Data-driven calibration of linear estimators with minimal penalties

Sylvain Arlot *
CNRS ; Willow Project-Team
Laboratoire d'Informatique de
l'Ecole Normale Superieure
(CNRS/ENS/INRIA UMR 8548)
23, avenue d'Italie, F-75013 Paris, France
sylvain.arlot@ens.fr

Francis Bach †
INRIA ; Willow Project-Team
Laboratoire d'Informatique de
l'Ecole Normale Superieure
(CNRS/ENS/INRIA UMR 8548)
23, avenue d'Italie, F-75013 Paris, France
francis.bach@ens.fr

Abstract

This paper tackles the problem of selecting among several linear estimators in non-parametric regression; this includes model selection for linear regression, the choice of a regularization parameter in kernel ridge regression or spline smoothing, and the choice of a kernel in multiple kernel learning. We propose a new algorithm which first estimates consistently the variance of the noise, based upon the concept of minimal penalty which was previously introduced in the context of model selection. Then, plugging our variance estimate in Mallows' C_L penalty is proved to lead to an algorithm satisfying an oracle inequality. Simulation experiments with kernel ridge regression and multiple kernel learning show that the proposed algorithm often improves significantly existing calibration procedures such as 10-fold cross-validation or generalized cross-validation.

1 Introduction

Kernel-based methods are now well-established tools for supervised learning, allowing to perform various tasks, such as regression or binary classification, with linear and non-linear predictors [1, 2]. A central issue common to all regularization frameworks is the choice of the regularization parameter: while most practitioners use cross-validation procedures to select such a parameter, data-driven procedures not based on cross-validation are rarely used. The choice of the kernel, a seemingly unrelated issue, is also important for good predictive performance: several techniques exist, either based on cross-validation, Gaussian processes or multiple kernel learning [3, 4, 5].

In this paper, we consider least-squares regression and cast these two problems as the problem of selecting among several *linear estimators*, where the goal is to choose an estimator with a quadratic risk which is as small as possible. This problem includes for instance model selection for linear regression, the choice of a regularization parameter in kernel ridge regression or spline smoothing, and the choice of a kernel in multiple kernel learning (see Section 2).

The main contribution of the paper is to extend the notion of *minimal penalty* [6, 7] to all discrete classes of linear operators, and to use it for defining a fully data-driven selection algorithm satisfying a non-asymptotic oracle inequality. Our new theoretical results presented in Section 4 extend similar results which were limited to unregularized least-squares regression (i.e., projection operators). Finally, in Section 5, we show that our algorithm improves the performances of classical selection procedures, such as GCV [8] and 10-fold cross-validation, for kernel ridge regression or multiple kernel learning, for moderate values of the sample size.

*http://www.di.ens.fr/∼arlot/
†http://www.di.ens.fr/∼fbach/

2 Linear estimators

In this section, we define the problem we aim to solve and give several examples of linear estimators.

2.1 Framework and notation

Let us assume that one observes

$$Y_i = f(x_i) + \varepsilon_i \in \mathbb{R} \qquad \text{for} \quad i = 1 \dots n \ ,$$

where $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. centered random variables with $\mathbb{E}[\varepsilon_i^2] = \sigma^2$ unknown, f is an unknown measurable function $\mathcal{X} \mapsto \mathbb{R}$ and $x_1, \dots, x_n \in \mathcal{X}$ are deterministic design points. No assumption is made on the set $\mathcal{X}$. The goal is to reconstruct the signal $F = (f(x_i))_{1 \leq i \leq n} \in \mathbb{R}^n$, with some estimator $\widehat{F} \in \mathbb{R}^n$, depending only on $(x_1, Y_1), \dots, (x_n, Y_n)$, and having a small quadratic risk $n^{-1}\|\widehat{F} - F\|_2^2$, where $\forall t \in \mathbb{R}^n$, we denote by $\|t\|_2$ the ℓ_2-norm of t, defined as $\|t\|_2^2 := \sum_{i=1}^n t_i^2$.

In this paper, we focus on *linear estimators* $\widehat{F}$ that can be written as a linear function of $Y = (Y_1, \dots, Y_n) \in \mathbb{R}^n$, that is, $\widehat{F} = AY$, for some (deterministic) $n \times n$ matrix A. Here and in the rest of the paper, vectors such as Y or F are assumed to be column-vectors. We present in Section 2.2 several important families of estimators of this form. The matrix A may depend on $x_1, \dots, x_n$ (which are known and deterministic), but not on Y, and may be parameterized by certain quantities—usually regularization parameter or kernel combination weights.

2.2 Examples of linear estimators

In this paper, our theoretical results apply to matrices A which are symmetric positive semi-definite, such as the ones defined below.

Ordinary least-squares regression / model selection. If we consider linear predictors from a design matrix $X \in \mathbb{R}^{n \times p}$, then $\widehat{F} = AY$ with $A = X(X^\top X)^{-1}X^\top$, which is a projection matrix (i.e., $A^\top A = A$); $\widehat{F} = AY$ is often called a *projection estimator*. In the variable selection setting, one wants to select a subset $J \subset \{1, \dots, p\}$, and matrices A are parameterized by J.

