Minimal penalties and the slope heuristics: a survey

Sylvain Arlot

Journal de la Société Française de Statistique

Minimal penalties and the slope heuristics: a survey

Titre: Pénalités minimales et heuristique de pente

Sylvain Arlot[1]

Abstract: Birgé and Massart proposed in 2001 the slope heuristics as a way to choose optimally from data an unknown multiplicative constant in front of a penalty. It is built upon the notion of minimal penalty, and it has been generalized since to some "minimal-penalty algorithms". This paper reviews the theoretical results obtained for such algorithms, with a self-contained proof in the simplest framework, precise proof ideas for further generalizations, and a few new results. Explicit connections are made with residual-variance estimators —with an original contribution on this topic, showing that for this task the slope heuristics performs almost as well as a residual-based estimator with the best model choice— and some classical algorithms such as L-curve or elbow heuristics, Mallows' C_p, and Akaike's FPE. Practical issues are also addressed, including two new practical definitions of minimal-penalty algorithms that are compared on synthetic data to previously-proposed definitions. Finally, several conjectures and open problems are suggested as future research directions.

Résumé : Birgé et Massart ont proposé en 2001 l'heuristique de pente, pour déterminer à l'aide des données une constante multiplicative optimale devant une pénalité en sélection de modèles. Cette heuristique s'appuie sur la notion de pénalité minimale, et elle a depuis été généralisée en "algorithmes à base de pénalités minimales". Cet article passe en revue les résultats théoriques obtenus sur ces algorithmes, avec une preuve complète dans le cadre le plus simple, des idées de preuves précises pour généraliser ce résultat au-delà des cadres déjà étudiés, et quelques résultats nouveaux. Des liens sont faits avec les méthodes d'estimation de la variance résiduelle (avec une contribution originale sur ce thème, qui démontre que l'heuristique de pente produit un estimateur de la variance quasiment aussi bon qu'un estimateur fondé sur les résidus d'un modèle oracle) ainsi qu'avec plusieurs algorithmes classiques tels que les heuristiques de coude (ou de courbe en L), C_p de Mallows et FPE d'Akaike. Les questions de mise en œuvre pratique sont également étudiées, avec notamment la proposition de deux nouvelles définitions pratiques pour des algorithmes à base de pénalités minimales et leur comparaison aux définitions précédentes sur des données simulées. Enfin, des conjectures et problèmes ouverts sont proposés comme pistes de recherche pour l'avenir.

Keywords: model selection, estimator selection, penalization, slope heuristics, minimal penalty, residual-variance estimation, L-curve heuristics, elbow heuristics, scree test, overpenalization
Mots-clés : sélection de modèles, sélection d'estimateurs, pénalisation, heuristique de pente, pénalité minimale, estimation de la variance résiduelle, heuristique de courbe en L, heuristique de coude, test scree, surpénalisation
AMS 2000 subject classifications: 62-02, 62G05, 62J05

Contents

[1] Laboratoire de Mathématiques d'Orsay
Univ. Paris-Sud, CNRS, Université Paris-Saclay
91405 Orsay, France
E-mail: sylvain.arlot@u-psud.fr

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

1. Introduction

Model selection attracts much attention in statistics since more than forty years [Akaike, 1973, Mallows, 1973, Burnham and Anderson, 2002, Massart, 2007]. A related and crucial question for machine learning is the data-driven choice of hyperparameters of learning algorithms. Both are particular instances of the estimator-selection problem: given a family of estimators, how to choose from data one among them whose risk is as small as possible?

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

One of the main strategies proposed for estimator (or model) selection is penalization, that is, choosing the estimator minimizing the sum of its empirical risk —how well it fits the data— and some penalty term —whose role is to avoid overfitting. Optimal penalties often depend on at least one parameter whose data-driven choice is challenging. In the early 2000s, Birgé and Massart [Birgé and Massart, 2001, Birgé and Massart, 2007] pointed out two key facts leading to a novel approach for an optimal data-driven choice of multiplicative constants in front of penalties. Birgé and Massart were considering a rather theoretical question: what is the *minimal* amount of penalization needed for avoiding a strong overfitting? For least-squares estimators in regression, they noticed that (i) the minimal penalty is equal to half the optimal penalty, and (ii) the minimal penalty is observable. These two facts are called "the slope heuristics" and lead to an algorithm for choosing multiplicative constants in front of penalties.

These ideas and the corresponding algorithm have been generalized since to several frameworks (see Section 3–4 and 8), with numerous applications in various fields such as biology [Akakpo, 2011, Reynaud-Bouret and Schbath, 2010, Rau et al., 2015, Devijver and Gallopin, 2018, Devijver et al., 2017, Bontemps and Toussile, 2013], energy [Devijver et al., 2015, Michel, 2008], or text analysis [Derman and Le Pennec, 2017] (see also Section 8).

In particular, for linear estimators in regression, the original slope heuristics does not work directly and can be modified successfully into a more general "minimal-penalty algorithm" [Arlot and Bach, 2009, Arlot and Bach, 2011] detailed in Section 3.

For least-squares regression with projection or linear estimators, the slope heuristics also provides a residual-variance estimator with nice properties (Section 6.1). In the general setting, the slope heuristics can also be seen as a way to give proper mathematical grounds to "L-curve" or "elbow heuristics" algorithms that are used for choosing regularization parameters in ill-posed problems [Hansen and O'Leary, 1993], as explained in Sections 6.4–6.5.

Goals The goals of this survey are the following:

1. to review recent theoretical results about the slope heuristics, and more generally all minimal-penalty algorithms (Sections 2–4);
2. to help identifying how such results could be generalized to other settings, possibly with new algorithms, by giving a precise account of existing proofs (Sections 2.7, 4.1, and 5) and by identifying several conjectures and open problems suggested by experimental results (Section 8);
3. to make connections between minimal penalties and other classical procedures for residual-variance estimation and for model or estimator selection (Section 6).

Practical issues are only briefly mentioned in Section 7, since more details can be found on these in the survey [Baudry et al., 2012].

There is currently no final answer to the question of generalizing minimal-penalty algorithms as much as possible, but we hope that this survey will motivate further theoretical and empirical work in this direction, which could have a great practical impact in statistics, machine learning, and data science in general.

Contributions Let us finally point out some original results appearing in this paper. In the framework of least-squares fixed-design regression with projection estimators and Gaussian

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

noise, Theorem 1 validates the slope heuristics in a stronger sense compared to previous results [Birgé and Massart, 2007]; it is inspired by [Arlot and Bach, 2011] but makes weaker assumptions. Its extension to sub-Gaussian noise (Remark 1 in Section 2.5) is original. As a corollary, Proposition 3 in Section 6.1 is the first precise statement on a slope-heuristics-based residual-variance estimator —more precise than the result that can be derived from [Arlot and Bach, 2011]—, showing that it is minimax optimal (up to $\log(n)$ factors) under mild assumptions. Proposition 3 provides non-asymptotic bounds (in expectation and with high probability) on this residual-variance estimator, that can be seen as some kind of oracle inequality for residual-variance estimation, which is interesting independently from the slope heuristics.

In the general framework, Propositions 1–2 in Section 5.2 propose two general approaches for justifying minimal-penalty algorithms. These approaches were previously proposed in specific settings [Lerasle and Takahashi, 2016, Garivier and Lerasle, 2011], but their generalization to the setting of Section 3.1 is new. For instance, the application of Proposition 1 to general minimum-contrast estimators with a bounded contrast is new, to the best of our knowledge.

On the practical side, as a complement to [Baudry et al., 2012], Section 7 shows original numerical experiments on synthetic data, assessing the performance of the slope heuristics in the least-squares regression framework, for both residual-variance estimation and model selection. Two new practical definitions of the slope heuristics (called 'median' and 'consensus') are proposed and compared to the classical ones. An efficient implementation of one previously-proposed definition is also provided and proved (Algorithm 8 and Proposition 6 in Appendix B.2).

2. The slope heuristics

This section presents the original "slope heuristics" [Birgé and Massart, 2001, Birgé and Massart, 2001, Birgé and Massart, 2007] in the framework of fixed-design regression, with the least-squares risk and projection estimators. By focusing on this framework, we get most of the flavor of the slope heuristics while keeping the exposition simple.

2.1. Framework

The framework considered in Section 2 is the following. We observe

$$Y = F + \varepsilon \in \mathbb{R}^n \tag{1}$$

where $\varepsilon_1, \dots, \varepsilon_n$ are independent and identically distributed with mean 0 and variance σ^2, and $F \in \mathbb{R}^n$ is some (deterministic) signal of interest. For instance, F can be equal to $(f(x_i))_{1 \leqslant i \leqslant n}$ for some deterministic design points $x_1, \dots, x_n \in \mathcal{X}$ and f some unknown measurable function $\mathcal{X} \mapsto \mathbb{R}$, with no assumption on the set $\mathcal{X}$.

The goal is to reconstruct F from Y, that is, to find some $t \in \mathbb{R}^n$ such that its quadratic risk

$$\frac{1}{n} \|t - F\|^2$$

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

is small, where for every $u \in \mathbb{R}^n$, $\|u\|^2 = \sum_{i=1}^n u_i^2$. To this end, for every linear subspace S of $\mathbb{R}^n$, the *projection estimator* or *least-squares estimator* on S is defined as

$$\widehat{F}_S \in \operatorname*{argmin}_{t \in S} \left\{ \frac{1}{n} \|t - Y\|^2 \right\}$$

where $n^{-1}\|t-Y\|^2$ is called the empirical risk of t. Since S is a linear subspace, $\widehat{F}_S$ exists and is unique: $\widehat{F}_S = \Pi_S Y$ where $\Pi_S : \mathbb{R}^n \to \mathbb{R}^n$ denotes the orthogonal projection onto S. In the following, any linear subspace S of $\mathbb{R}^n$ is called a *model*.

