lambda_j = 1 \right\}, \qquad \Lambda_{\mathrm{Cv}} = \left\{ \lambda \in \mathbb{R}_+^M, \sum_{j=1}^{M} \lambda_j = 1 \right\}, \qquad \Lambda_{\mathrm{L}} = \mathbb{R}^M.$$

When $\Lambda = \Lambda_{\mathrm{MS}}$, $\mathbb{F}_\Lambda$ is the set $\{\phi_1, \dots, \phi_M\}$ consisting of the initial estimators. When $\Lambda = \Lambda_{\mathrm{Cv}}$, $\mathbb{F}_\Lambda$ is the convex hull of the ϕ_j. In the literature, one may also find

$$\Lambda'_{\mathrm{Cv}} = \left\{ \lambda \in [0,1]^M, \sum_{j=1}^{M} \lambda_j \leq 1 \right\}$$

in place of Λ_{Cv} in which case $\mathbb{F}_\Lambda$ is the convex hull of $\{0, \phi_1, \dots, \phi_M\}$. Finally, when $\Lambda = \Lambda_{\mathrm{L}}$, $\mathbb{F}_\Lambda$ is the linear span of the ϕ_j.

Each of these three aggregation problems are solved *separately* if for each $\Lambda \in \{\Lambda_{\mathrm{MS}}, \Lambda_{\mathrm{Cv}}, \Lambda_{\mathrm{L}}\}$ one can design an estimator $\widehat{f} = \widehat{f}(\Lambda)$ satisfying

$$\mathbb{E}\big[\|f - \widehat{f}\|^2\big] - C \inf_{g \in \mathbb{F}_\Lambda} \|f - g\|^2 \leq C' \psi_{n,\Lambda} \sigma^2 \tag{5.1}$$

with $C = 1$, $C' > 0$ free of f, n, M and

$$\psi_{n,\Lambda} = \begin{cases} M & \text{if} \Lambda = \Lambda_{\mathrm{L}}, \\ \sqrt{n}\log(\mathrm{e}M/\sqrt{n}) & \text{if } \Lambda = \Lambda_{\mathrm{Cv}} \text{ and } \sqrt{n} \le M, \\ M & \text{if } \Lambda = \Lambda_{\mathrm{Cv}} \text{ and } \sqrt{n} \ge M, \\ \log M & \text{if } \Lambda = \Lambda_{\mathrm{MS}}. \end{cases} \tag{5.2}$$

These problems have only been considered when the variance is known. The quantity $\psi_{n,\Lambda}$ then corresponds to the best possible upper bound in (5.1) over all possible $f \in \mathbb{R}^n$ and preliminary estimators ϕ_j and is called the *optimal rate of aggregation*. For a more precise definition, we refer the reader to Tsybakov [47]. Bunea et al. [13] considered the problem of solving these three problems *simultaneously* by building an estimator $\widehat{f}$ which satisfies (5.1) simultaneously for all $\Lambda \in \{\Lambda_{\mathrm{MS}}, \Lambda_{\mathrm{Cv}}, \Lambda_{\mathrm{L}}\}$ and some constant $C > 1$. This is an interesting issue since it is impossible to know in practice which aggregation device should be used to achieve the smallest risk bound: as Λ grows (for the inclusion), the bias $\inf_{g\in\mathbb{F}_\Lambda} \|f - g\|^2$ decreases while the rate $\psi_{n,\Lambda}$ increases.

The aim of this section is to show that our procedure provides a way of solving (or nearly solving) the three aggregation problems both *separately* and *simultaneously* when the variance is unknown.

Throughout this section, we consider the family $\overline{\mathbb{S}}$ consisting of the S_m defined for each $m \subset \{1, \dots, M\} \setminus \{\varnothing\}$ as the linear span of the ϕ_j for $j \in m$ and $S_\varnothing = \{0\}$. Along this section, we shall use the weight function Δ defined on $\overline{\mathbb{S}}$ by

$$\Delta(S_m) = |m| + \log\left[\binom{M}{|m|}\right] \quad \text{if } m \neq \varnothing \quad \text{and} \quad \Delta(S_\varnothing) = 1$$

take $\alpha = 1/2$ and $\mathrm{pen}(\cdot) = 1.1\mathrm{pen}_\Delta(\cdot)$ taking thus $K = 1.1$. We make these choices of α and K only to fix up the ideas. Note that Δ satisfies Assumption 2 with $\Sigma < 1$. To avoid trivialities, we assume all along $n \ge 4$.

5.1. *Solving the three aggregation problems separately*

5.1.1. *Linear Aggregation*

Problem (L) is the easiest to solve. Let us take $\mathbb{F} = \mathbb{F}_\Lambda$ with $\Lambda = \Lambda_{\mathrm{L}}$ and

$$\mathbb{S} = \mathbb{S}_{\mathrm{L}} = \{S_{\{1,\dots,M\}}\} \tag{5.3}$$

and $\mathbb{S}_\lambda = \mathbb{S}_{\mathrm{L}}$ for all $\lambda \in \Lambda_{\mathrm{L}}$. Minimizing $\mathrm{crit}_\alpha(f_\lambda)$ over $f_\lambda \in \mathbb{F}_\Lambda$ amounts to minimizing $\|Y - f_\lambda\|^2$ over $f_\lambda \in S_{\{1,\dots,M\}}$ and hence, the resulting estimator is merely $\widehat{f}_{\mathrm{L}} = \Pi_{S_{\{1,\dots,M\}}} Y$. The risk of $\widehat{f}_{\mathrm{L}}$ satisfies

$$\mathbb{E}\big[\|f - \widehat{f}_{\mathrm{L}}\|\big] \le \inf_{g\in\mathbb{F}_\Lambda} \|f - g\|^2 + (M \wedge n)\sigma^2$$

whatever n and M. This solves the problem of *Linear Aggregation*.