Kernel ridge regression / spline smoothing. We assume that a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is given, and we are looking for a function $f : \mathcal{X} \to \mathbb{R}$ in the associated reproducing kernel Hilbert space (RKHS) $\mathcal{F}$, with norm $\|\cdot\|_{\mathcal{F}}$. If K denotes the $n \times n$ kernel matrix, defined by $K_{ab} = k(x_a, x_b)$, then the ridge regression estimator—a.k.a. spline smoothing estimator for spline kernels [9]—is obtained by minimizing with respect to $f \in \mathcal{F}$ [2]:

$$\frac{1}{n}\sum_{i=1}^n (Y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{F}}^2 \ .$$

The unique solution is equal to $\widehat{f} = \sum_{i=1}^n \alpha_i k(\cdot, x_i)$, where $\alpha = (K + n\lambda I)^{-1}Y$. This leads to the smoothing matrix $A_\lambda = K(K + n\lambda I_n)^{-1}$, parameterized by the regularization parameter $\lambda \in \mathbb{R}_+$.

Multiple kernel learning / Group Lasso / Lasso. We now assume that we have p different kernels k_j, feature spaces $\mathcal{F}_j$ and feature maps $\Phi_j : \mathcal{X} \to \mathcal{F}_j$, $j = 1, \dots, p$. The group Lasso [10] and multiple kernel learning [11, 5] frameworks consider the following objective function

$$J(f_1, \dots, f_p) = \tfrac{1}{n}\sum_{i=1}^n \big(y_i - \textstyle\sum_{j=1}^p \langle f_j, \Phi_j(x_i)\rangle\big)^2 + 2\lambda \sum_{j=1}^p \|f_j\|_{\mathcal{F}_j} = L(f_1, \dots, f_p) + 2\lambda \sum_{j=1}^p \|f_j\|_{\mathcal{F}_j} \ .$$

Note that when $\Phi_j(x)$ is simply the j-th coordinate of $x \in \mathbb{R}^p$, we get back the penalization by the ℓ^1-norm and thus the regular Lasso [12].

Using $a^{1/2} = \min_{b \geqslant 0} \frac{1}{2}\{\frac{a}{b} + b\}$, we obtain a variational formulation of the sum of norms $2\sum_{j=1}^p \|f_j\| = \min_{\eta \in \mathbb{R}_+^p} \sum_{j=1}^p \left\{ \frac{\|f_j\|^2}{\eta_j} + \eta_j \right\}$. Thus, minimizing $J(f_1, \dots, f_p)$ with respect to $(f_1, \dots, f_p)$ is equivalent to minimizing with respect to $\eta \in \mathbb{R}_+^p$ (see [5] for more details):

$$\min_{f_1, \dots, f_p} L(f_1, \dots, f_p) + \lambda \sum_{j=1}^p \frac{\|f_j\|^2}{\eta_j} + \lambda \sum_{j=1}^p \eta_j = \frac{1}{n} y^\top \big(\textstyle\sum_{j=1}^p \eta_j K_j + n\lambda I_n\big)^{-1} y + \lambda \sum_{j=1}^p \eta_j \ ,$$

where I_n is the $n \times n$ identity matrix. Moreover, given η, this leads to a smoothing matrix of the form

$$A_{\eta,\lambda} = (\sum_{j=1}^p \eta_j K_j)(\sum_{j=1}^p \eta_j K_j + n\lambda I_n)^{-1} \ , \tag{1}$$

parameterized by the regularization parameter $\lambda \in \mathbb{R}_+$ and the kernel combinations in $\mathbb{R}_+^p$—note that it depends only on $\lambda^{-1}\eta$, which can be grouped in a single parameter in $\mathbb{R}_+^p$.

Thus, the Lasso/group lasso can be seen as particular (convex) ways of optimizing over η. In this paper, we propose a non-convex alternative with better statistical properties (oracle inequality in Theorem 1). Note that in our setting, finding the solution of the problem is hard in general since the optimization is not convex. However, while the model selection problem is by nature combinatorial, our optimization problems for multiple kernels are all differentiable and are thus amenable to gradient descent procedures—which only find local optima.

Non symmetric linear estimators. Other linear estimators are commonly used, such as nearest-neighbor regression or the Nadaraya-Watson estimator [13]; those however lead to non symmetric matrices A, and are not entirely covered by our theoretical results.

3 Linear estimator selection

In this section, we first describe the statistical framework of linear estimator selection and introduce the notion of minimal penalty.

3.1 Unbiased risk estimation heuristics

Usually, several estimators of the form $\widehat{F} = AY$ can be used. The problem that we consider in this paper is then to select one of them, that is, to choose a matrix A. Let us assume that a family of matrices $(A_\lambda)_{\lambda\in\Lambda}$ is given (examples are shown in Section 2.2), hence a family of estimators $(\widehat{F}_\lambda)_{\lambda\in\Lambda}$ can be used, with $\widehat{F}_\lambda := A_\lambda Y$. The goal is to choose *from data* some $\widehat{\lambda} \in \Lambda$, so that the quadratic risk of $\widehat{F}_{\widehat{\lambda}}$ is as small as possible.