Let $(S_m)_{m \in \mathcal{M}}$ be some collection of models, and for every $m \in \mathcal{M}$, let

$$\widehat{F}_m = \widehat{F}_{S_m} = \Pi_{S_m} Y \quad \text{and} \quad \Pi_m = \Pi_{S_m} .$$

In this survey, we assume that the goal of model selection is to choose from data some $\widehat{m} \in \mathcal{M}$ such that the quadratic risk of $\widehat{F}_{\widehat{m}}$ is minimal. The best choice would be the *oracle*:

$$m^\star \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 \right\} ,$$

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

$$\frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 \leqslant K_n \inf_{m \in \mathcal{M}} \left\{ \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 \right\} + R_n \tag{2}$$

with large probability, where the leading constant K_n should be close to 1 —at least for large n— and the remainder term R_n should be small compared to the oracle risk $\inf_{m \in \mathcal{M}} \{ n^{-1} \|\widehat{F}_m - F\|^2 \}$.

2.2. *Optimal penalty*

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

$$\forall m \in \mathcal{M}, \qquad \mathbb{E}\big[\mathrm{crit}(m)\big] \approx \mathbb{E}\left[\frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 \right] ,$$

then $\widehat{m}$ satisfies with large probability an oracle inequality such as Eq. (2) with an optimal constant $K_n = 1 + \mathrm{o}(1)$. This can be proved by showing a concentration inequality for $\|\Pi_m \varepsilon\|^2$ and $\langle \varepsilon, (I_n - \Pi_m) F \rangle$ around their expectations for all $m \in \mathcal{M}$, see Section 2.7. For instance, cross-validation [Allen, 1974, Stone, 1974] and generalized cross-validation (GCV) [Craven and Wahba, 1978] 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, that is, using a criterion of the form:

$$\mathrm{crit}(m) = \frac{1}{n} \left\| \widehat{F}_m - Y \right\|^2 + \mathrm{pen}(m) . \tag{3}$$

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

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

$$\mathrm{pen}_{\mathrm{opt},0}(m) := \mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] - \mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2\right]. \tag{4}$$

When $\widehat{F}_m = \Pi_m Y$, we have

$$\left\|\widehat{F}_m - F\right\|^2 = \left\|(I_n - \Pi_m)F\right\|^2 + \|\Pi_m \varepsilon\|^2 \tag{5}$$

and

$$\left\|\widehat{F}_m - Y\right\|^2 = \left\|\widehat{F}_m - F\right\|^2 + \|\varepsilon\|^2 - 2\langle \varepsilon, \Pi_m \varepsilon\rangle + 2\left\langle \varepsilon, (I_n - \Pi_m)F\right\rangle, \tag{6}$$

where $\forall t, u \in \mathbb{R}^n$, $\langle t, u\rangle = \sum_{i=1}^n t_i u_i$. Since the ε_i are independent, centered, with variance σ^2, Eq. (5) and Eq. (6) imply that

$$\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] = \frac{1}{n}\left\|(I_n - \Pi_m)F\right\|^2 + \frac{\sigma^2 D_m}{n}, \tag{7}$$

$$\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2\right] = \frac{1}{n}\left\|(I_n - \Pi_m)F\right\|^2 + \frac{\sigma^2 (n - D_m)}{n}, \tag{8}$$

and

$$\mathrm{pen}_{\mathrm{opt},0}(m) + \sigma^2 = \frac{2\sigma^2 D_m}{n} =: \mathrm{pen}_{\mathrm{opt}}(m), \tag{9}$$

where $D_m := \dim(S_m)$. Note that the optimal penalties (9) and (4) differ by an additive constant σ^2, which does not change the argmin of the penalized criterion (3); this choice simplifies formulas involving $\mathrm{pen}_{\mathrm{opt}}$.

Eq. (7) is classically known as a *bias-variance decomposition of the risk*: the first term — called *approximation error* or bias— decreases when S_m gets larger, while the second term — called *estimation error* or variance— increases when S_m gets larger, see Figure 1 left. Eq. (8) shows that the expectation of the empirical risk decreases when S_m gets larger, as expected since $\widehat{F}_m$ is defined as a minimizer of the empirical risk, see Figure 1 left.

The expression of the optimal penalty in Eq. (9) leads to Mallows' C_p [Mallows, 1973], where σ^2 is replaced by some estimator $\widehat{\sigma^2}$. Several approaches exist for estimating σ^2, see Section 6.1. The slope heuristics provides a data-driven estimation of the unknown constant σ^2 in front of the penalty shape D_m/n thanks to the notion of minimal penalty.

2.3. *Minimal penalty and the slope heuristics*

Eq. (9) shows that the shape $\mathrm{pen}_1(m) = D_m/n$ of the optimal penalty is known, even when σ^2 is unknown. A natural question is to determine the minimal value of the constant that should be put in front of $\mathrm{pen}_1(m)$. More precisely, if for every $C \geqslant 0$

$$\widehat{m}(C) \in \operatorname*{argmin}_{m \in \mathscr{M}} \left\{ \frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2 + C\frac{D_m}{n} \right\}, \tag{10}$$

what is the minimal value of C such that $\widehat{m}(C)$ stays a "reasonable" choice, that is, avoids strong overfitting, or equivalently, satisfies an oracle inequality like Eq. (2) with $K_n = \mathscr{O}(1)$ as n tends to infinity?

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

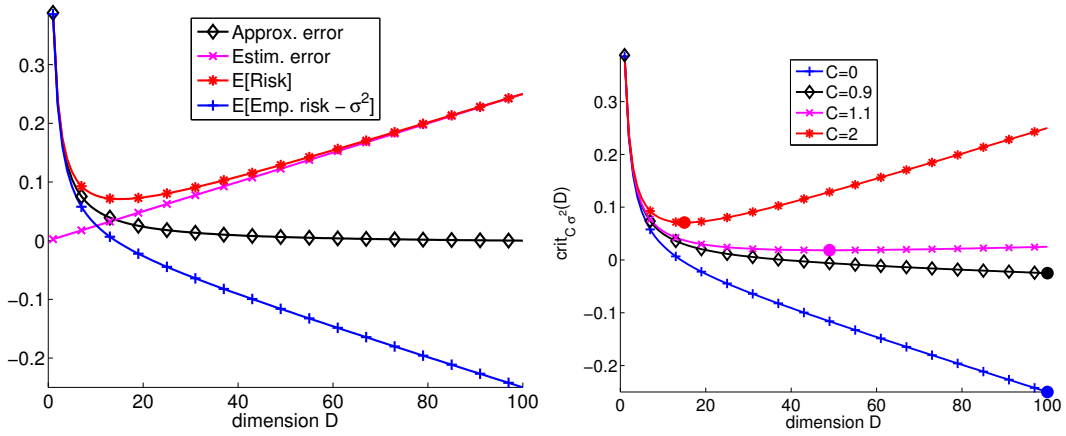


FIGURE 1. *Left*: *Expectations of the risk and empirical risk, bias-variance decomposition of the risk. Right*: $\mathrm{crit}_{C\sigma^2}(m)$, *defined by Eq.* (11), *for* $C \in \{0, 0.9, 1.1, 2\}$; *its minimal value at* $m^\star(C\sigma^2)$ *is shown by a plain dot. 'Easy setting', see Appendix D for detailed information.*

In order to understand how $\widehat{m}(C)$ behaves as a function of C, let us consider, for every $C \geqslant 0$,

$$m^\star(C) \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \mathbb{E}\left[\frac{1}{n} \left\| \widehat{F}_m - Y \right\|^2 + C \frac{D_m}{n} \right] \right\} = \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \mathrm{crit}_C(m) \right\}$$
$$\text{with} \quad \mathrm{crit}_C(m) := \frac{1}{n} \left[\left\| (I_n - \Pi_m) F \right\|^2 + (C - \sigma^2) D_m \right], \tag{11}$$

by Eq. (8). Provided that we can prove some uniform concentration inequalities for $\|\widehat{F}_m - Y\|^2$, we can expect $m^\star(C)$ to be close to $\widehat{m}(C)$. Let us also assume, for simplicity, that the approximation error $n^{-1}\|(I_n - \Pi_m)F\|^2$ is a decreasing function of D_m —which holds for instance if the S_m are nested— and is almost constant for D_m large enough. Then, two cases can be distinguished with respect to C:

— if $C < \sigma^2$, then $\mathrm{crit}_C(m)$ is a decreasing function of D_m, and $D_{m^\star(C)}$ is huge: $m^\star(C)$ overfits.
— if $C > \sigma^2$, then $\mathrm{crit}_C(m)$ increases with D_m for D_m large enough, so $D_{m^\star(C)}$ is much smaller than when $C < \sigma^2$.

This behavior is illustrated on the right part of Figure 1. In other words,

$$\mathrm{pen}_{\min}(m) := \frac{\sigma^2 D_m}{n} \tag{12}$$

seems to be the minimal amount of penalization needed so that a minimizer $\widehat{m}$ of the penalized criterion (3) does not clearly overfit. The above arguments will be made rigorous in Section 2.5, showing that $\sigma^2 D_m / n$ is indeed a minimal penalty in the current framework.

We can now summarize the slope heuristics into two major facts. First, from Eq. (9) and (12), we get a *relationship between the optimal and minimal penalties*:

$$\mathrm{pen}_{\mathrm{opt}}(m) = 2\,\mathrm{pen}_{\min}(m)\,. \tag{13}$$

Second, *the minimal penalty is observable*, since $D_{\widehat{m}(C)}$ decreases "smoothly" as a function of C everywhere except around $C = \sigma^2$ where it jumps.