5.1.2. *Model Selection Aggregation*

To tackle Problem (MS), we take $\mathbb{F} = \mathbb{F}_\Lambda$ with $\Lambda = \Lambda_{\mathrm{MS}}$, that is, $\mathbb{F}_\Lambda = \{\phi_1, \dots, \phi_M\}$,

$$\mathbb{S} = \mathbb{S}_{\mathrm{MS}} = \{S_{\{1\}}, \dots, S_{\{M\}}\} \tag{5.4}$$

and associate to each $f_\lambda = \phi_j$ the collection $\mathbb{S}_\lambda$ reduced to $\{S_{\{j\}}\}$. Note that $\dim(S) \le 1$ and $\Delta(S) = \log(\mathrm{e}M) \ge \dim(S)$ for all $S \in \mathbb{S}_{\mathrm{MS}}$, so that under the assumption that $\log(\mathrm{e}M) \le n/2$ we may apply Corollary 1 with $\delta = 0$ (since $\mathbb{F}_\Lambda$ is finite), $\kappa = 1/2$ and get that for some constant $C > 0$ the resulting estimator $\widehat{f}_{\mathrm{MS}}$ satisfies

$$C\mathbb{E}\big[\|f - \widehat{f}_{\mathrm{MS}}\|^2\big] \le \inf_{g\in\mathbb{F}_\Lambda} \|f - g\|^2 + \log(M)\sigma^2.$$

This risk bound is of the form (5.1) except for the constant C which is not equal to 1.

5.1.3. *Convex Aggregation*

Let us consider the family of estimators $\mathbb{F} = \mathbb{F}_\Lambda$ with $\Lambda = \Lambda_{\mathrm{Cv}}$ and

$$\mathbb{S} = \mathbb{S}_{\mathrm{Cv}} = \mathbb{S}_\lambda = \big\{S_m \in \overline{\mathbb{S}}, |m| \le d(n, M)\big\} \quad \forall \lambda \in \Lambda_{\mathrm{Cv}}, \tag{5.5}$$

where $d(n, M) = n/(2\log(\mathrm{e}M))$. The set Λ_{Cv} being compact, $\lambda \mapsto \mathrm{crit}_\alpha(f_\lambda)$ admits a minimum $\widehat{\lambda}$ over Λ_{Cv} and we set $\widehat{f}_{\mathrm{Cv}} = \widehat{f}_{\widehat{\lambda}}$. For such an estimator, the following holds.

Proposition 2. *Assume that $M \le \mathrm{e}^{n/4-1}$ and let $\rho = \sup_{j=1,\dots,M} \|\phi_j\|/\sigma$. The estimator $\widehat{f}_{\mathrm{Cv}}$ satisfies for some universal constant $C > 0$*

$$C\mathbb{E}\big[\|f - \widehat{f}_{\mathrm{Cv}}\|^2\big] \le \inf_{g\in\mathbb{F}_\Lambda} \|f - g\|^2 + B_{\mathrm{Cv}}\sigma^2,$$

where B_{Cv} is defined as follows:

$$B_{\mathrm{Cv}} = \begin{cases} \rho\sqrt{\log(\mathrm{e}M/\rho)} \wedge (\rho^2 \vee 1) & \text{if } \rho \le M \wedge d(n, M), \\ M & \text{if } \rho > M \wedge d(n, M) \text{ and } M \le d(n, M), \\ \rho^2/d(n, M) + d(n, M)\log(\mathrm{e}M/d(n, M)) & \text{if } \rho > M \wedge d(n, M) \text{ and } M > d(n, M). \end{cases} \tag{5.6}$$

In the literature, only the dependency of the aggregation rate with respect to n and M is emphasized and that with respect to $L = \rho/\sqrt{n}$ omitted. If one considers L as a constant, so that ρ is of order $\sqrt{n}$, and assumes that M remains small enough compared to n (more precisely, $M \le d(n, M)$), B_{Cv} is of order $\min\{M, \sqrt{n\log(\mathrm{e}M/\sqrt{n})}\}$ and one recovers the usual aggregation rate. However, one may get different rates by considering L as a function of n. This is a somewhat reasonable point of view since typically the ϕ_j are estimators of f. Unfortunately, the value of L is unknown to the statistician, since it depends on σ, and it is therefore impossible to design an aggregation rule based on L in order to achieve these rates. A particular feature of our aggregation strategy, which does not depend on σ, lies in the fact that it does not rely on the *prior* knowledge of L.

5.2. *Solving the three problems simultaneously*

Consider now three estimators $\widehat{f}_{\mathrm{L}}, \widehat{f}_{\mathrm{MS}}, \widehat{f}_{\mathrm{Cv}}$ with values respectively in $S_{\{1,\dots,M\}}$, $\bigcup_{j=1}^M S_{\{j\}}$ and the convex hull $\mathcal{C}$ of the ϕ_j (we use a new notation for this convex hull to avoid ambiguity). One may take the estimators defined in Section 5.1 but any others would suit. The aim of this section is to select the one with the smallest risk to estimate f. To do so, we apply our selection procedure with $\mathbb{F} = \{\widehat{f}_{\mathrm{L}}, \widehat{f}_{\mathrm{MS}}, \widehat{f}_{\mathrm{Cv}}\}$, taking thus $\Lambda = \{\mathrm{L}, \mathrm{MS}, \mathrm{Cv}\}$, and associate to each of these three estimators the families $\mathbb{S}_{\mathrm{L}}, \mathbb{S}_{\mathrm{MS}}, \mathbb{S}_{\mathrm{Cv}}$ defined by (5.3), (5.4) and (5.5) respectively and choose $\mathbb{S} = \mathbb{S}_{\mathrm{L}} \cup \mathbb{S}_{\mathrm{MS}} \cup \mathbb{S}_{\mathrm{Cv}}$.