The best choice would be the *oracle*:

$$\lambda^\star \in \arg\min_{\lambda\in\Lambda} \left\{ n^{-1}\|\widehat{F}_\lambda - F\|_2^2 \right\} \ ,$$

which cannot be used since it depends on the unknown signal F. Therefore, the goal is to define a data-driven $\widehat{\lambda}$ satisfying an *oracle inequality*

$$n^{-1}\|\widehat{F}_{\widehat{\lambda}} - F\|_2^2 \le C_n \inf_{\lambda\in\Lambda} \left\{ n^{-1}\|\widehat{F}_\lambda - F\|_2^2 \right\} + R_n \ , \tag{2}$$

with large probability, where the leading constant C_n should be close to 1 (at least for large n) and the remainder term R_n should be negligible compared to the risk of the oracle.

Many classical selection methods are built upon the "unbiased risk estimation" heuristics: If $\widehat{\lambda}$ minimizes a criterion $\mathrm{crit}(\lambda)$ such that

$$\forall \lambda \in \Lambda, \qquad \mathbb{E}\left[\mathrm{crit}(\lambda)\right] \approx \mathbb{E}\left[n^{-1}\|\widehat{F}_\lambda - F\|_2^2 \right] \ ,$$

then $\widehat{\lambda}$ satisfies an oracle inequality such as in Eq. (2) with large probability. For instance, cross-validation [14, 15] and generalized cross-validation (GCV) [8] are built upon this heuristics.

One way of implementing this heuristics is penalization, which consists in minimizing the sum of the empirical risk and a penalty term, i.e., using a criterion of the form:

$$\mathrm{crit}(\lambda) = n^{-1}\|\widehat{F}_\lambda - Y\|_2^2 + \mathrm{pen}(\lambda) \ .$$

The unbiased risk estimation heuristics, also called Mallows' heuristics, then leads to the *ideal (deterministic) penalty*

$$\mathrm{pen}_{\mathrm{id}}(\lambda) := \mathbb{E}\left[n^{-1}\|\widehat{F}_\lambda - F\|_2^2 \right] - \mathbb{E}\left[n^{-1}\|\widehat{F}_\lambda - Y\|_2^2 \right] \ .$$

When $\widehat{F}_\lambda = A_\lambda Y$, we have:

$$\|\widehat{F}_\lambda - F\|_2^2 = \|(A_\lambda - I_n)F\|_2^2 + \|A_\lambda \varepsilon\|_2^2 + 2\,\langle A_\lambda \varepsilon,\, (A_\lambda - I_n)F\rangle \quad , \tag{3}$$

$$\|\widehat{F}_\lambda - Y\|_2^2 = \|\widehat{F}_\lambda - F\|_2^2 + \|\varepsilon\|_2^2 - 2\,\langle \varepsilon,\, A_\lambda \varepsilon\rangle + 2\,\langle \varepsilon,\, (I_n - A_\lambda)F\rangle \quad , \tag{4}$$

where $\varepsilon = Y - F \in \mathbb{R}^n$ and $\forall t, u \in \mathbb{R}^n$, $\langle t,\, u\rangle = \sum_{i=1}^n t_i u_i$. Since ε is centered with covariance matrix $\sigma^2 I_n$, Eq. (3) and Eq. (4) imply that

$$\mathrm{pen}_{\mathrm{id}}(\lambda) = \frac{2\sigma^2 \operatorname{tr}(A_\lambda)}{n} \quad , \tag{5}$$

up to the term $-\mathbb{E}[n^{-1}\|\varepsilon\|_2^2] = -\sigma^2$, which can be dropped off since it does not vary with λ.

Note that $\mathrm{df}(\lambda) = \operatorname{tr}(A_\lambda)$ is called the *effective dimensionality* or *degrees of freedom* [16], so that the ideal penalty in Eq. (5) is proportional to the dimensionality associated with the matrix A_λ— for projection matrices, we get back the dimension of the subspace, which is classical in model selection.

The expression of the ideal penalty in Eq. (5) led to several selection procedures, in particular Mallows' C_L (called C_p in the case of projection estimators) [17], where σ^2 is replaced by some estimator $\widehat{\sigma^2}$. The estimator of σ^2 usually used with C_L is based upon the value of the empirical risk at some λ_0 with $\mathrm{df}(\lambda_0)$ large; it has the drawback of overestimating the risk, in a way which depends on λ_0 and F [18]. GCV, which implicitly estimates σ^2, has the drawback of overfitting if the family $(A_\lambda)_{\lambda\in\Lambda}$ contains a matrix too close to I_n [19]; GCV also overestimates the risk even more than C_L for most A_λ (see (7.9) and Table 4 in [18]).

In this paper, we define an estimator of σ^2 directly related to the selection task which does not have similar drawbacks. Our estimator relies on the concept of minimal penalty, introduced by Birgé and Massart [6] and further studied in [7].