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

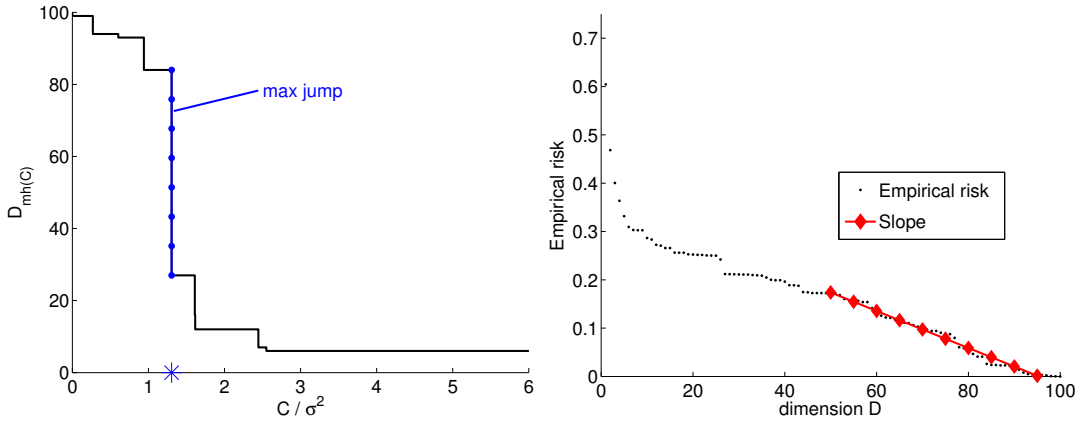


FIGURE 2. *Illustration of Algorithms 1 and 2 on the same sample ('easy' setting, see Appendix D for details). Left: Plot of* $C \mapsto D_{\widehat{m}(C)}$ *and visualization of* $\widehat{C}_{\text{jump}}$. *Right: Plot of* $D_m \mapsto n^{-1}\|Y - \widehat{F}_m\|^2$ *and visualization of* $-\widehat{C}_{\text{slope}}/n$.

2.4. *Data-driven penalty algorithm*

The two major facts of the slope heuristics described above directly lead to a data-driven penalization algorithm, which can be formalized in two ways.

2.4.1. *Dimension jump*

First, we can estimate the minimal penalty by looking for a jump of $C \mapsto D_{\widehat{m}(C)}$, and make use of Eq. (13) to get an estimator of the optimal penalty.

Algorithm 1. `Input:` $\left(\|\widehat{F}_m - Y\|^2\right)_{m\in\mathscr{M}}$.

1. *Compute* $(\widehat{m}(C))_{C\geqslant 0}$, *where* $\widehat{m}(C)$ *is defined by Eq.* (10).
2. *Find* $\widehat{C}_{\text{jump}} > 0$ *corresponding to the "unique large jump" of* $C \mapsto D_{\widehat{m}(C)}$.
3. *Select* $\widehat{m}_{\text{Alg.1}} \in \operatorname{argmin}_{m\in\mathscr{M}}\left\{n^{-1}\|\widehat{F}_m - Y\|^2 + 2\widehat{C}_{\text{jump}} D_m/n\right\}$.

`Output:` $\widehat{m}_{\text{Alg.1}}$.

The left part of Figure 2 shows one instance of the plot of $C \mapsto D_{\widehat{m}(C)}$, with one clear jump corresponding to $\widehat{C}_{\text{jump}}$. Computational issues are discussed in Section 7.2; in particular, step 1 of Algorithm 1 can be done efficiently, see Appendix B.1. Step 2 of Algorithm 1 can be done in several ways, see Section 7.1. The practical problems arising with step 2 of Algorithm 1 can motivate the use of an alternative algorithm that we detail below.

2.4.2. *Slope estimation*

As explained in Section 2.3, the reason why $D_{\widehat{m}(C)}$ jumps around $C \approx \sigma^2$ is that by Eq. (8),

$$\frac{1}{n}\mathbb{E}\left[\left\|\widehat{F}_m - Y\right\|^2\right] = \frac{a(m) - \sigma^2 D_m}{n}$$

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

where $a(m) := \|(I_n - \Pi_m)F\|^2 + n\sigma^2$ can be assumed almost constant for all m such that D_m is large enough. Therefore, considering only models with a large dimension, the empirical risk approximately has a linear behavior as a function of D_m, with slope $-\sigma^2/n$. Since the empirical risk is observable, one can estimate this slope in order to get an estimator of σ^2, and plug it in the optimal penalty given by Eq. (9).

Algorithm 2. `Input`: $(\|\widehat{F}_m - Y\|^2)_{m\in\mathcal{M}}$.

1. *Estimate the slope* $\widehat{S}$ *of* $\|\widehat{F}_m - Y\|^2$ *as a function of* D_m *for all* $m \in \mathcal{M}$ *with* D_m *"large enough", for instance by (robust) linear regression, and define* $\widehat{C}_{\text{slope}} = -n\widehat{S}$.
2. *Select* $\widehat{m}_{\text{Alg.2}} \in \operatorname{argmin}_{m\in\mathcal{M}}\{n^{-1}\|\widehat{F}_m - Y\|^2 + 2\widehat{C}_{\text{slope}} D_m/n\}$.

`Output`: $\widehat{m}_{\text{Alg.2}}$.

The right part of Figure 2 shows an instance of the plot of $n^{-1}\|\widehat{F}_m - Y\|^2$ as a function of D_m. Algorithm 2 relies on the choice of what is a "large enough" dimension, and how the slope $\widehat{C}_{\text{slope}}$ is estimated. Therefore, Algorithms 1 and 2 both have pros and cons, and there is no universal choice between them. The links between Algorithms 1 and 2, as well as their differences, are discussed in Section 7.1.

2.5. *What can be proved mathematically*

A major interest of the slope heuristics is that it can be made rigorous. For instance, in the framework of the present section, we can prove the next theorem.

Theorem 1. *In the framework described in Section 2.1, assume that* $\mathcal{M}$ *is finite, contains at least one model of dimension at most* $n/20$*, and that*

$$\exists m_1 \in \mathcal{M}\,, \quad S_{m_1} = \mathbb{R}^n \tag{HId}$$

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)\,. \tag{HG}$$

Recall that for every $C \geqslant 0$, $\widehat{m}(C)$ *is defined by Eq. (10). Then, for every* $\gamma \geqslant 0$*, some* $n_0(\gamma)$ *exists such that if* $n \geqslant n_0(\gamma)$*, with probability at least* $1 - 4\operatorname{card}(\mathcal{M})n^{-\gamma}$*, the following inequalities hold simultaneously*:

$$\forall C \leqslant (1-\eta_n^-)\,\sigma^2\,, \qquad D_{\widehat{m}(C)} \geqslant \frac{9n}{10}\,, \tag{14}$$

$$\forall C \leqslant (1-\eta_n^-)\,\sigma^2\,, \qquad \frac{1}{n}\left\|\widehat{F}_{\widehat{m}(C)} - F\right\|^2 \geqslant \frac{7\sigma^2}{8}\,, \tag{15}$$

$$\forall C \geqslant (1+\eta_n^+)\,\sigma^2\,, \qquad D_{\widehat{m}(C)} \leqslant \frac{n}{10}\,, \tag{16}$$

$$\forall C > \sigma^2\,, \qquad \frac{1}{n}\left\|\widehat{F}_{\widehat{m}(C)} - F\right\|^2 \leqslant h\left(\frac{C}{\sigma^2}\right)\left[\inf_{m\in\mathcal{M}}\left\{\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right\} + \frac{20\sigma^2 x}{n}\right], \tag{17}$$

and for every $\eta \in (0,1/2]$ *and* $C \in \left[(2-\eta)\sigma^2, (2+\eta)\sigma^2\right]$,

$$\frac{1}{n}\left\|\widehat{F}_{\widehat{m}(C)} - F\right\|^2 \leqslant (1+3\eta) \inf_{m\in\mathscr{M}} \left\{\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right\} + \frac{880\sigma^2\gamma\log(n)}{\eta n}, \tag{18}$$

where $\sigma^2\eta_n^+ = 40 \inf_{m\in\mathscr{M}\,/\,D_m\leqslant n/20}\left\{\frac{1}{n}\left\|(I_n-\Pi_m)F\right\|^2\right\} + 82\sigma^2\sqrt{\frac{\gamma\log(n)}{n}}$,

$\eta_n^- = 41\sqrt{\frac{\gamma\log(n)}{n}}$, *and* $\forall u>1\,,\, h(u) = \frac{10}{(u-1)^4}\mathbb{1}_{u\in(1,2)} + u^3\mathbb{1}_{u\geqslant 2}$.

Theorem 1 is proved in Section 2.7. It revisits results first obtained by [Birgé and Massart, 2007], formulating them similarly to [Arlot and Bach, 2011] but with milder assumptions.

What Theorem 1 proves about Algorithms 1– 2 Eq. (14) and (16) do not show exactly that there is a single large jump in $C \mapsto D_{\widehat{m}(C)}$, as in the heuristics reasoning of Section 2.3. We cannot hope to prove it since numerical experiments show that the global jump of $D_{\widehat{m}(C)}$ can be split into several small jumps within a small interval of values of C, see Figure 5 in Section 7.1. Nevertheless, Eq. (14) and (16) imply that the variation of $D_{\widehat{m}(C)}$ over a geometric window of C is extremely strong around σ^2: if $\widehat{C}_{\text{jump}}$ in Algorithm 1 is defined as

$$\widehat{C}_{\text{window}} = \widehat{C}_{\text{window}}(\eta) \in \operatorname*{argmax}_{C>0}\left\{D_{\widehat{m}(C/(1+\eta))} - D_{\widehat{m}(C(1+\eta))}\right\} \tag{19}$$

with $\eta = \max\{\eta_n^-, \eta_n^+\}$, then $\widehat{C}_{\text{window}}$ is close to σ^2 —see Proposition 3 in Section 6.1 for a precise statement—, and Eq. (18) implies a first-order optimal oracle inequality for the model-selection procedure of Algorithm 1. Note that $\widehat{C}_{\text{window}}$ can be computed efficiently, see Section 7.2 and Appendix B.2. In addition, Eq. (14) and (16) imply that

$$\widehat{C}_{\text{thr.}} = \widehat{C}_{\text{thr.}}(T_n) := \inf\left\{C \geqslant 0\,/\,D_{\widehat{m}(C)} \leqslant T_n\right\} \tag{20}$$

is close to σ^2 when $T_n \in [n/10, 9n/10]$ —precise statements are provided by Proposition 3 in Section 6.1—, and Eq. (18) implies a first-order optimal oracle inequality for the corresponding model-selection procedure. See Section 7.1 for practical comments about these variants of Algorithm 1.

Theorem 1 does not prove that Algorithm 2 works, and it seems difficult to prove such a result without adding some assumptions. Indeed, the key heuristics behind Algorithm 2 is a linear behavior of the empirical risk as a function of the dimension, at least for large models. In the proof of Theorem 1, we control the deviations of the empirical risk around its expectation, but this is not sufficient for justifying Algorithm 2 without a strong uniform control on the approximation errors of the models, an assumption much stronger than the ones of Theorem 1.