Proposition 3. *Assume that $M \le \min\{\mathrm{e}^{n/4-1}, d(n, M)\}$ where $d(n, M) = n/(2\log(\mathrm{e}M))$. There exists a universal constant $C > 0$ such that whatever $\widehat{f}_{\mathrm{L}}, \widehat{f}_{\mathrm{MS}}$ and $\widehat{f}_{\mathrm{Cv}}$ with values in $S_{\{1,\dots,M\}}$, $\bigcup_{j=1}^M S_{\{j\}}$ and $\mathcal{C}$ respectively, the selected estimator $\widehat{f}_{\widehat{\lambda}}$ satisfies for all $f \in \mathbb{R}^n$*,

$$C\mathbb{E}\big[\|f - \widehat{f}_{\widehat{\lambda}}\|^2\big] \le \inf_{\lambda\in\{\mathrm{L},\mathrm{MS},\mathrm{Cv}\}} \big[\mathbb{E}\big[\|f - \widehat{f}_\lambda\|^2\big] + B_\lambda\sigma^2\big],$$

where $B_{\mathrm{L}} = M$, $B_{\mathrm{MS}} = \log M$ and B_{Cv} is given by (5.6). *In particular, if $\widehat{f}_{\mathrm{L}}, \widehat{f}_{\mathrm{MS}}$ and $\widehat{f}_{\mathrm{Cv}}$ are the estimators defined in Sections* 5.1.1, 5.1.2 *and* 5.1.3 *respectively then*

$$C\mathbb{E}\big[\|f - \widehat{f}_{\widehat{\lambda}}\|^2\big] \le \inf_{\lambda\in\{\mathrm{L},\mathrm{MS},\mathrm{Cv}\}} \Big[\inf_{g\in\mathbb{F}_\lambda} \|f - g\|^2 + B_\lambda\sigma^2\Big],$$

where $\mathbb{F}_\lambda$ stands for $\mathbb{F}_\Lambda$ when $\Lambda = \Lambda_\lambda$.

6. Proofs

6.1. *Proof of Theorem* 2.1

We denote by $\langle\cdot,\cdot\rangle$ the inner product of $\mathbb{R}^n$ and for all $\lambda\in\Lambda$ and $S\in\mathbb{S}_\lambda$, write

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

where

$$\mathfrak{pen}(S) = \mathrm{pen}(S)\widehat{\sigma}_S^2/\sigma^2 \quad \text{for all } S\in\mathbb{S}. \tag{6.1}$$

For all $\lambda\in\Lambda$, let $S(\lambda)\in\mathbb{S}_\lambda$ be such that

$$\mathrm{crit}_\alpha\big(\widehat{f}_\lambda, S(\lambda)\big) \le \mathrm{crit}_\alpha(\widehat{f}_\lambda) + \delta.$$

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

$$\mathrm{crit}_\alpha\big(\widehat{f}_{\widehat{\lambda}}, S(\widehat{\lambda})\big) \le \mathrm{crit}_\alpha(\widehat{f}_{\widehat{\lambda}}) + \delta \le \mathrm{crit}_\alpha(\widehat{f}_\lambda) + 2\delta \le \mathrm{crit}_\alpha(\widehat{f}_\lambda, S) + 2\delta$$

and simple algebra that

$$\begin{aligned}
&\|f-\Pi_{S(\widehat{\lambda})}\widehat{f}_{\widehat{\lambda}}\|^2 + \alpha\|\widehat{f}_{\widehat{\lambda}} - \Pi_{S(\widehat{\lambda})}\widehat{f}_{\widehat{\lambda}}\|^2 \\
&\quad \le \|f-\Pi_S\widehat{f}_\lambda\|^2 + \alpha\|\widehat{f}_\lambda - \Pi_S\widehat{f}_\lambda\|^2 + 2\sigma^2\mathfrak{pen}(S) + 2\delta \\
&\quad + 2\langle\varepsilon, \Pi_{S(\widehat{\lambda})}\widehat{f}_{\widehat{\lambda}} - f\rangle - \sigma^2\mathfrak{pen}\big(S(\widehat{\lambda})\big) + 2\langle\varepsilon, f-\Pi_S\widehat{f}_\lambda\rangle - \sigma^2\mathfrak{pen}(S).
\end{aligned}$$

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

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

the second inequality following from $ab\le K^{-1}a^2 + Kb^2$ for all positive K.

Let $\tilde{\Sigma}$ be defined as follows

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

By using (2.4) and (6.1), and noting that for each $S\in\mathbb{S}$,

$$\frac{\|Y-\Pi_SY\|^2}{n-\dim(S)} \ge \frac{\|Y-\Pi_{\overline{S}}Y\|^2}{n-\dim(S)},$$

we get

$$\begin{aligned}
&\big(1-K^{-1}\big)\|f-\Pi_{S(\widehat{\lambda})}\widehat{f}_{\widehat{\lambda}}\|^2+\alpha\|\widehat{f}_{\widehat{\lambda}}-\Pi_{S(\widehat{\lambda})}\widehat{f}_{\widehat{\lambda}}\|^2\\
&\quad\le\big(1+K^{-1}\big)\|f-\Pi_S\widehat{f}_\lambda\|^2+\alpha\|\widehat{f}_\lambda-\Pi_S\widehat{f}_\lambda\|^2+2\sigma^2\mathfrak{pen}(S)+\tilde{\Sigma}+2\delta\\
&\quad\le 2\big(1+K^{-1}\big)\|f-\widehat{f}_\lambda\|^2+2\delta\\
&\qquad+\big(\alpha+2\big(1+K^{-1}\big)\big)\|\widehat{f}_\lambda-\Pi_S\widehat{f}_\lambda\|^2+2\sigma^2\mathfrak{pen}(S)+\tilde{\Sigma}.
\end{aligned}\tag{6.2}$$

Now, since the variable $\|Y-\Pi_{\overline{S}}Y\|^2$ is independent of $\|\Pi_{\overline{S}}\varepsilon\|^2$ and is stochastically larger than (or equal to) $\|\varepsilon-\Pi_{\overline{S}}\varepsilon\|^2$, we deduce from the definition of $\mathrm{pen}_\Delta(S)$ and (2.1), that on the one hand $\mathbb{E}(\tilde{\Sigma})\le 2K\sigma^2\Sigma$.