3.2 Minimal and optimal penalties

We deduce from Eq. (3) the *bias-variance decomposition* of the risk:

$$\mathbb{E}\left[n^{-1}\|\widehat{F}_\lambda - F\|_2^2\right] = n^{-1}\,\|(A_\lambda - I_n)F\|_2^2 + \frac{\operatorname{tr}(A_\lambda^\top A_\lambda)\sigma^2}{n} = \text{bias} + \text{variance} \quad , \tag{6}$$

and from Eq. (4) the expectation of the empirical risk:

$$\mathbb{E}\left[n^{-1}\|\widehat{F}_\lambda - Y\|_2^2 - \|\varepsilon\|_2^2\right] = n^{-1}\,\|(A_\lambda - I_n)F\|_2^2 - \frac{\left(2\operatorname{tr}(A_\lambda) - \operatorname{tr}(A_\lambda^\top A_\lambda)\right)\sigma^2}{n} \quad . \tag{7}$$

Note that the variance term in Eq. (6) is not proportional to the effective dimensionality $\mathrm{df}(\lambda) = \operatorname{tr}(A_\lambda)$ but to $\operatorname{tr}(A_\lambda^\top A_\lambda)$. Although several papers argue these terms are of the same order (for instance, they are equal when A_λ is a projection matrix), this may not hold in general. If A_λ is symmetric with a spectrum $\mathrm{Sp}(A_\lambda) \subset [0,1]$, as in all the examples of Section 2.2, we only have

$$0 \le \operatorname{tr}(A_\lambda^\top A_\lambda) \le \operatorname{tr}(A_\lambda) \le 2\operatorname{tr}(A_\lambda) - \operatorname{tr}(A_\lambda^\top A_\lambda) \le 2\operatorname{tr}(A_\lambda) \quad . \tag{8}$$

In order to give a first intuitive interpretation of Eq. (6) and Eq. (7), let us consider the kernel ridge regression example and assume that the risk and the empirical risk behave as their expectations in Eq. (6) and Eq. (7); see also Fig. 1. Completely rigorous arguments based upon concentration inequalities are developed in [20] and summarized in Section 4, leading to the same conclusion as the present informal reasoning.

First, as proved in [20], the bias $n^{-1}\,\|(A_\lambda - I_n)F\|_2^2$ is a decreasing function of the dimensionality $\mathrm{df}(\lambda) = \operatorname{tr}(A_\lambda)$, and the variance $\operatorname{tr}(A_\lambda^\top A_\lambda)\sigma^2 n^{-1}$ is an increasing function of $\mathrm{df}(\lambda)$, as well as $2\operatorname{tr}(A_\lambda) - \operatorname{tr}(A_\lambda^\top A_\lambda)$. Therefore, Eq. (6) shows that the optimal λ realizes the best trade-off between bias (which decreases with $\mathrm{df}(\lambda)$) and variance (which increases with $\mathrm{df}(\lambda)$), which is a classical fact in model selection.

Second, the expectation of the empirical risk in Eq. (7) can be decomposed into the bias and a negative variance term which is the opposite of

$$\mathrm{pen}_{\min}(\lambda) := n^{-1}\left(2\operatorname{tr}(A_\lambda) - \operatorname{tr}(A_\lambda^\top A_\lambda)\right)\sigma^2 \quad . \tag{9}$$

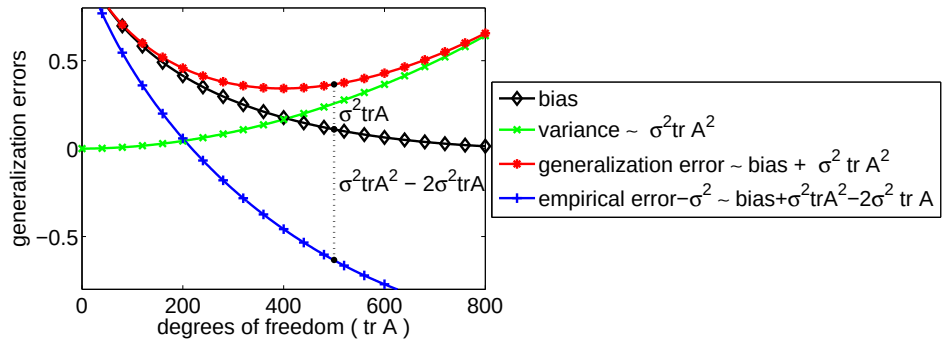

Figure 1: Bias-variance decomposition of the generalization error, and minimal/optimal penalties.

As suggested by the notation $\mathrm{pen}_{\min}$, we will show it is a *minimal penalty* in the following sense. If

$$\forall C \geq 0, \qquad \widehat{\lambda}_{\min}(C) \in \arg\min_{\lambda \in \Lambda} \left\{ n^{-1} \|\widehat{F}_\lambda - Y\|_2^2 + C\,\mathrm{pen}_{\min}(\lambda) \right\} \ ,$$

then, up to concentration inequalities that are detailed in Section 4.2, $\widehat{\lambda}_{\min}(C)$ behaves like a minimizer of

$$g_C(\lambda) = \mathbb{E}\left[n^{-1} \|\widehat{F}_\lambda - Y\|_2^2 + C\,\mathrm{pen}_{\min}(\lambda) \right] - n^{-1}\sigma^2 = n^{-1}\,\|(A_\lambda - I_n)F\|_2^2 + (C-1)\,\mathrm{pen}_{\min}(\lambda) \ .$$

Therefore, two main cases can be distinguished:

- if $C < 1$, then $g_C(\lambda)$ decreases with $\mathrm{df}(\lambda)$ so that $\mathrm{df}(\widehat{\lambda}_{\min}(C))$ is huge: $\widehat{\lambda}_{\min}(C)$ overfits.
- if $C > 1$, then $g_C(\lambda)$ increases with $\mathrm{df}(\lambda)$ when $\mathrm{df}(\lambda)$ is large enough, so that $\mathrm{df}(\widehat{\lambda}_{\min}(C))$ is much smaller than when $C < 1$.