Note finally that Eq. (15) and (17) are not necessary for justifying Algorithm 1, but they are interesting for theory since they justify the term "minimal penalty". Eq. (15) is a straightforward consequence of Eq. (14), and results like Eq. (17) are easier to obtain than Eq. (18), see Section 4.4.

Variant of Theorem 1 If $\mathscr{M}$ contains at least one model of dimension at most $c_n \in [0,n)$, on the event defined in Theorem 1, we can actually prove that more results hold true: we can change Eq. (14) and (16) respectively into

$$\forall a_n < n\,, \quad \forall C \leqslant \left[1-\eta_n^-(a_n)\right]\sigma^2\,, \quad D_{\widehat{m}(C)} \geqslant a_n \tag{21}$$

$$\forall b_n > c_n\,, \quad \forall C \geqslant \left[1+\eta_n^+(b_n,c_n)\right]\sigma^2\,, \quad D_{\widehat{m}(C)} \leqslant b_n \tag{22}$$

where

$$\sigma^2\eta_n^+(b_n,c_n) := \frac{n}{b_n-c_n}\left(2\mathscr{B}(c_n)+4.1\sigma^2\sqrt{\frac{\gamma\log(n)}{n}}\right),$$

$$\mathscr{B}(c_n) := \inf_{m\in\mathscr{M}\,/\,D_m\leqslant c_n}\left\{\frac{1}{n}\left\|(I_n-\Pi_m)F\right\|^2\right\}, \quad \text{and} \quad \eta_n^-(a_n) := 4.1\left(1-\frac{a_n}{n}\right)^{-1}\sqrt{\frac{\gamma\log(n)}{n}}\,.$$

In particular, under the assumptions of Theorem 1, taking $a_n \in (9n/10,n)$, $b_n \in (n/20,n/10)$ and $c_n = n/20$, we get a larger jump of $D_{\widehat{m}(C)}$ —hence easier to detect— by considering a larger window of values of C, hence reducing the precision of the estimation of σ^2.

Relaxation of the noise assumption Assumption (**HG**) is a classical noise model for proving non-asymptotic oracle inequalities. In Theorem 1, it is only used for proving some concentration inequalities at the beginning of the proof —Eq. (23)–(24) in Section 2.7—, so it could be changed into any noise assumption ensuring that similar concentration inequalities hold true. For instance, Theorem 1 can be generalized to the case of sub-Gaussian noise, as formalized below.

Remark 1 (Generalization of Theorem 1 to sub-Gaussian noise)**.** *Assume that the* $(\varepsilon_i)_{1\leqslant i\leqslant n}$ *are centered, independent, and* $(\phi^2\sigma^2)$*-sub-Gaussian for some* $\phi > 0$ *—with any definition among the classical ones since they are all equivalent up to numerical constants [Boucheron et al., 2013, Section 2.3]. Then, by the Cramér-Chernoff method [Boucheron et al., 2013, Section 2.2], Eq. (24) holds true with probability at least* $1-2\exp(x/\phi^2)$*. In addition, [Bellec, 2014, Theorem 3.4] shows that Eq. (23) holds true with probability at least* $1-2\exp[x/(L\phi^2)]$ *for some numerical constant L. Therefore, the event* $\Omega_{L\phi^2 x}$ *defined in the proof of Theorem 1 has a probability at least* $1-4\,\mathrm{card}(\mathscr{M})\mathrm{e}^{-x}$*. So, the result of Theorem 1 holds true with x (resp.* γ*) replaced by* $L\phi^2 x$ *(resp.* $L\phi^2\gamma$*) in* n_0*,* η_n^-*,* η_n^+*, and in the risk bounds (17)–(18). The same generalization holds for Eq. (21)–(22) and for consequences of Theorem 1 such as Proposition 3.*

Comments on the assumptions on $\mathscr{M}$ Assumption (**HId**) is barely an assumption since we can always add such a model to the collection considered (and it will never be selected by the procedure). It is used in the proof of Eq. (14) where we need to make sure that a model of large dimension and small bias exists.

Theorem 1 implicitly assumes that $\mathscr{M}$ contains a model of dimension at most $n/20$ with a small approximation error. This is much milder than the assumption of the corresponding results of [Arlot and Bach, 2009, Arlot and Bach, 2011, Arlot and Massart, 2009], which is that $\mathscr{M}$ contains a model of dimension at most $\sqrt{n}$ with an approximation error upper bounded by $\sigma^2\sqrt{\log(n)/n}$. Here, having a model of dimension $n/20$ with approximation error $\sigma^2/\log(n)$ is sufficient to get a consistent estimation of σ^2 and a first-order optimal model-selection procedure.

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

Finally, $\mathscr{M}$ is assumed to be finite, but Theorem 1 implicitly assumes a little more, since the event on which the result holds only has a large probability if $\mathrm{card}(\mathscr{M})n^{-\gamma}$ is small, which requires to take γ large enough. Since γ appears in all the bounds, assuming that it can be chosen fixed as n grows is equivalent to assuming that $\mathrm{card}(\mathscr{M})$ grows at most like a power of n, which excludes model collections of exponential complexity —that is, $\mathrm{card}(\mathscr{M}) \propto a^n$ for some $a > 0$. The case of exponential collections is discussed in Sections 4.7 and 8.3.4.

2.6. *Historical remarks*

Algorithms The slope heuristics and the corresponding data-driven penalty were first proposed by Birgé and Massart in a preprint [Birgé and Massart, 2001] and the subsequent article [Birgé and Massart, 2007]. They are also exposed in [Massart, 2005], [Blanchard and Massart, 2006, Section 2], [Massart, 2007, Section 8.5.2], and [Massart, 2008].

The term "slope" corresponds to the linear behavior of the empirical risk as a function of the dimension, as Algorithm 2 exploits.

The first implementation of data-driven penalties built upon the slope heuristics was expressed as a slope estimation, as in Algorithm 2; it was done by Letué [Letué, 2000, Section A.4] for penalized maximum likelihood, inspired by a preliminary version of [Birgé and Massart, 2001].

Several practical issues with Algorithm 2 were underlined in the context of change-point detection by Lebarbier [Lebarbier, 2002, Chapter 4], who then suggested to prefer the "dimension jump" formulation of Algorithm 1 which was present in the final version of [Birgé and Massart, 2001], as well as in [Massart, 2005, Birgé and Massart, 2007]. The drawbacks of Algorithm 1 were also underlined by [Lebarbier, 2002, Lebarbier, 2005] where some automatic ways to detect the dimension jump were proposed and tested on some synthetic data. Later on, Baudry, Maugis and Michel [Baudry et al., 2012] studied more deeply the practical use of Algorithms 1 and 2, with several variants (see also Section 7). The first proposition of detecting a jump over some sliding window was made by [Bontemps and Toussile, 2013], where only a finite set of values of C is considered; to the best of our knowledge, the continuous formulation for $\widehat{C}_{\mathrm{window}}$ is new, as well as the corresponding algorithm in Appendix B.2.

Theory The first theoretical results about the slope heuristics were proved in the setting of the present section, that is, regression on a fixed design with the least-squares risk and projection (least-squares) estimators. In [Birgé and Massart, 2001, Birgé and Massart, 2007], the first results obtained were similar to Eq. (14), (15), (17), and (18), making slightly stronger assumptions. A result similar to Eq. (15) was even published previously in [Birgé and Massart, 2001], but only in the restrictive case $F = 0$.

The first result showing the existence of a jump —that is, Eq. (14) and (16) holding simultaneously for all C on the same large-probability event— was obtained for least-squares regression on a random design with regressogram estimators [Arlot and Massart, 2009]. It was then proved in the fixed-design setting with more general estimators including projection estimators [Arlot and Bach, 2009, Arlot and Bach, 2011].

Eq. (17) is a corollary of a classical non-asymptotic oracle inequality for C_p-like penalties; similar results were known before the introduction of the slope heuristics, see for instance [Barron et al., 1999]. Eq. (18) is more precise because of the constant $1 + \mathrm{o}(1)$ in front of the oracle

risk, which was first obtained by [Birgé and Massart, 2001, Birgé and Massart, 2007].

The extension of Theorem 1 to sub-Gaussian noise (Remark 1 in Section 2.5) is new, to the best of our knowledge.

2.7. *Proof of Theorem 1*

The proof mixes ideas from [Birgé and Massart, 2007, Arlot and Bach, 2011]. We split it into three main steps, the last two ones being split themselves into several substeps: (1) using concentration inequalities, (2) proving the existence of a dimension jump (Eq. (14)–(16)), and (3) proving risk bounds thanks to a general oracle inequality (Eq. (17)–(18)).

We define $n_0(\gamma)$ as the smallest integer such that $\gamma\log(n)/n \leqslant 1/80^2$ for every $n \geqslant n_0(\gamma)$. At various places in the proof (in steps 2.3, 3.2, and 3.3), we make use of the inequality $2\sqrt{ab} \leqslant \theta a + \theta^{-1} b$ for all $a, b, \theta > 0$.

Step 1: concentration inequalities As explained in Section 2.3, the slope heuristics relies on the fact that $\|\widehat{F}_m - Y\|^2$ is close to its expectation. Let $x \geqslant 0$ be fixed. Given Eq. (5)–(6), for every $m \in \mathscr{M}$, we consider the event $\Omega_{m,x}$ on which the following two inequalities hold simultaneously:

$$\left|\langle \varepsilon, \Pi_m \varepsilon\rangle - \sigma^2 D_m\right| \leqslant 2\sigma^2\sqrt{xD_m} + 2x\sigma^2 \tag{23}$$

$$\left|\left\langle \varepsilon, (I_n - \Pi_m)F\right\rangle\right| \leqslant \sigma\sqrt{2x}\left\|(I_n - \Pi_m)F\right\|. \tag{24}$$

Under (**HG**), by standard Gaussian concentration results —for instance [Arlot and Bach, 2011, Propositions 4 and 6]—, we have

$$\mathbb{P}\left(\Omega_{m,x}\right) \geqslant 1 - 4\mathrm{e}^{-x}.$$

Then, defining $\Omega_x := \bigcap_{m\in\mathscr{M}} \Omega_{m,x}$, the union bound gives

$$\mathbb{P}\left(\Omega_x\right) \geqslant 1 - 4\,\mathrm{card}(\mathscr{M})\mathrm{e}^{-x}$$

and it is sufficient to prove that Eq. (14)–(18) hold true on Ω_x with $x = \gamma\log(n)$.