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

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

we deduce from (6.2) that for all $\lambda\in\Lambda$,

$$\|f-\widehat{f}_{\widehat{\lambda}}\|^2\le C^{-1}\big[\|f-\widehat{f}_\lambda\|^2+A(\widehat{f}_\lambda,\mathbb{S}_\lambda)+\tilde{\Sigma}+\delta\big]\tag{6.3}$$

with

$$C^{-1}=C^{-1}(K,\alpha)=2\frac{(1+\alpha-K^{-1})(\alpha+2(1+K^{-1}))}{\alpha(1-K^{-1})},\tag{6.4}$$

and (2.10) follows by taking the expectation on both sides of (6.3). Note that provided that

$$\inf_{\lambda\in\Lambda}\big[\|f-\widehat{f}_\lambda\|^2+A(\widehat{f}_\lambda,\mathbb{S}_\lambda)\big]$$

is measurable, we have actually proved the stronger inequality

$$C\mathbb{E}\big[\|f-\widehat{f}_{\widehat{\lambda}}\|^2\big]\le\mathbb{E}\Big[\inf_{\lambda\in\Lambda}\big\{\|f-\widehat{f}_\lambda\|^2+A(\widehat{f}_\lambda,\mathbb{S}_\lambda)\big\}\Big]+\sigma^2\Sigma+\delta.\tag{6.5}$$

Let us now turn to the second part of the theorem. Since equality holds in (2.4), under Assumption 3 by (2.3)

$$\mathrm{pen}(S)=K\mathrm{pen}_\Delta(S)\le C(\kappa,K)\big(\dim(S)\vee\Delta(S)\big)\quad\forall S\in\mathbb{S}.$$

Combining this bound with Assumption 3 we obtain from simple algebra that for all $S\in\mathbb{S}$ and $\lambda\in\Lambda$

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

where C is a positive constant depending on K and κ only. Putting together this bound with (6.5) and $\mathbb{E}[(\|\varepsilon\|^2-2n\sigma^2)_+]\le 3\sigma^2$, gives (2.11).

6.2. *Proof of Proposition* 1

For all $\lambda\in\Lambda$ and $f\in S$, $\|f-\widehat{f}_\lambda\|\ge\|\Pi_S\widehat{f}_\lambda-\widehat{f}_\lambda\|$ and hence,

$$\|f-\widehat{f}_{\widehat{\lambda}}\|^2\ge\inf_{\lambda\in\Lambda}\|f-\widehat{f}_\lambda\|^2\ge\frac{1}{2}\inf_{\lambda\in\Lambda}\big[\|f-\widehat{f}_\lambda\|^2+\|\Pi_S\widehat{f}_\lambda-\widehat{f}_\lambda\|^2\big].$$

Besides, since the minimax rate of estimation over S is of order $\dim(S)\sigma^2$, for some universal constant C,

$$C\sup_{f\in S}\mathbb{E}\big[\|f-\widehat{f}_{\widehat{\lambda}}\|^2\big]\geq \dim(S)\sigma^2.$$

Putting these bounds together lead to the result.

6.3. *Proof of Corollary 2*

Since Assumptions 1 to 4 are fulfilled and $\mathbb{F}$ is finite, we may apply Theorem 2.1 and take $\delta=0$. By using (2.11), we have for some C depending on K,α and κ,

$$\begin{aligned}&C\mathbb{E}\big[\|f-\widehat{f}_{\widehat{\lambda}}\|^2\big]\\&\quad\leq\inf_{\lambda\in\Lambda}\big\{\mathbb{E}\big[\|f-\widehat{f}_\lambda\|^2\big]+\mathbb{E}\big[\|\widehat{f}_\lambda-\Pi_{S_\lambda}\widehat{f}_\lambda\|^2\big]+a\big(1+\dim(S_\lambda)\big)\sigma^2\big\}.\end{aligned}$$

For all $\lambda\in\Lambda$,

$$\begin{aligned}\mathbb{E}\big[\|f-\widehat{f}_\lambda\|^2\big]&=\|f-A_\lambda f\|^2+\mathbb{E}\big[\|A_\lambda\varepsilon\|^2\big]\\&=\|f-A_\lambda f\|^2+\operatorname{Tr}\big(A_\lambda^*A_\lambda\big)\sigma^2\\&\geq\max\big\{\|f-A_\lambda f\|^2,\operatorname{Tr}\big(A_\lambda^*A_\lambda\big)\sigma^2\big\}\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\big[\|\widehat{f}_\lambda-\Pi_{S_\lambda}\widehat{f}_\lambda\|^2\big]&=\big\|(I-\Pi_{S_\lambda})A_\lambda f\big\|^2+\mathbb{E}\big[\big\|(I-\Pi_{S_\lambda})A_\lambda\varepsilon\big\|^2\big]\\&\leq 2\max\big\{\big\|(I-\Pi_{S_\lambda})A_\lambda f\big\|^2,\mathbb{E}\big[\|A_\lambda\varepsilon\|^2\big]\big\}\\&=2\max\big\{\big\|(I-\Pi_{S_\lambda})A_\lambda f\big\|^2,\operatorname{Tr}\big(A_\lambda^*A_\lambda\big)\sigma^2\big\}\end{aligned}$$

and hence, Corollary 2 follows from the next lemma.

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

(i) $\|(I-\Pi_{S_\lambda})A_\lambda f\|\leq\|f-A_\lambda f\|$,
(ii) $\dim(S_\lambda)\leq 4\operatorname{Tr}\big(A_\lambda^*A_\lambda\big)$.

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

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

where $s_1\geq\cdots\geq s_{m_\lambda}$ are the singular values of $A_\lambda^+-\overline{\Pi}_\lambda$ counted with their multiplicity and $(v_1,\dots,v_{m_\lambda})$ is an orthonormal family of right-singular vectors associated to $(s_1,\dots,s_{m_\lambda})$. If $s_1<1$, then $S_\lambda=\mathbb{R}^n$ and we have $\|f-A_\lambda f\|\geq\|(I-\Pi_{S_\lambda})A_\lambda f\|=0$. Otherwise, $s_1\geq 1$, we may consider k_λ as the largest k such that $s_k\geq 1$ and derive that

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

which proves the assertion (i).