As a conclusion, $\mathrm{pen}_{\min}(\lambda)$ is the minimal amount of penalization needed so that a minimizer $\widehat{\lambda}$ of a penalized criterion is not clearly overfitting.

Following an idea first proposed in [6] and further analyzed or used in several other papers such as [21, 7, 22], we now propose to use that $\mathrm{pen}_{\min}(\lambda)$ is a minimal penalty for estimating σ^2 and plug this estimator into Eq. (5). This leads to the algorithm described in Section 4.1.

Note that the minimal penalty given by Eq. (9) is new; it generalizes previous results [6, 7] where $\mathrm{pen}_{\min}(A_\lambda) = n^{-1}\,\mathrm{tr}(A_\lambda)\sigma^2$ because all A_λ were assumed to be projection matrices, i.e., $A_\lambda^\top A_\lambda = A_\lambda$. Furthermore, our results generalize the slope heuristics $\mathrm{pen}_{\mathrm{id}} \approx 2\,\mathrm{pen}_{\min}$ (only valid for projection estimators [6, 7]) to general linear estimators for which $\mathrm{pen}_{\mathrm{id}} \,/\, \mathrm{pen}_{\min} \in (1, 2]$.

4 Main results

In this section, we first describe our algorithm and then present our theoretical results.

4.1 Algorithm

The following algorithm first computes an estimator of $\widehat{C}$ of σ^2 using the minimal penalty in Eq. (9), then considers the ideal penalty in Eq. (5) for selecting λ.

Input: Λ a finite set with $\mathrm{Card}(\Lambda) \leq Kn^\alpha$ for some $K, \alpha \geq 0$, and matrices A_λ.

- $\forall C > 0$, compute $\widehat{\lambda}_0(C) \in \arg\min_{\lambda\in\Lambda}\{\|\widehat{F}_\lambda - Y\|_2^2 + C\left(2\,\mathrm{tr}(A_\lambda) - \mathrm{tr}(A_\lambda^\top A_\lambda)\right)\}$.
- Find $\widehat{C}$ such that $\mathrm{df}(\widehat{\lambda}_0(\widehat{C})) \in \left[n^{3/4}, n/10\right]$.
- Select $\widehat{\lambda} \in \arg\min_{\lambda\in\Lambda}\{\|\widehat{F}_\lambda - Y\|_2^2 + 2\widehat{C}\,\mathrm{tr}(A_\lambda)\}$.

In the steps 1 and 2 of the above algorithm, in practice, a grid in log-scale is used, and our theoretical results from the next section suggest to use a step-size of order $n^{-1/4}$. Note that it may not be

possible in all cases to find a C such that $\mathrm{df}(\widehat{\lambda}_0(C)) \in [n^{3/4}, n/10]$; therefore, our condition in step 2, could be relaxed to finding a $\widehat{C}$ such that for all $C > \widehat{C} + \delta$, $\mathrm{df}(\widehat{\lambda}_0(C)) < n^{3/4}$ and for all $C < \widehat{C} - \delta$, $\mathrm{df}(\widehat{\lambda}_0(C)) > n/10$, with $\delta = n^{-1/4+\xi}$, where $\xi > 0$ is a small constant.

Alternatively, using the same grid in log-scale, we can select $\widehat{C}$ with maximal jump between successive values of $\mathrm{df}(\widehat{\lambda}_0(C))$—note that our theoretical result then does not entirely hold, as we show the presence of a jump around σ^2, but do not show the absence of similar jumps elsewhere.

4.2 Oracle inequality

Theorem 1 *Let $\widehat{C}$ and $\widehat{\lambda}$ be defined as in the algorithm of Section 4.1, with $\mathrm{Card}(\Lambda) \leq Kn^\alpha$ for some $K, \alpha \geq 0$. Assume that $\forall \lambda \in \Lambda$, A_λ is symmetric with $\mathrm{Sp}(A_\lambda) \subset [0,1]$, that ε_i are i.i.d. Gaussian with variance $\sigma^2 > 0$, and that $\exists \lambda_1, \lambda_2 \in \Lambda$ with*

$$\mathrm{df}(\lambda_1) \geq \frac{n}{2},\ \mathrm{df}(\lambda_2) \leq \sqrt{n},\ \textit{and}\ \ \forall i \in \{1,2\},\ n^{-1}\left\|(A_{\lambda_i} - I_n)F\right\|_2^2 \leq \sigma^2\sqrt{\frac{\ln(n)}{n}}\ . \quad (\mathbf{A_{1-2}})$$

Then, a numerical constant C_a and an event of probability at least $1 - 8Kn^{-2}$ exist on which, for every $n \geq C_a$,

$$\left(1 - 91(\alpha+2)\sqrt{\frac{\ln(n)}{n}}\right)\sigma^2 \leq \widehat{C} \leq \left(1 + \frac{44(\alpha+2)\sqrt{\ln(n)}}{n^{1/4}}\right)\sigma^2\ . \quad (10)$$