From now on, we restrict ourselves to the event Ω_x.

Step 2: existence of a dimension jump For proving Eq. (14) and (16), we show that $\widehat{m}(C)$ minimizes a quantity $G_C(m)$ close to $\mathrm{crit}_C(m)$, and then we show that $G_C(m_1)$ (resp. $G_C(m_2)$, for some well-chosen $m_2 \in \mathscr{M}$) is smaller than $G_C(m)$ for any model m with $D_m < 9n/10$ (resp. $D_m > n/10$).

Step 2.1: control of the difference between $\mathrm{crit}_C(m)$ and the quantity minimized by $\widehat{m}(C)$ Let $C \geqslant 0$. By Eq. (10), (5), and (6), since $\|\varepsilon\|^2$ doesn't depend from m, $\widehat{m}(C)$ minimizes

$$\begin{aligned}
G_C(m) &:= \frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2 + C\frac{D_m}{n} - \frac{1}{n}\|\varepsilon\|^2 \\
&= \frac{1}{n}\left\|(I_n - \Pi_m)F\right\|^2 - \frac{1}{n}\langle\varepsilon, \Pi_m\varepsilon\rangle + C\frac{D_m}{n} + \frac{2}{n}\left\langle\varepsilon, (I_n - \Pi_m)F\right\rangle \\
&= \mathrm{crit}_C(m) - \left(\frac{1}{n}\langle\varepsilon, \Pi_m\varepsilon\rangle - \sigma^2 D_m\right) + \frac{2}{n}\left\langle\varepsilon, (I_n - \Pi_m)F\right\rangle
\end{aligned}$$

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

where crit_C is defined by Eq. (11). Therefore, by Eq. (23)–(24) and using $D_m \leqslant n$, for every $m \in \mathscr{M}$,

$$\left|G_C(m) - \mathrm{crit}_C(m)\right| \leqslant 2\sigma^2\left(\sqrt{\frac{x}{n}} + \frac{x}{n}\right) + \frac{2\sigma\sqrt{2x}}{n}\left\|(I_n - \Pi_m)F\right\|. \tag{25}$$

Step 2.2: lower bound on $D_{\widehat{m}(C)}$ when C is too small (proof of Eq. (14)) Let $C \in [0, \sigma^2)$. Since $\widehat{m}(C)$ minimizes $G_C(m)$ over $m \in \mathscr{M}$, it is sufficient to prove that if C is far enough from σ^2,

$$G_C(m_1) < \inf_{m \in \mathscr{M}, D_m < 9n/10}\left\{G_C(m)\right\} \tag{26}$$

where m_1 is given by (**HId**). On the one hand, by Eq. (25),

$$G_C(m_1) \leqslant \mathrm{crit}_C(m_1) + 2\sigma^2\left(\sqrt{\frac{x}{n}} + \frac{x}{n}\right) = C - \sigma^2 + 2\sigma^2\left(\sqrt{\frac{x}{n}} + \frac{x}{n}\right). \tag{27}$$

On the other hand, by Eq. (25), for any $m \in \mathscr{M}$ such that $D_m < 9n/10$,

$$\begin{aligned} G_C(m) &\geqslant \frac{(C-\sigma^2)D_m}{n} - 2\sigma^2\left(\sqrt{\frac{x}{n}} + \frac{x}{n}\right) + \frac{1}{n}\left\|(I_n - \Pi_m)F\right\|^2 - \frac{2\sigma\sqrt{2x}}{n}\left\|(I_n - \Pi_m)F\right\| \\ &> \frac{9}{10}(C - \sigma^2) - 2\sigma^2\left(\sqrt{\frac{x}{n}} + \frac{2x}{n}\right). \end{aligned} \tag{28}$$

To conclude, the upper bound in Eq. (27) is smaller than the lower bound in Eq. (28) when

$$C \leqslant \sigma^2\left(1 - 40\sqrt{\frac{x}{n}} - 60\frac{x}{n}\right) =: \widetilde{C}_1(x). \tag{29}$$

Taking $x = \gamma\log(n)$, for $n \geqslant n_0(\gamma)$, we have $\widetilde{C}_1(x) \geqslant \sigma^2(1 - \eta_n^-)$ hence Eq. (14).

Remark that the same reasoning with $9n/10$ replaced by any $a_n \in [0, n)$ proves that $D_{\widehat{m}(C)} \geqslant a_n$ for every

$$C \leqslant \sigma^2\left(1 - \frac{4\sqrt{\frac{x}{n}} + 6\frac{x}{n}}{1 - \frac{a_n}{n}}\right) =: C_1(x; a_n). \tag{30}$$

We get Eq. (21) by taking $x = \gamma\log(n)$ and using that $x/n \leqslant 1/60^2$ since $n \geqslant n_0(\gamma)$.

Step 2.3: upper bound on $D_{\widehat{m}(C)}$ when C is large enough (proof of Eq. (16)) Let $C > \sigma^2$. Similarly to the proof of Eq. (14), it is sufficient to prove that if C is far enough from σ^2,

$$G_C(m_2) < \inf_{m \in \mathscr{M}, D_m > n/10}\left\{G_C(m)\right\} \tag{31}$$

where $m_2 \in \mathrm{argmin}_{m \in \mathscr{M}\,/\,D_m \leqslant n/20}\left\{\mathbb{E}\left[\|(I_n - \Pi_m)F\|^2\right]\right\}$ exists by assumption. For any $c_n \in [0, n]$, let us define

$$\mathscr{B}(c_n) := \inf_{m \in \mathscr{M}\,/\,D_m \leqslant c_n}\left\{\frac{1}{n}\left\|(I_n - \Pi_m)F\right\|^2\right\},$$

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

so that m_2 has an approximation error equal to $\mathscr{B}(n/20)$. On the one hand, by Eq. (25),

$$\begin{aligned} G_C(m_2) &\leqslant \mathrm{crit}_C(m_2) + 2\sigma^2\left(\sqrt{\frac{x}{n}} + \frac{x}{n}\right) + \frac{2\sigma\sqrt{2x}}{n}\left\|(I_n - \Pi_{m_2})F\right\| \\ &\leqslant \frac{2}{n}\left\|(I_n - \Pi_{m_2})F\right\|^2 + \frac{(C-\sigma^2)D_{m_2}}{n} + 2\sigma^2\left(\sqrt{\frac{x}{n}} + \frac{2x}{n}\right) \\ &\leqslant 2\mathscr{B}\left(\frac{n}{20}\right) + (C-\sigma^2)\frac{n/20}{n} + 2\sigma^2\left(\sqrt{\frac{x}{n}} + \frac{2x}{n}\right). \end{aligned} \tag{32}$$

On the other hand, by Eq. (25), for any $m \in \mathscr{M}$ such that $D_m > n/10$,

$$G_C(m) > \frac{n/10}{n}(C-\sigma^2) - 2\sigma^2\left(\sqrt{\frac{x}{n}} + \frac{2x}{n}\right). \tag{33}$$

To conclude, the upper bound in Eq. (32) is smaller than the lower bound in Eq. (33) when

$$C \geqslant \sigma^2\left[1 + 80\left(\sqrt{\frac{x}{n}} + \frac{2x}{n}\right)\right] + 40\mathscr{B}\left(\frac{n}{20}\right) =: \widetilde{C}_2(x). \tag{34}$$

Taking $x = \gamma\log(n)$, for $n \geqslant n_0(\gamma)$, we have $\widetilde{C}_2(x) \leqslant \sigma^2(1+\eta_n^+)$ hence Eq. (16).

Remark that if $\mathscr{M}$ contains a model of dimension at most $c_n \in [0,n)$, the same reasoning with $n/10$ replaced by any $b_n \in (c_n, n]$ and $n/20$ replaced by c_n proves that $D_{\widehat{m}(C)} \leqslant b_n$ for every

$$C \geqslant \sigma^2\left[1 + \frac{4n}{b_n - c_n}\left(\sqrt{\frac{x}{n}} + \frac{2x}{n}\right)\right] + \frac{2n}{b_n - c_n}\mathscr{B}(c_n) =: C_2(x; b_n; c_n). \tag{35}$$

We get Eq. (22) by taking $x = \gamma\log(n)$ and using that $x/n \leqslant 1/80^2$ since $n \geqslant n_0(\gamma)$.

Until the end of the proof, we fix $x = \gamma\log(n)$.

Step 2.4: lower bound on the risk of large models (proof of Eq. (15)) This is a straightforward consequence of Eq. (14). Indeed, on Ω_x, for any $m \in \mathscr{M}$ such that $D_m \geqslant 9n/10$,

$$\begin{aligned} \frac{1}{n}\left\|F - \widehat{F}_m\right\|^2 &= \frac{1}{n}\left\|(I_n - \Pi_m)F\right\|^2 + \frac{1}{n}\langle\varepsilon, \Pi_m\varepsilon\rangle \\ &\geqslant \frac{\sigma^2}{n}\left(D_m - 2\sqrt{xD_m} - 2x\right) = \frac{\sigma^2}{n}\left[\left(\sqrt{D_m} - \sqrt{x}\right)^2 - 3x\right] \geqslant \frac{7\sigma^2}{8}, \end{aligned}$$

where we use that $x/n \leqslant 1/77^2$ since $n \geqslant n_0(\gamma)$.

Step 3: upper bounds on the risk For proving Eq. (17)–(18), we prove a slightly more general oracle inequality —Eq. (43)— using the classical approach used for instance by [Birgé and Massart, 2001, Massart, 2007, Arlot and Bach, 2011].