For the bound (ii), 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

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

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

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

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

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

which conclude the proof of Lemma 1. □

6.4. *Proof of Corollary* 4

Along the section, we write S_* for S_{m^*} and $\widehat{S}_\lambda$ for $S_{\widehat{m}(\lambda)}$ for short. First, note that when $D_{\max} \leq \kappa n/(2\log p)$, Assumption 3 holds. Since $\Sigma \leq 1 + \log(1+p)$, by using (2.11) with $\delta = 0$ we have

$$C\mathbb{E}\big[\|f - \widehat{f}_{\widehat{\lambda}}\|^2\big] \leq \mathbb{E}\Big[\inf_{\lambda\in\Lambda} \|f - \Pi_{\widehat{S}_\lambda} Y\|^2 + \dim(\widehat{S}_\lambda)\log(p)\sigma^2\Big] + \big(1 + \log(p+1)\big)\sigma^2,$$

for some constant $C > 0$ depending on K and κ only. Writing B for the event $B = \{m^* \notin \widehat{\mathcal{M}}\}$, we have

$$\mathbb{E}\Big[\inf_{\lambda\in\Lambda}\big\{\|f - \Pi_{\widehat{S}_\lambda} Y\|^2 + \dim(\widehat{S}_\lambda)\log(p)\sigma^2\big\}\Big] \leq A_n + R_n',$$

where

$$\begin{aligned}A_n &= \mathbb{E}\big[\|f - \Pi_{S_*} Y\|^2 + \dim(S_*)\log(p)\sigma^2\big] = \big(1 + \log(p)\big)\dim(S_*)\sigma^2,\\ R_n' &= \mathbb{E}\Big[\inf_{\lambda\in\Lambda}\big\{\|f - \Pi_{\widehat{S}_\lambda} Y\|^2 + \dim(\widehat{S}_\lambda)\log(p)\sigma^2\big\}\mathbf{1}_B\Big].\end{aligned}$$

Let us bound R'_n. For all $\lambda \in \Lambda$, $\|f - \Pi_{\widehat{S}_\lambda} Y\|^2 \le \|f\|^2 + \|\varepsilon\|^2$. Since for all $S \in \mathbb{S}$, $\dim(S) \le D_{\max} \le \kappa n/(2\log p)$, by using (2.3), we have for all $\lambda \in \Lambda$, $\dim(\widehat{S}_\lambda)\log(p)\sigma^2 \le n\sigma^2$ and hence,

$$R'_n \le \mathbb{E}\big[\big(\|f\|^2 + \|\varepsilon\|^2 + n\sigma^2\big)\mathbf{1}_B\big].$$

Straightforward calculation shows that $\mathbb{E}[(\|f\|^2 + \|\varepsilon\|^2)^2] \le (\|f\|^2 + 2n\sigma^2)^2$ and hence, by Cauchy–Schwarz inequality

$$R'_n \le \big(\|f\|^2 + 3n\sigma^2\big)\sqrt{\mathbb{P}(B)}.$$

The result follows.

6.5. *Proof of Proposition* 2

Since $\mathrm{e}M \le \mathrm{e}^{n/4}$, $d(n,M) = n/(2\log(\mathrm{e}M)) \ge 2$ and hence $\mathbb{S}$ is not empty. Besides, for all $S_m \in \mathbb{S}_{\mathrm{Cv}}$

$$\big(\dim(S_m) \vee 1\big) \le \Delta(S_m) = |m| + \log\left[\binom{M}{|m|}\right] \le |m|(1 + \log M) \le \frac{n}{2},$$

and hence Assumptions 1 to 4 are satisfied with $\Sigma = 1$ and $\kappa = 1/2$. Besides, the set Λ_{Cv} being compact, $\lambda \mapsto \mathrm{crit}_\alpha(f_\lambda)$ admits a minimum over Λ_{Cv} (we shall come back to the minimization of this criterion at the end of the subsection) and hence we can take $\delta = 0$. By applying Theorem 2.1 and using (2.11), the resulting estimator $\widehat{f}_{\mathrm{Cv}} = \widehat{f}_{\widehat{\lambda}}$ satisfies for some universal constant $C > 0$

$$C\mathbb{E}\big[\|f - \widehat{f}_{\mathrm{Cv}}\|^2\big] \le \inf_{g \in \mathbb{F}_\Lambda}\big\{\|f - g\|^2 + \overline{A}(g, \mathbb{S})\big\}, \tag{6.6}$$

where

$$\overline{A}(g, \mathbb{S}) = \inf_{S \in \mathbb{S}}\big[\|g - \Pi_S g\|^2 + \big(\dim(S) \vee \Delta(S)\big)\sigma^2\big]. \tag{6.7}$$

Case $\rho \le M \wedge d(n,M)$
If $\rho \le 2\sqrt{\log(\mathrm{e}M/\rho)}$, we choose $S = \{0\}$ in (6.9). By convexity, for all g in the convex hull of the ϕ_j, $\sigma^{-1}\|g\| \le \rho$ and therefore,

$$\overline{A}(g, \mathbb{S}) \le \left(\frac{\|g\|^2}{\sigma^2} + 1\right)\sigma^2 \le 2\big(\rho^2 \vee 1\big)\sigma^2. \tag{6.8}$$

Let us now turn to the situation where $2\sqrt{\log(\mathrm{e}M/\rho)} < \rho \le M \wedge d(n,M)$. We bound $\overline{A}(g, \mathbb{S})$ from above by using the following approximation result which is due to Maurey. A proof is available in Makovoz [38].

Lemma 2. *For all g in the convex hull $\mathbb{F}_\Lambda$ of the ϕ_j and all $D \ge 1$, there exists $m \subset \{1, \dots, M\}$ such that $|m| = (2D) \wedge M$ and*

$$\|g - \Pi_{S_m} g\|^2 \le 4D^{-1} \sup_{j=1,\dots,M} \|\phi_j\|^2.$$

By using this lemma, we get that

$$\overline{A}(g, \mathbb{S}) \le \inf_D\left[\frac{4\rho^2}{D} + \big[(2D) \wedge M\big]\left(1 + \log\left(\frac{\mathrm{e}M}{[(2D) \wedge M]}\right)\right)\right]\sigma^2, \tag{6.9}$$

where the infimum runs among all $D \ge 1$ such that $(2D) \wedge M \le d(n,M)$. We choose $D = D^*$ as the integer part of

$$d^* = \frac{\rho}{2\sqrt{\log(\mathrm{e}M/\rho)}} \ge 1.$$

Note that $D^* \geq 1$ and under the assumption $\rho \leq d(n, M) \wedge M$ we have

$$\left(2D^*\right) \wedge M \leq 2D^* \leq \frac{\rho}{\sqrt{\log(\mathrm{e}M/\rho)}} \leq \rho \leq d(n, M).$$

For such a choice of $D = D^*$ in (6.9) and suitable numerical constants $C > 0$ we get

$$\overline{A}(g, \mathbb{S}) \leq C\sigma^2\rho\sqrt{\log(\mathrm{e}M/\rho)}. \tag{6.10}$$

Combining the two bounds (6.8) and (6.10), we finally obtain

$$\overline{A}(g, \mathbb{S}) \leq C'\sigma^2 \min\left(\rho\sqrt{\log(\mathrm{e}M/\rho)}, \rho^2 \vee 1\right)$$

when $\rho \leq M \wedge d(n, M)$.