Furthermore, if

$$\exists \kappa \geq 1,\ \forall \lambda \in \Lambda, \quad n^{-1}\mathrm{tr}(A_\lambda)\sigma^2 \leq \kappa \mathbb{E}\left[n^{-1}\|\widehat{F}_\lambda - F\|_2^2\right]\ , \quad (\mathbf{A_3})$$

then, a constant C_b depending only on κ exists such that for every $n \geq C_b$, on the same event,

$$n^{-1}\|\widehat{F}_{\widehat{\lambda}} - F\|_2^2 \leq \left(1 + \frac{40\kappa}{\ln(n)}\right)\inf_{\lambda\in\Lambda}\left\{n^{-1}\|\widehat{F}_\lambda - F\|_2^2\right\} + \frac{36(\kappa+\alpha+2)\ln(n)\sigma^2}{n}\ . \quad (11)$$

Theorem 1 is proved in [20]. The proof mainly follows from the informal arguments developed in Section 3.2, completed with the following two concentration inequalities: If $\xi \in \mathbb{R}^n$ is a standard Gaussian random vector, $\alpha \in \mathbb{R}^n$ and M is a real-valued $n \times n$ matrix, then for every $x \geq 0$,

$$\mathbb{P}\left(|\langle \alpha, \xi\rangle| \leq \sqrt{2x}\,\|\alpha\|_2\right) \geq 1 - 2e^{-x} \quad (12)$$

$$\mathbb{P}\left(\forall \theta > 0,\ \left|\|M\xi\|_2^2 - \mathrm{tr}(M^\top M)\right| \leq \theta\,\mathrm{tr}(M^\top M) + 2(1+\theta^{-1})\|M\|^2 x\right) \geq 1 - 2e^{-x}\ , \quad (13)$$

where $\|M\|$ is the operator norm of M. A proof of Eq. (12) and (13) can be found in [20].

4.3 Discussion of the assumptions of Theorem 1

Gaussian noise. When ε is sub-Gaussian, Eq. (12) and Eq. (13) can be proved for $\xi = \sigma^{-1}\varepsilon$ at the price of additional technicalities, which implies that Theorem 1 is still valid.

Symmetry. The assumption that matrices A_λ must be symmetric can certainly be relaxed, since it is only used for deriving from Eq. (13) a concentration inequality for $\langle A_\lambda \xi, \xi\rangle$. Note that $\mathrm{Sp}(A_\lambda) \subset [0,1]$ barely is an assumption since it means that A_λ actually shrinks Y.

Assumptions $(\mathbf{A_{1-2}})$. $(\mathbf{A_{1-2}})$ holds if $\max_{\lambda\in\Lambda}\{\mathrm{df}(\lambda)\} \geq n/2$ and the bias is smaller than $c\,\mathrm{df}(\lambda)^{-d}$ for some $c, d > 0$, a quite classical assumption in the context of model selection. Besides, $(\mathbf{A_{1-2}})$ is much less restrictive and can even be relaxed, see [20].

Assumption $(\mathbf{A_3})$. The upper bound $(\mathbf{A_3})$ on $\mathrm{tr}(A_\lambda)$ is certainly the strongest assumption of Theorem 1, but it is only needed for Eq. (11). According to Eq. (6), $(\mathbf{A_3})$ holds with $\kappa = 1$ when A_λ is a projection matrix since $\mathrm{tr}(A_\lambda^\top A_\lambda) = \mathrm{tr}(A_\lambda)$. In the kernel ridge regression framework, $(\mathbf{A_3})$ holds as soon as the eigenvalues of the kernel matrix K decrease like $j^{-\alpha}$—see [20]. In general, $(\mathbf{A_3})$ means that $\widehat{F}_\lambda$ should not have a risk smaller than the parametric convergence rate associated with a model of dimension $\mathrm{df}(\lambda) = \mathrm{tr}(A_\lambda)$.

When $(\mathbf{A_3})$ does not hold, selecting among estimators whose risks are below the parametric rate is a rather difficult problem and it may not be possible to attain the risk of the oracle in general.

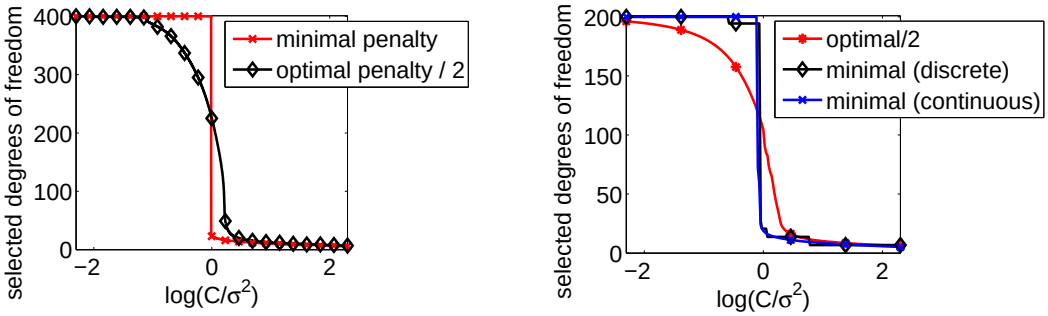

Figure 2: Selected degrees of freedom vs. penalty strength $\log(C/\sigma^2)$: note that when penalizing by the minimal penalty, there is a strong jump at $C = \sigma^2$, while when using half the optimal penalty, this is not the case. Left: single kernel case, Right: multiple kernel case.