Step 3.1: general approach for proving an oracle inequality Following Section 2.2, an ideal penalty is

$$\mathrm{pen}_{\mathrm{id}}(m) := \frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 - \frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2 + \|\varepsilon\|^2$$

which has expectation $2\sigma^2 D_m/n = \mathrm{pen}_{\mathrm{opt}}(m)$. A key argument for getting an oracle inequality is that $\mathrm{pen}_{\mathrm{id}}(m)$ concentrates around its expectation. Indeed, let us define

$$\Delta(m) := \mathrm{pen}_{\mathrm{id}}(m) - \frac{2\sigma^2 D_m}{n} = \frac{2}{n}\left(\langle \varepsilon, \Pi_m \varepsilon\rangle - \sigma^2 D_m\right) - \frac{2}{n}\left\langle \varepsilon, (I_n - \Pi_m)F\right\rangle, \tag{36}$$

where the second formulation is a consequence of Eq. (6). Then, by Eq. (10), for any $C \geqslant 0$ and $m \in \mathscr{M}$,

$$\frac{1}{n}\left\|\widehat{F}_{\widehat{m}(C)} - Y\right\|^2 + \frac{C D_{\widehat{m}(C)}}{n} \leqslant \frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2 + \frac{C D_m}{n}$$

which is equivalent to

$$\frac{1}{n}\left\|\widehat{F}_{\widehat{m}(C)} - F\right\|^2 - \Delta\big(\widehat{m}(C)\big) + \frac{(C-2\sigma^2)D_{\widehat{m}(C)}}{n} \leqslant \frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 - \Delta(m) + \frac{(C-2\sigma^2)D_m}{n}. \tag{37}$$

It remains to show that $\Delta(m)$ and $(C-2\sigma^2)D_m/n$ are small compared to $n^{-1}\|\widehat{F}_m - F\|^2$ for all $m \in \mathscr{M}$. Recall that we restrict ourselves to the event Ω_x until the end of the proof, with $x = \gamma \log(n)$.

Step 3.2: control of $\Delta(m)$ By Eq. (23), (24), and (36), for every $m \in \mathscr{M}$ and $\theta > 0$,

$$\begin{aligned}\left|\Delta(m)\right| &\leqslant \frac{2}{n}\left[2\sigma^2\sqrt{xD_m} + 2\sigma^2 x + \sigma\sqrt{2x}\left\|(I_n - \Pi_m)F\right\|\right] \\ &\leqslant 2\theta\,\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] + \frac{\sigma^2 x}{n}\left(3\theta^{-1} + 4\right).\end{aligned} \tag{38}$$

Step 3.3: upper bound on the expected risk in terms of risk By Eq. (5) and (24), for every $m \in \mathscr{M}$ and $\theta' > 0$,

$$\begin{aligned}\left\|\widehat{F}_m - F\right\|^2 &= \mathbb{E}\left[\left\|\widehat{F}_m - F\right\|^2\right] + \langle \varepsilon, \Pi_m \varepsilon\rangle - \sigma^2 D_m \\ &\geqslant \mathbb{E}\left[\left\|\widehat{F}_m - F\right\|^2\right] - \sigma^2\left(2\sqrt{xD_m} + 2x\right) \\ &\geqslant \left(1-\theta'\right)\mathbb{E}\left[\left\|\widehat{F}_m - F\right\|^2\right] - x\sigma^2\left(2 + \theta'^{-1}\right)\end{aligned}$$

so that, for every $\theta' \in (0,1)$,

$$\mathbb{E}\left[\left\|\widehat{F}_m - F\right\|^2\right] \leqslant \frac{1}{1-\theta'}\left\|\widehat{F}_m - F\right\|^2 + \kappa(\theta')x\sigma^2 \qquad \text{with} \quad \kappa(\theta') := \frac{2 + \frac{1}{\theta'}}{1-\theta'}. \tag{39}$$

Step 3.4: control of the remainder terms appearing in Eq. (37) Combining Eq. (38) and (39), we get on the one hand that for every $m \in \mathscr{M}$, $\theta > 0$, $\theta' \in (0,1)$,

$$\begin{aligned}
&\Delta(m) + \frac{(2\sigma^2 - C)D_m}{n} \\
&\leqslant \left[2\theta + \left(2 - \frac{C}{\sigma^2}\right)_+\right] \mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] + \frac{\sigma^2 x}{n}\left(\frac{3}{\theta} + 4\right) \\
&\leqslant \frac{2\theta + \left(2 - \frac{C}{\sigma^2}\right)_+}{1-\theta'} \frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 + \frac{\sigma^2 x}{n}\left(\frac{3}{\theta} + 4 + \kappa(\theta')\left[2\theta + \left(2 - \frac{C}{\sigma^2}\right)_+\right]\right). \qquad (40)
\end{aligned}$$

On the other hand, similarly, for every $m \in \mathscr{M}$, $\theta > 0$, $\theta' \in (0,1)$,

$$\begin{aligned}
&-\Delta(m) + \frac{(C - 2\sigma^2)D_m}{n} \\
&\leqslant \frac{2\theta + \left(\frac{C}{\sigma^2} - 2\right)_+}{1-\theta'} \frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 + \frac{\sigma^2 x}{n}\left(\frac{3}{\theta} + 4 + \kappa(\theta')\left[2\theta + \left(\frac{C}{\sigma^2} - 2\right)_+\right]\right). \qquad (41)
\end{aligned}$$

Step 3.5: proof of a general oracle inequality Combining Eq. (37), (40), and (41) with $\theta' = 2\theta \in (0,1)$, we get that for every $\theta \in (0,1/2)$ and $m \in \mathscr{M}$,

$$\begin{aligned}
&\left[1 - \frac{2\theta + \left(2 - \frac{C}{\sigma^2}\right)_+}{1-2\theta}\right] \frac{1}{n}\left\|\widehat{F}_{\widehat{m}(C)} - F\right\|^2 \\
&\qquad \leqslant \left[1 + \frac{2\theta + \left(\frac{C}{\sigma^2} - 2\right)_+}{1-2\theta}\right] \frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 + \frac{\sigma^2 x}{n} R_1(\theta, C\sigma^{-2}) \qquad (42) \\
\text{with} \qquad R_1(\theta, C\sigma^{-2}) &:= \frac{6}{\theta} + 8 + \kappa(2\theta)\left(4\theta + \left|\frac{C}{\sigma^2} - 2\right|\right).
\end{aligned}$$

Let us assume $C > \sigma^2$. For any $\delta \in (0,1]$, we choose

$$\theta = \theta^\star(\delta, C\sigma^{-2}) := \frac{\delta}{4} \frac{\left[1 - \left(2 - \frac{C}{\sigma^2}\right)_+\right]^2}{1 + \left(\frac{C}{\sigma^2} - 2\right)_+ + \delta\left[1 - \left(2 - \frac{C}{\sigma^2}\right)_+\right]} < \frac{\delta}{4} \leqslant \frac{1}{4}.$$

So, if $C \geqslant (1+\delta)\sigma^2$, we have $C > (1+4\theta)\sigma^2$ hence we can divide both sides of Eq. (42) by

$$1 - \frac{2\theta + \left(2 - \frac{C}{\sigma^2}\right)_+}{1-2\theta} > 0.$$

Remark that

$$\left[1 + \frac{2\theta + \left(\frac{C}{\sigma^2} - 2\right)_+}{1-2\theta}\right] \times \left[1 - \frac{2\theta + \left(2 - \frac{C}{\sigma^2}\right)_+}{1-2\theta}\right]^{-1} = \frac{1 + \left(\frac{C}{\sigma^2} - 2\right)_+}{1 - 4\theta - \left(2 - \frac{C}{\sigma^2}\right)_+} = \frac{1 + \left(\frac{C}{\sigma^2} - 2\right)_+}{1 - \left(2 - \frac{C}{\sigma^2}\right)_+} + \delta$$

where the last equality uses $\theta = \theta^\star(\delta, C\sigma^{-2})$. So, if $C \geqslant (1+\delta)\sigma^2$, Eq. (42) leads to

$$\frac{1}{n}\left\|\widehat{F}_{\widehat{m}(C)} - F\right\|^2 \leqslant \left(\frac{1+\left(\frac{C}{\sigma^2}-2\right)_+}{1-\left(2-\frac{C}{\sigma^2}\right)_+} + \delta\right)\left[\inf_{m\in\mathscr{M}}\left\{\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right\} + \frac{\sigma^2 x}{n} R_2\left(\delta, \frac{C}{\sigma^2}\right)\right] \tag{43}$$

where for every $\delta \in (0,1]$ and $u \in (1,+\infty)$,

$$\begin{aligned} R_2(\delta,u) &= R_1\left(\theta^\star(\delta,u),u\right) \\ &\leqslant \left(10+2|u-2|\right)\theta^\star(\delta,u)^{-1}. \end{aligned}$$

Therefore, for every $C \geqslant (1+\delta)\sigma^2$,

$$R_2\left(\delta, \frac{C}{\sigma^2}\right) \leqslant \frac{8}{\delta}\left(5+|C\sigma^{-2}-2|\right)\max\left\{2+(C\sigma^{-2}-2)_+, \frac{2}{\left[1-(2-C\sigma^{-2})_+\right]^2}\right\}.$$

Step 3.6: risk bound for $\widehat{m}(C)$ when C is large enough (proof of Eq. (17)) In this step, we assume $C > \sigma^2$. When $C/\sigma^2 \in (1,2]$, Eq. (43) with $\delta = C\sigma^{-2} - 1 \in (0,1]$ yields

$$\frac{1}{n}\left\|\widehat{F}_{\widehat{m}(C)} - F\right\|^2 \leqslant 2\left(\frac{C}{\sigma^2}-1\right)^{-1}\left[\inf_{m\in\mathscr{M}}\left\{\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right\} + \frac{96\sigma^2 x}{n}\left(\frac{C}{\sigma^2}-1\right)^{-3}\right].$$

When $C/\sigma^2 \geqslant 2$, Eq. (43) with $\delta = 1$ yields

$$\frac{1}{n}\left\|\widehat{F}_{\widehat{m}(C)} - F\right\|^2 \leqslant \frac{C}{\sigma^2}\left[\inf_{m\in\mathscr{M}}\left\{\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right\} + \frac{20\sigma^2 x}{n}\left(\frac{C}{\sigma^2}\right)^2\right].$$