Case $\rho > M \wedge d(n, M)$
If $M \leq d(n, M)$, we choose $S = S_{\{1,\dots,M\}}$ (which belongs to $\mathbb{S}$) and get

$$\overline{A}(g, \mathbb{S}) \leq 0 + \Delta(S_{\{1,\dots,M\}})\sigma^2 = M\sigma^2.$$

Otherwise, $M > d(n, M)$ and we choose D as the integer part of $d(n, M)/2$ and get from (6.9)

$$\overline{A}(g, \mathbb{S}) \leq C\left[\frac{\rho^2}{d(n, M)} + d(n, M)\log\left(\frac{\mathrm{e}M}{d(n, M)}\right)\right]\sigma^2,$$

which concludes the proof.

Computation of $\widehat{f}_{\mathrm{Cv}}$
Finally, concerning the computation of $\widehat{f}_{\mathrm{Cv}}$, note that

$$\begin{aligned}\inf_{\lambda\in\Lambda} \mathrm{crit}_\alpha(f_\lambda) &= \inf_{\lambda\in\Lambda}\inf_{S\in\mathbb{S}_{\mathrm{Cv}}}\left[\|Y - \Pi_S f_\lambda\|^2 + \alpha\|f_\lambda - \Pi_S f_\lambda\|^2 + \mathrm{pen}(S)\widehat{\sigma}_S^2\right] \\ &= \inf_{S\in\mathbb{S}_{\mathrm{Cv}}}\left\{\left[\inf_{\lambda\in\Lambda}\left(\|Y - \Pi_S f_\lambda\|^2 + \alpha\|f_\lambda - \Pi_S f_\lambda\|^2\right)\right] + \mathrm{pen}(S)\widehat{\sigma}_S^2\right\},\end{aligned}$$

and hence, one can solve the problem of minimizing $\mathrm{crit}_\alpha(f_\lambda)$ over $\lambda \in \Lambda$ by proceeding into two steps. First, for each S in the finite set $\mathbb{S}_{\mathrm{Cv}}$ minimize the convex criterion

$$\mathrm{crit}_\alpha(S, f_\lambda) = \|Y - \Pi_S f_\lambda\|^2 + \alpha\|f_\lambda - \Pi_S f_\lambda\|^2$$

over the convex (and compact set) Λ_{Cv}. Denote by $\widehat{f}_{\mathrm{Cv},S}$ the resulting minimizers. Then, minimize the quantity $\mathrm{crit}_\alpha(S, \widehat{f}_{\mathrm{Cv},S}) + \mathrm{pen}(S)\widehat{\sigma}_S^2$ for S varying among $\mathbb{S}_{\mathrm{Cv}}$. Denoting by $\widehat{S}$ such a minimizer, we have that $\widehat{f}_{\mathrm{Cv}} = \widehat{f}_{\mathrm{Cv},\widehat{S}}$.

6.6. *Proof of Proposition 3*

Under the assumption $M \leq \min\{\mathrm{e}^{n/4-1}, d(n, M)\}$, the families $\mathbb{S}_\lambda$ with $\lambda \in \{\mathrm{L}, \mathrm{MS}\}$ are subsets of $\mathbb{S} = \mathbb{S}_{\mathrm{Cv}}$ and Assumption 3 holds. We may therefore apply Theorem 2.1 (more precisely (2.11)) and get

$$C\mathbb{E}\left[\|f - \widehat{f}_{\widehat{\lambda}}\|^2\right] \leq \inf_{\lambda\in\{\mathrm{L},\mathrm{MS},\mathrm{Cv}\}}\left[\mathbb{E}\left[\|f - \widehat{f}_\lambda\|^2\right] + \mathbb{E}\left[\overline{A}(\widehat{f}_\lambda, \mathbb{S}_\lambda)\right]\right],$$

where $\overline{A}(\cdot, \cdot)$ is given by (6.7). It remains to bound from above the quantity $\mathbb{E}[\overline{A}(\widehat{f}_\lambda, \mathbb{S}_\lambda)]$ for each $\lambda \in \{\mathrm{L}, \mathrm{MS}, \mathrm{Cv}\}$. For $\lambda = \mathrm{L}$, $\widehat{f}_{\mathrm{L}} \in S_{\{1,\dots,M\}}$, $\Delta(S_{\{1,\dots,M\}}) = M$ and hence,

$$\mathbb{E}\left[\overline{A}(\widehat{f}_{\mathrm{L}}, \mathbb{S}_{\mathrm{L}})\right] = \mathbb{E}\left[\|f - \widehat{f}_{\mathrm{L}}\|^2\right] + M\sigma^2.$$

For $\lambda = \mathrm{MS}$, $\widehat{f}_{\mathrm{MS}} \in \mathbb{S}_{\mathrm{MS}}$ and for all $S_m \in \mathbb{S}_{\mathrm{MS}}$, $\dim(S_m) \le \Delta(S_m) = \log(\mathrm{e}M)$. Therefore

$$\mathbb{E}\bigl[A(\widehat{f}_{\mathrm{MS}}, \mathbb{S}_{\mathrm{MS}})\bigr] \le \mathbb{E}\bigl[\|f - \widehat{f}_{\mathrm{MS}}\|^2\bigr] + \log(\mathrm{e}M)\sigma^2.$$