Nevertheless, an oracle inequality can still be proved without $(\mathbf{A_3})$, at the price of enlarging $\widehat{C}$ slightly and adding a small fraction of $\sigma^2 n^{-1} \operatorname{tr}(A_\lambda)$ in the right-hand side of Eq. (11), see [20]. Enlarging $\widehat{C}$ is necessary in general: If $\operatorname{tr}(A_\lambda^\top A_\lambda) \ll \operatorname{tr}(A_\lambda)$ for most $\lambda \in \Lambda$, the minimal penalty is very close to $2\sigma^2 n^{-1} \operatorname{tr}(A_\lambda)$, so that according to Eq. (10), overfitting is likely as soon as $\widehat{C}$ underestimates σ^2, even by a very small amount.

4.4 Main consequences of Theorem 1 and comparison with previous results

Consistent estimation of σ^2. The first part of Theorem 1 shows that $\widehat{C}$ is a consistent estimator of σ^2 in a general framework and under mild assumptions. Compared to classical estimators of σ^2, such as the one usually used with Mallows' C_L, $\widehat{C}$ does not depend on the choice of some model assumed to have almost no bias, which can lead to overestimating σ^2 by an unknown amount [18].

Oracle inequality. Our algorithm satisfies an oracle inequality with high probability, as shown by Eq. (11): The risk of the selected estimator $\widehat{F}_{\widehat{\lambda}}$ is close to the risk of the oracle, up to a remainder term which is negligible when the dimensionality $\operatorname{df}(\lambda^\star)$ grows with n faster than $\ln(n)$, a typical situation when the bias is never equal to zero, for instance in kernel ridge regression.

Several oracle inequalities have been proved in the statistical literature for Mallows' C_L with a consistent estimator of σ^2, for instance in [23]. Nevertheless, except for the model selection problem (see [6] and references therein), all previous results were asymptotic, meaning that n is implicitly assumed to be larged compared to each parameter of the problem. This assumption can be problematic for several learning problems, for instance in multiple kernel learning when the number p of kernels may grow with n. On the contrary, Eq. (11) is *non-asymptotic*, meaning that it holds for every fixed n as soon as the assumptions explicitly made in Theorem 1 are satisfied.

Comparison with other procedures. According to Theorem 1 and previous theoretical results [23, 19], C_L, GCV, cross-validation and our algorithm satisfy similar oracle inequalities in various frameworks. This should not lead to the conclusion that these procedures are completely equivalent. Indeed, second-order terms can be large for a given n, while they are hidden in asymptotic results and not tightly estimated by non-asymptotic results. As showed by the simulations in Section 5, our algorithm yields statistical performances as good as existing methods, and often quite better.

Furthermore, our algorithm never overfits too much because $\operatorname{df}(\widehat{\lambda})$ is by construction smaller than the effective dimensionality of $\widehat{\lambda}_0(\widehat{C})$ at which the jump occurs. This is a quite interesting property compared for instance to GCV, which is likely to overfit if it is not corrected because GCV minimizes a criterion proportional to the empirical risk.

5 Simulations

Throughout this section, we consider exponential kernels on $\mathbb{R}^d$, $k(x,y) = \prod_{i=1}^d e^{-|x_i - y_i|}$, with the x's sampled i.i.d. from a standard multivariate Gaussian. The functions f are then selected randomly as $\sum_{i=1}^m \alpha_i k(\cdot, z_i)$, where both α and z are i.i.d. standard Gaussian (i.e., f belongs to the RKHS).

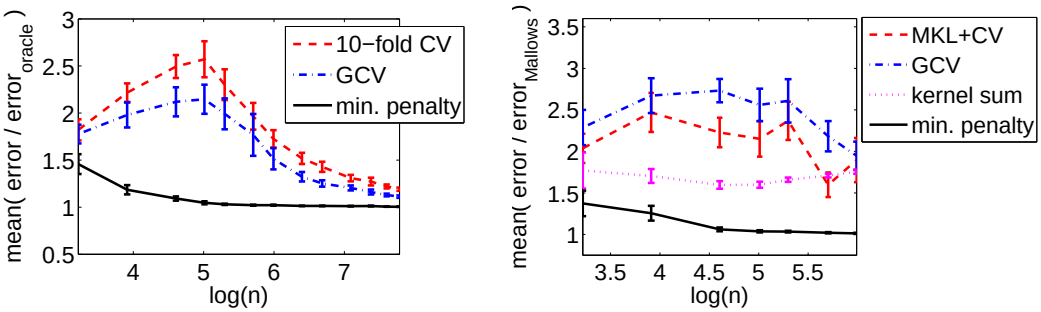

Figure 3: Comparison of various smoothing parameter selection (minikernel, GCV, 10-fold cross validation) for various values of numbers of observations, averaged over 20 replications. Left: single kernel, right: multiple kernels.