Step 3.7: oracle inequality for $\widehat{m}(C)$ when C is close to $2\sigma^2$ (proof of Eq. (18)) Now, we assume $C/\sigma^2 \in [2-\eta, 2+\eta]$ with $\eta \in [0,1/2]$. Taking $\delta = \eta$ in Eq. (43) yields

$$\begin{aligned} \frac{1}{n}\left\|\widehat{F}_{\widehat{m}(C)} - F\right\|^2 &\leqslant \left(\max\left\{1+\eta, \frac{1}{1-\eta}\right\} + \eta\right) \\ &\quad \times \left[\inf_{m\in\mathscr{M}}\left\{\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right\} + \frac{\sigma^2 x}{n}\frac{8}{\eta}(5+\eta)\max\left\{2+\eta, \frac{2}{(1-\eta)^2}\right\}\right] \\ &\leqslant (1+3\eta)\inf_{m\in\mathscr{M}}\left\{\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right\} + \frac{\sigma^2 x}{n}\frac{8}{\eta}(1+3\eta)(5+\eta)\max\left\{2+\eta, \frac{2}{(1-\eta)^2}\right\} \\ &\leqslant (1+3\eta)\inf_{m\in\mathscr{M}}\left\{\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right\} + \frac{880\sigma^2 x}{\eta n}, \end{aligned}$$

using that $1/(1-\eta) \leqslant 1+2\eta$ for every $\eta \in [0,1/2]$. □

3. Generalizing the slope heuristics

The slope heuristics has first been formulated and theoretically validated in the framework of Section 2. Then, it rapidly became a more general heuristics for building data-driven optimal penalties. This section discusses two possible formulations for its generalization.

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

3.1. General framework

Before going any further, we need to introduce a general model/estimator-selection framework. Let $\mathbb{S}$ be some set, $\mathscr{R} : \mathbb{S} \mapsto [0, +\infty)$ be some risk function, and assume that our goal is to build from data some estimator $\widehat{s} \in \mathbb{S}$ such that $\mathscr{R}(\widehat{s})$ is as small as possible. Let $(\widehat{s}_m)_{m \in \mathscr{M}}$ be a collection of estimators. The goal of estimator selection is to choose from data some $\widehat{m} \in \mathscr{M}$ such that the risk of $\widehat{s}_{\widehat{m}}$ is as small as possible, that is, satisfying an oracle inequality

$$\mathscr{R}\left(\widehat{s}_{\widehat{m}}\right) - \mathscr{R}(s^\star) \leqslant K_n \inf_{m \in \mathscr{M}} \left\{ \mathscr{R}(\widehat{s}_m) - \mathscr{R}(s^\star) \right\} + R_n \tag{44}$$

with large probability, where $\mathscr{R}(s^\star) := \inf_{t \in \mathbb{S}} \mathscr{R}(t)$. In the following, K_n is called "the leading constant" of the oracle inequality (44). Let $\widehat{\mathscr{R}}_n : \mathbb{S} \mapsto [0, +\infty)$ be the empirical risk associated with $\mathscr{R}$, that is, we assume throughout Sections 3–5 that $\forall t \in \mathbb{S}$, $\mathbb{E}[\widehat{\mathscr{R}}_n(t)] = \mathscr{R}(t)$.

This framework includes the one of Section 2 by taking $\mathbb{S} = \mathbb{R}^n$, $\mathscr{R}(t) = n^{-1}\|t - F\|^2 - \sigma^2$, $\widehat{s}_m = \widehat{F}_m$ the projection estimator associated with some model S_m for every $m \in \mathscr{M}$, and $\widehat{\mathscr{R}}_n(t) = n^{-1}\|t - Y\|^2$. Many other classical settings also fit into this framework, such as density estimation with the Kullback risk or the L^2 risk, random-design regression with the L^2 risk, and classification with the $0-1$ risk, see [Arlot and Celisse, 2010, Section 1] for details.

3.2. Penalties known up to some constant factor

The most natural extension of the slope heuristics is to generalize it to all frameworks where a penalty is known up to some multiplicative constant [Massart, 2005, Blanchard and Massart, 2006], that is, if theoretical results show that a good penalty is $C^\star \mathrm{pen}_1(m)$ with pen_1 known but $C^\star$ unknown. Penalties known up to a constant factor appear in several frameworks, for four main reasons:

1. A penalty satisfying an optimal oracle inequality —that is, with a leading constant $1 + o(1)$— is theoretically known, but involves *unknown quantities* in practice, such as the noise-level σ^2 for Mallows' C_p and C_L [Mallows, 1973], see Sections 2 and 3.3.
2. An optimal penalty pen_1 is known theoretically and in practice, but only *asymptotically*, that is, the (unknown) nonasymptotic optimal penalty is $C_n^\star \mathrm{pen}_1$ with $C_n^\star \to 1$ as the sample size n tends to infinity, but $C_n^\star$ is unknown and can be far from 1 for finite sample sizes. For instance, AIC [Akaike, 1973] and BIC [Schwarz, 1978] penalties for maximum likelihood rely on asymptotic computations. Section 8.4 explains why such a problem can arise in almost any framework.
3. An optimal penalty is obtained by resampling, hence depending on some multiplicative factor that might depend on unknown quantities or be correct only for n large enough, see [Arlot, 2009] and Remark 3 in Section 4.2.
4. A penalty $C\,\mathrm{pen}_1$ satisfying an oracle inequality with a leading constant $\mathscr{O}(1)$ when C is well-chosen is known theoretically, but theoretical results are not precise enough to specify the optimal value $C^\star$ of C. This occurs for instance for change-point detection [Comte and Rozenholc, 2004, Lebarbier, 2005], density estimation with Gaussian mixtures [Maugis and Michel, 2011b], and local Rademacher complexities in classification

[Bartlett et al., 2005, Koltchinskii, 2006]. In some frameworks, some partial information is available about the optimal value of the constant: in binary classification, global Rademacher complexities differ from a factor two between theory [Koltchinskii, 2001] and practice [Lozano, 2000]. Note that in such cases, it might happen that $C^\star \operatorname{pen}_1$ is not exactly an optimal penalty, so that no oracle inequality with leading constant $1+o(1)$ can be obtained; nevertheless, choosing the constant C in the penalty $C\operatorname{pen}_1$ remains an important practical problem.

Then, if for every $m \in \mathscr{M}$, $\mathscr{C}_m$ measures the "complexity" of $\widehat{s}_m$, the slope heuristics suggests to generalize Algorithm 1 into the following.

Algorithm 3. `Input:` $\left(\widehat{\mathscr{R}}_n(\widehat{s}_m)\right)_{m\in\mathscr{M}}$, $\left(\operatorname{pen}_1(m)\right)_{m\in\mathscr{M}}$, *and* $(\mathscr{C}_m)_{m\in\mathscr{M}}$.

1. *Compute* $\left(\widehat{m}_1(C)\right)_{C\geqslant 0}$, *where for every* $C \geqslant 0$,
$$\widehat{m}_1(C) \in \operatorname*{argmin}_{m\in\mathscr{M}} \left\{ \widehat{\mathscr{R}}_n(\widehat{s}_m) + C\operatorname{pen}_1(m) \right\}. \tag{45}$$
2. *Find* $\widehat{C}_{\mathrm{jump}} > 0$ *corresponding to the "unique large jump" of* $C \mapsto \mathscr{C}_{\widehat{m}_1(C)}$.
3. *Select* $\widehat{m}_{\mathrm{Alg.3}} \in \operatorname{argmin}_{m\in\mathscr{M}} \{\widehat{\mathscr{R}}_n(\widehat{s}_m) + 2\widehat{C}_{\mathrm{jump}}\operatorname{pen}_1(m)\}$.

`Output:` $\widehat{m}_{\mathrm{Alg.3}}$.

Algorithm 3 relies on two ideas: (i) Eq. (13) $\operatorname{pen}_{\mathrm{opt}} = 2\operatorname{pen}_{\mathrm{min}}$ is valid in a more general framework than least-squares regression and projection estimators, and (ii) if a proper complexity measure $\mathscr{C}_m$ is used instead of the dimension D_m of the models, the minimal penalty can be caracterized empirically by a jump of $\mathscr{C}_{\widehat{m}_1(C)}$.

3.3. *Algorithm 3 fails for linear estimator selection*

We now illustrate on an example why Algorithm 3 can fail, before showing how to correct it in Section 3.4. Let us consider the fixed-design regression framework of Section 2.1 with linear estimators instead of projection estimators, that is, for every $m \in \mathscr{M}$,
$$\widehat{F}_m = A_m Y$$
for some deterministic linear mapping $A_m : \mathbb{R}^n \to \mathbb{R}^n$. For instance, projection estimators are linear estimators since the orthogonal projection $A_m = \Pi_m$ onto a linear space S_m is linear. Other examples include kernel ridge regression or spline smoothing, nearest-neighbor regression, Nadaraya-Watson estimators, see [Arlot and Bach, 2011] for more examples and references.