Finally, let us turn to the case $\lambda = \mathrm{Cv}$ and denote by g the best approximation of f in $\mathcal{C}$. Since $\widehat{f}_{\mathrm{Cv}} \in \mathcal{C}$, for all $S \in \mathbb{S}_{\mathrm{Cv}}$,

$$\begin{aligned}\|\widehat{f}_{\mathrm{Cv}} - \Pi_S \widehat{f}_{\mathrm{Cv}}\| \le \|\widehat{f}_{\mathrm{Cv}} - \Pi_S g\| &= \|\widehat{f}_{\mathrm{Cv}} - f + f - g + g - \Pi_S g\| \\ &\le 2\|f - \widehat{f}_{\mathrm{Cv}}\| + \|g - \Pi_S g\|,\end{aligned}$$

and hence

$$8^{-1}\mathbb{E}\bigl[\overline{A}(\widehat{f}_{\mathrm{Cv}}, \mathbb{S}_{\mathrm{Cv}})\bigr] \le \mathbb{E}\bigl[\|f - \widehat{f}_{\mathrm{Cv}}\|^2\bigr] + \overline{A}(g, \mathbb{S}_{\mathrm{Cv}}).$$

By arguing as in Section (5.1.3), we deduce that under the assumption that $\mathrm{e}M \le \mathrm{e}^{n/4}$,

$$C'\mathbb{E}\bigl[A(\widehat{f}_{\mathrm{Cv}}, \mathbb{S}_{\mathrm{Cv}})\bigr] \le \mathbb{E}\bigl[\|f - \widehat{f}_{\mathrm{Cv}}\|^2\bigr] + B_{\mathrm{Cv}}\sigma^2.$$

By putting these bounds together we get the result.

Appendix

A.1. *Computation of* $\mathrm{pen}_\Delta(S)$

The penalty $\mathrm{pen}_\Delta(S)$, defined at equation (2.2), is linked to the EDkhi function introduced in Baraud et al. [7] (see Definition 3), via the following formula:

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

Therefore, according to the result given in Section 6.1 in Baraud et al. [7], $\mathrm{pen}_\Delta(S)$ is the solution in x of the equation

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

A.2. *Simulated examples*

The collection $\mathcal{E}$ is composed of several collections $\mathcal{E}_1, \dots, \mathcal{E}_{11}$ 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 $e \in \{1, \dots, 10\}$, 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:

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

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

$$\mathcal{I} = \bigl\{(100, 50), (100, 100), (100, 1000), (200, 100), (200, 200)\bigr\} \tag{A.1}$$

in Section 4.3.2, and

$$\mathcal{I} = \bigl\{(100, 50), (100, 100), (200, 100), (200, 200)\bigr\} \tag{A.2}$$

in Section 4.3.3.

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 $\rho_1 = 0.39$ and $\rho_2 = 0.23$. 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 (A.2), and 825 for $\mathcal{I}$ chosen as in (A.1). For some of the examples, the Lasso estimators were highly biased leading to high values of the ratio $O_{\mathrm{ex}}/n\sigma^2$, see Equation (4.4). In these cases, our procedure that tends to choose an estimator with small dimension, leads to very high value of the risk. 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, the others correspond to matrices X that are nearly orthogonal.

Computation time. The computation time for tuning the Lasso parameter depends on n, p, the maximum number of steps in the Lasso algorithm, `max.steps`, and, for our procedure, it depends on the cardinality of $\mathbb{S}$ or equivalently on $D_{\max}$ (see Equation (4.1)). For example, for $n = p =$ `max.steps` $= 100$, the CV procedure using `elasticnet`, takes 4 s, and the pen_Δ procedure, taking $D_{\max} = \min\{p, n/\log(p)\}$, takes 0.2 s.

A.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 4.3.2. The collection $\widehat{\mathcal{M}}_{\text{Lasso}} = \{\widehat{m}(1), \ldots, \widehat{m}(D_{\max})\}$ where $\widehat{m}(h)$ is the set of indices corresponding to the predictors returned by the LARS-Lasso algorithm at step $h \in \{1, \ldots, D_{\max}\}$ (see Section 4.3.2).

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 [30]. Tibshirani [46] 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 $h \in H_{\text{ridge}}$, 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 $|\widehat{\beta}_{j_1}(h)| > \cdots > |\widehat{\beta}_{j_p}(h)|$ and set

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

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

The elastic net procedure proposed by Zou and Hastie [58] mixes the ℓ_1 and ℓ_2 penalties of the Lasso and the ridge procedures. Let H_{ridge} be a grid of values for the tuning parameter h of the ℓ_2 penalty. We choose $\widehat{\mathcal{M}}_{\text{en}} = \{M_{(\text{en},h)} : h \in H_{\text{ridge}}\}$ where $M_{(\text{en},h)}$ 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 $h \in H_{\text{ridge}}$ the collection $M_{(\text{en},h)}$ can be conveniently computed by first calculating the ridge regression coefficients and then applying the LARS-Lasso algorithm, see Zou and Hastie [58].

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 [11]. 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 [27]. 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 [57] 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 $|\beta_j|/|\widetilde{\beta}_j|^\gamma$, $j = 1, \ldots, p$ 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. [31] 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 [12] 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 [20], Genuer et al. [22], Strobl et al. [43,44].

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

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

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

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

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 $D_{\max} \le \min(n-2, p)$. 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, $H_{\text{pls}} = 1, \ldots, 5$.

References

[1] S. Arlot. Rééchantillonnage et Sélection de modèles. Ph.D. thesis, Univ. Paris XI, 2007.

[2] S. Arlot. Model selection by resampling penalization. *Electron. J. Stat.* **3** (2009) 557–624. MR2519533

[3] S. Arlot and F. Bach. Data-driven calibration of linear estimators with minimal penalties, 2011. Available at arXiv:0909.1884v2.

[4] S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Stat. Surv.* **4** (2010) 40–79. MR2602303

[5] Y. Baraud. Model selection for regression on a fixed design. *Probab. Theory Related Fields* **117** (2000) 467–493. MR1777129

[6] Y. Baraud. Estimator selection with respect to Hellinger-type risks. *Probab. Theory Related Fields* **151** (2011) 353–401. MR2834722

[7] Y. Baraud, C. Giraud and S. Huet. Gaussian model selection with an unknown variance. *Ann. Statist.* **37** (2009) 630–672. MR2502646

[8] Y. Baraud, C. Giraud and S. Huet. Estimator selection in the Gaussian setting, 2010. Available at arXiv:1007.2096v1.

[9] L. Birgé. Model selection via testing: An alternative to (penalized) maximum likelihood estimators. *Ann. Inst. Henri Poincaré Probab. Stat.* **42** (2006) 273–325. MR2219712

[10] L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)* **3** (2001) 203–268. MR1848946

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

[12] L. Breiman. Random forests. *Mach. Learn.* **45** (2001) 5–32.