Jump. In Figure 2 (left), we consider data $x_i \in \mathbb{R}^6$, $n = 1000$, and study the size of the jump in Figure 2 for kernel ridge regression. With half the optimal penalty (which is used in traditional variable selection for linear regression), we do not get any jump, while with the minimal penalty we always do. In Figure 2 (right), we plot the same curves for the multiple kernel learning problem with two kernels on two different 4-dimensional variables, with similar results. In addition, we show two ways of optimizing over $\lambda \in \Lambda = \mathbb{R}_+^2$, by discrete optimization with n different kernel matrices—a situation covered by Theorem 1—or with continuous optimization with respect to η in Eq. (1), by gradient descent—a situation not covered by Theorem 1.

Comparison of estimator selection methods. In Figure 3, we plot model selection results for 20 replications of data ($d = 4$, $n = 500$), comparing GCV [8], our minimal penalty algorithm, and cross-validation methods. In the left part (single kernel), we compare to the oracle (which can be computed because we can enumerate Λ), and use for cross-validation all possible values of λ. In the right part (multiple kernel), we compare to the performance of Mallows' C_L when σ^2 is known (i.e., penalty in Eq. (5)), and since we cannot enumerate all λ's, we use the solution obtained by MKL with CV [5]. We also compare to using our minimal penalty algorithm with the sum of kernels.

6 Conclusion

A new light on the slope heuristics. Theorem 1 generalizes some results first proved in [6] where all A_λ are assumed to be projection matrices, a framework where assumption $(\mathbf{A_3})$ is automatically satisfied. To this extent, Birgé and Massart's slope heuristics has been modified in a way that sheds a new light on the "magical" factor 2 between the minimal and the optimal penalty, as proved in [6, 7]. Indeed, Theorem 1 shows that for general linear estimators,

$$\frac{\mathrm{pen}_{\mathrm{id}}(\lambda)}{\mathrm{pen}_{\min}(\lambda)} = \frac{2\operatorname{tr}(A_\lambda)}{2\operatorname{tr}(A_\lambda) - \operatorname{tr}(A_\lambda^\top A_\lambda)} \ , \tag{14}$$

which can take any value in $(1, 2]$ in general; this ratio is only equal to 2 when $\operatorname{tr}(A_\lambda) \approx \operatorname{tr}(A_\lambda^\top A_\lambda)$, hence mostly when A_λ is a projection matrix.

Future directions. In the case of projection estimators, the slope heuristics still holds when the design is random and data are heteroscedastic [7]; we would like to know whether Eq. (14) is still valid for heteroscedastic data with general linear estimators. In addition, the good empirical performances of elbow heuristics based algorithms (i.e., based on the sharp variation of a certain quantity around good hyperparameter values) suggest that Theorem 1 can be generalized to many learning frameworks (and potentially to non-linear estimators), probably with small modifications in the algorithm, but always relying on the concept of minimal penalty.

Another interesting open problem would be to extend the results of Section 4, where $\operatorname{Card}(\Lambda) \leq Kn^\alpha$ is assumed, to continuous sets Λ such as the ones appearing naturally in kernel ridge regression and multiple kernel learning. We conjecture that Theorem 1 is valid without modification for a "small" continuous Λ, such as in kernel ridge regression where taking a grid of size n in log-scale is almost equivalent to taking $\Lambda = \mathbb{R}_+$. On the contrary, in applications such as the Lasso with $p \gg n$ variables, the natural set Λ cannot be well covered by a grid of cardinality n^α with α small, and our minimal penalty algorithm and Theorem 1 certainly have to be modified.

References

[1] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[2] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2001.

[3] O. Chapelle and V. Vapnik. Model selection for support vector machines. In *Advances in Neural Information Processing Systems (NIPS)*, 1999.

[4] C. E. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[5] F. Bach. Consistency of the group Lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

[6] L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73, 2007.

[7] S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.*, 10:245–279, 2009.

[8] P. Craven and G. Wahba. Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, 31(4):377–403, 1978/79.

[9] G. Wahba. *Spline Models for Observational Data*. SIAM, 1990.

[10] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of The Royal Statistical Society Series B*, 68(1):49–67, 2006.

[11] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *J. Mach. Learn. Res.*, 5:27–72, 2003/04.

[12] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of The Royal Statistical Society Series B*, 58(1):267–288, 1996.

[13] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer-Verlag, 2001.

[14] D. M. Allen. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16:125–127, 1974.

[15] M. Stone. Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B*, 36:111–147, 1974.

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

[17] C. L. Mallows. Some comments on C_p. *Technometrics*, 15:661–675, 1973.

[18] B. Efron. How biased is the apparent error rate of a prediction rule? *J. Amer. Statist. Assoc.*, 81(394):461–470, 1986.

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

[20] S. Arlot and F. Bach. Data-driven calibration of linear estimators with minimal penalties, September 2009. Long version. arXiv:0909.1884v1.

[21] É. Lebarbier. Detecting multiple change-points in the mean of a gaussian process by model selection. *Signal Proces.*, 85:717–736, 2005.

[22] C. Maugis and B. Michel. Slope heuristics for variable selection and clustering via gaussian mixtures. Technical Report 6550, INRIA, 2008.

[23] K.-C. Li. Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. *Ann. Statist.*, 15(3):958–975, 1987.