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

[14] E. Candès and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** (2007) 2313–2351. MR2382644

[15] Y. Cao and Y. Golubev. On oracle inequalities related to smoothing splines. *Math. Methods Statist.* **15** (2006) 398–414. MR2301659

[16] O. Catoni. Mixture approach to universal model selection. Technical report, Ecole Normale Supérieure, France, 1997.

[17] O. Catoni. 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, Berlin, 2004. MR2163920

[18] A. Celisse. Model selection via cross-validation in density estimation, regression, and change-points detection. Ph.D. thesis, Univ. Paris XI, 2008.

[19] S. S. Chen, D. L. Donoho and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.* **20** (1998) 33–61 (electronic). MR1639094

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

[21] B. Efron, T. Hastie, I. Johnstone and R. Tibshirani. Least angle regression. *Ann. Statist.* **32** (2004) 407–499. With discussion, and a rejoinder by the authors. MR2060166

[22] R. Genuer, J.-M. Poggi and C. Tuleau-Malot. Variable selection using random forests. *Patter Recognition Lett.* **31** (2010) 2225–2236.

[23] C. Giraud. Mixing least-squares estimators when the variance is unknown. *Bernoulli* **14** (2008) 1089–1107. MR2543587

[24] C. Giraud, S. Huet and N. Verzelen. High-dimensional regression with unknown variance. *Statist. Sci.* **27** (2013) 500–518.

[25] A. Goldenshluger. A universal procedure for aggregating estimators. *Ann. Statist.* **37** (2009) 542–568. MR2488362

[26] A. Goldenshluger and O. Lepski. Structural adaptation via $\mathbb{L}_p$-norm oracle inequalities. *Probab. Theory Related Fields* **143** (2009) 41–71. MR2449122

[27] I. Helland. Partial least squares regression. In *Encyclopedia of Statistical Sciences*, 2nd edition **9** 5957–5962. S. Kotz, N. Balakrishnan, C. Read, B. Vidakovic and N. Johnston (Eds.). Wiley, New York, 2006.

[28] I. Helland. Some theoretical aspects of partial least squares regression. *Chemometrics and Intelligent Laboratory Systems* **58** (2001) 97–107.

[29] A. Hoerl and R. Kennard. Ridge regression: Bayes estimation for nonorthogonal problems. *Technometrics* **12** (1970) 55–67.

[30] A. Hoerl and R. Kennard. Ridge regression. In *Encyclopedia of Statistical Sciences*, 2nd edition **11** 7273–7280. S. Kotz, N. Balakrishnan, C. Read, B. Vidakovic and N. Johnston (Eds.). Wiley, New York, 2006.

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

[32] A. Juditsky and A. Nemirovski. Functional aggregation for nonparametric regression. *Ann. Statist.* **28** (2000) 681–712. MR1792783

[33] O. V. Lepskiĭ. A problem of adaptive estimation in Gaussian white noise. *Teor. Veroyatnost. i Primenen.* **35** (1990) 459–470. MR1091202

[34] O. V. Lepskiĭ. Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Teor. Veroyatnost. i Primenen.* **36** (1991) 645–659. MR1147167

[35] O. V. Lepskiĭ. Asymptotically minimax adaptive estimation. II. Schemes without optimal adaptation. Adaptive estimates. *Teor. Veroyatnost. i Primenen.* **37** (1992) 468–481. MR1214353

[36] O. V. Lepskiĭ. On problems of adaptive estimation in white Gaussian noise. In *Topics in Nonparametric Estimation* 87–106. *Adv. Soviet Math.* **12**. Amer. Math. Soc., Providence, RI, 1992. MR1191692

[37] G. Leung and A. R. Barron. Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory* **52** (2006) 3396–3410. MR2242356

[38] Y. Makovoz. Random approximants and neural networks. *J. Approx. Theory* **85** (1996) 98–109. MR1382053

[39] E. A. Nadaraya. On estimating regression. *Theory Probab. Appl.* **9** (1964) 141–142.

[40] A. Nemirovski. Topics in non-parametric statistics. In *Lectures on probability theory and statistics (Saint-Flour, 1998)* 85–277. *Lecture Notes in Math.* **1738**. Springer, Berlin, 2000. MR1775640

[41] P. Rigollet and A. B. Tsybakov. Linear and convex aggregation of density estimators. *Math. Methods Statist.* **16** (2007) 260–280. MR2356821

[42] J. Salmon and A. Dalalyan. Optimal aggregation of affine estimators. *J. Mach. Learn. Res.* **19** (2011) 635–660.

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

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

[45] M. Tenenhaus. *La régression PLS*. Éditions Technip, Paris. Théorie et pratique, 1998. [Theory and application]. MR1645125

[46] R. Tibshirani. Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58** (1996) 267–288. MR1379242

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

[48] G. S. Watson. Smooth regression analysis. *Sankhyā Ser. A* **26** (1964) 359–372. MR0185765

[49] M. Wegkamp. Model selection in nonparametric regression. *Ann. Statist.* **31** (2003) 252–273. MR1962506

[50] Y. Yang. Model selection for nonparametric regression. *Statist. Sinica* **9** (1999) 475–499. MR1707850

[51] Y. Yang. Combining different procedures for adaptive regression. *J. Multivariate Anal.* **74** (2000) 135–161. MR1790617

[52] Y. Yang. Mixing strategies for density estimation. *Ann. Statist.* **28** (2000) 75–87. MR1762904

[53] Y. Yang. Adaptive regression by mixing. *J. Amer. Statist. Assoc.* **96** (2001) 574–588. MR1946426

[54] T. Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural Comput.* **17** (2005) 2077–2098. MR2175849

[55] T. Zhang. Adaptive forward-backward greedy algorithm for learning sparse representations. Technical report, Rutgers Univ., NJ, 2008.

[56] P. Zhao and B. Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** (2006) 2541–2563. MR2274449

[57] H. Zou. The adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** (2006) 1418–1429. MR2279469

[58] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** (2005) 301–320. MR2137327

[59] H. Zou, T. Hastie and R. Tibshirani On the "degrees of freedom" of the Lasso. *Ann. Statist.* **35** (2007) 2173–2192. MR2363967