As in Section 2.2, expectations of the risk and empirical risk of a linear estimator can be computed as follows:
$$\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] = \frac{1}{n}\left\|(A_m - I_n)F\right\|^2 + \frac{\sigma^2 \operatorname{tr}\left(A_m^\top A_m\right)}{n}, \tag{46}$$
$$\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2\right] = \frac{1}{n}\left\|(A_m - I_n)F\right\|^2 + \frac{\sigma^2\left[n + \operatorname{tr}\left(A_m^\top A_m\right) - 2\operatorname{tr}(A_m)\right]}{n}, \tag{47}$$
$$\text{and} \qquad \operatorname{pen}_{\mathrm{opt}}(m) = \mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] - \mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2\right] + \sigma^2 = \frac{2\sigma^2 \operatorname{tr}(A_m)}{n}. \tag{48}$$

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

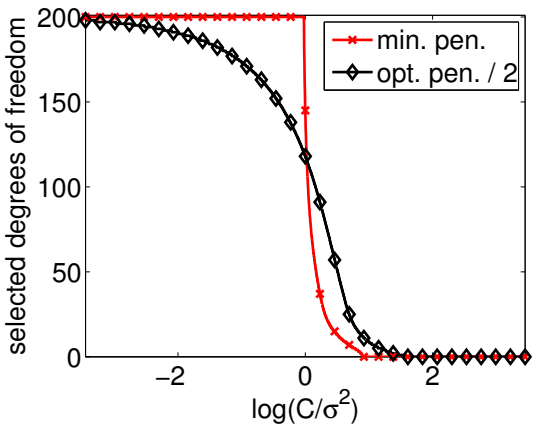


FIGURE 3. *The minimal penalty is not proportional to* $\mathrm{tr}(A_m)$ *for kernel ridge estimators* (*Figure taken from [Arlot and Bach, 2011], 'kernel ridge' framework, see Appendix D for details*): $C \mapsto \mathscr{C}_{\widehat{m}_1(C)}$ *for Algorithm 3 with* $\mathrm{pen}_1(m) = \mathrm{tr}(A_m)/n$ *and* $\mathscr{C}_m = \mathrm{tr}(A_m)$ (*black curve / diamonds*), *and* $C \mapsto \mathrm{tr}(A_{\widehat{m}^{\mathrm{lin}}_{\min}(C)})$ *for Algorithm 4* (*red curve / crosses*) *with linear estimators* (*kernel ridge*).

Eq. (46) can be interpreted as a bias-variance decomposition similarly to Eq. (7). The optimal penalty given by Eq. (48) has been called C_L by Mallows [Mallows, 1973] and is similar to C_p, with the dimension D_m replaced by the *degrees of freedom* $\mathrm{tr}(A_m)$. It also depends on σ^2 which is unknown, so one could think of using Algorithm 3 with $\widehat{s}_m = \widehat{F}_m$, $\widehat{\mathscr{R}}_n(t) = n^{-1}\|t - Y\|^2$, $\mathrm{pen}_1(m) = \mathrm{tr}(A_m)/n$, and $\mathscr{C}_m = \mathrm{tr}(A_m)$. Then, plotting $\mathscr{C}_{\widehat{m}_1(C)}$ as a function of C, what we typically get is shown in Figure 3: no clear jump of the complexity is observed around σ^2, contrary to what Algorithm 3 predicts.

3.4. *Minimal-penalty heuristics for linear estimators*

Following [Arlot and Bach, 2009, Arlot and Bach, 2011], the correct minimal penalty in the linear estimators framework is

$$\mathrm{pen}^{\mathrm{lin}}_{\min}(m) := \frac{\sigma^2\left[2\,\mathrm{tr}(A_m) - \mathrm{tr}\left(A_m^\top A_m\right)\right]}{n}\,.$$

Indeed, as in Section 2.3, let us consider, for every $C \geqslant 0$,

$$\widehat{m}^{\mathrm{lin}}_{\min}(C) \in \operatorname*{argmin}_{m\in\mathscr{M}} \left\{ \frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2 + C\frac{2\,\mathrm{tr}(A_m) - \mathrm{tr}\left(A_m^\top A_m\right)}{n} \right\}, \tag{49}$$

$$\text{and} \quad m^{\star\mathrm{lin}}_{\min}(C) \in \operatorname*{argmin}_{m\in\mathscr{M}} \left\{ \mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2 + C\frac{2\,\mathrm{tr}(A_m) - \mathrm{tr}\left(A_m^\top A_m\right)}{n}\right] \right\}$$

$$= \operatorname*{argmin}_{m\in\mathscr{M}} \left\{\mathrm{crit}^{\mathrm{lin}}_C(m)\right\}$$

$$\text{with} \quad \mathrm{crit}^{\mathrm{lin}}_C(m) := \frac{1}{n}\left(\left\|(I_n - \Pi_m)F\right\|^2 + (C - \sigma^2)\left[2\,\mathrm{tr}(A_m) - \mathrm{tr}(A_m^\top A_m)\right]\right),$$

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

by Eq. (47). Let us assume that the approximation error term $n^{-1}\|(A_m - I_n)F\|^2$, which appears in Eq. (46), is a decreasing function of the degrees of freedom $\mathscr{C}_m = \mathrm{tr}(A_m)$. Then, we can distinguish two cases:

— if $C < \sigma^2$, then $\mathrm{crit}_C^{\mathrm{lin}}(m)$ is a decreasing function of $\mathscr{C}_m$, and $\mathscr{C}_{m^{\star\mathrm{lin}}_{\mathrm{min}}(C)}$ is huge: $m^{\star\mathrm{lin}}_{\mathrm{min}}(C)$ overfits.

— if $C > \sigma^2$, then $\mathrm{crit}_C^{\mathrm{lin}}(m)$ increases with $\mathscr{C}_m$ for $\mathscr{C}_m$ large enough, so $\mathscr{C}_{m^{\star\mathrm{lin}}_{\mathrm{min}}(C)}$ is much smaller than when $C < \sigma^2$.

This behavior is also the one of $\widehat{m}^{\mathrm{lin}}_{\mathrm{min}}(C)$, as illustrated in Figure 3 (red curve / crosses), which leads to the following algorithm.

Algorithm 4. `Input:` $\left(\|\widehat{F}_m - Y\|^2\right)_{m\in\mathscr{M}}$.

1. *Compute* $\left(\widehat{m}^{\mathrm{lin}}_{\mathrm{min}}(C)\right)_{C\geqslant 0}$, *where* $\widehat{m}^{\mathrm{lin}}_{\mathrm{min}}(C)$ *is defined by Eq.* (49).
2. *Find* $\widehat{C}_{\mathrm{jump}} > 0$ *corresponding to the "unique large jump" of* $C \mapsto \mathrm{tr}\left(A_{\widehat{m}^{\mathrm{lin}}_{\mathrm{min}}(C)}\right)$.
3. *Select* $\widehat{m}_{\mathrm{Alg.4}} \in \mathrm{argmin}_{m\in\mathscr{M}}\left\{n^{-1}\|\widehat{F}_m - Y\|^2 + 2\widehat{C}_{\mathrm{jump}}\,\mathrm{tr}(A_m)/n\right\}$.

`Output:` $\widehat{m}_{\mathrm{Alg.4}}$.

Theorem 1 can be extended to Algorithm 4, up to some minor changes in the assumptions and results [Arlot and Bach, 2009, Arlot and Bach, 2011]. For kernel ridge regression, Algorithm 4 is proved to work also for choosing over a continuous set $\mathscr{M}$ [Arlot and Bach, 2011], provided the kernel is fixed.

Note that when $\mathrm{tr}(A_m^\top A_m) = \mathrm{tr}(A_m)$, $\mathrm{pen}_{\mathrm{opt}}(m) = 2\,\mathrm{pen}^{\mathrm{lin}}_{\mathrm{min}}(m)$. This occurs for least-squares estimators —we then recover the setting of Section 2 and Algorithm 1, in which $A_m = A_m^\top A_m = \Pi_m$— and for k-nearest neighbors estimators.

3.5. *General minimal-penalty algorithm*

We now go back to the general setting of Section 3.1, and propose a generalization of Algorithms 1 and 4. Here, we suggest to take $\widehat{C}_{\mathrm{jump}} = \widehat{C}_{\mathrm{window}}$, but any other formal definition of $\widehat{C}_{\mathrm{jump}}$ could be used instead.

Algorithm 5. `Input:` $\left(\widehat{\mathscr{R}}_n(\widehat{s}_m)\right)_{m\in\mathscr{M}}$, $\left(\mathrm{pen}_0(m)\right)_{m\in\mathscr{M}}$, $\left(\mathrm{pen}_1(m)\right)_{m\in\mathscr{M}}$, $(\mathscr{C}_m)_{m\in\mathscr{M}}$, *and* $\eta \geqslant 0$.

1. *Compute* $\left(\widehat{m}^{(0)}_{\mathrm{min}}(C)\right)_{C\geqslant 0}$, *where for every* $C \geqslant 0$,

$$\widehat{m}^{(0)}_{\mathrm{min}}(C) \in \operatorname*{argmin}_{m\in\mathscr{M}}\left\{\widehat{\mathscr{R}}_n(\widehat{s}_m) + C\,\mathrm{pen}_0(m)\right\}. \tag{50}$$

2. *Find* $\widehat{C}_{\mathrm{jump}} > 0$ *corresponding to the "unique large jump" of* $C \mapsto \mathscr{C}_{\widehat{m}^{(0)}_{\mathrm{min}}(C)}$, *for instance,*

$$\widehat{C}_{\mathrm{window}} \in \operatorname*{argmax}_{C>0}\left\{\mathscr{C}_{\widehat{m}^{(0)}_{\mathrm{min}}(C/(1+\eta))} - \mathscr{C}_{\widehat{m}^{(0)}_{\mathrm{min}}(C(1+\eta))}\right\}.$$

3. *Select* $\widehat{m}_{\mathrm{Alg.5}} \in \mathrm{argmin}_{m\in\mathscr{M}}\left\{\widehat{\mathscr{R}}_n(\widehat{s}_m) + \widehat{C}_{\mathrm{jump}}\,\mathrm{pen}_1(m)\right\}$.

`Output:` $\widehat{m}_{\mathrm{Alg.5}}$.

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

Algorithm 5 implicitly assumes that the minimal and the optimal penalty are respectively equal to $C^\star \operatorname{pen}_0$ and $C^\star \operatorname{pen}_1$, with pen_0 and pen_1 known, but $C^\star$ unknown. We refer to Section 7.1 for practical remarks about the choice of $\widehat{C}_{\text{jump}}$. Computational issues are discussed in Section 7.2.

In the "slope heuristics" setting (Section 2), $\operatorname{pen}_1 = 2\operatorname{pen}_0$, and Algorithm 5 reduces to Algorithm 3.

Similarly to Algorithm 2, we can also propose a "slope" formulation of Algorithm 5.

Algorithm 6. `Input:` $\left(\widehat{\mathscr{R}}_n(\widehat{s}_m)\right)_{m\in\mathscr{M}}$, $\left(\operatorname{pen}_0(m)\right)_{m\in\mathscr{M}}$, $\left(\operatorname{pen}_1(m)\right)_{m\in\mathscr{M}}$, *and* $(\mathscr{C}_m)_{m\in\mathscr{M}}$.

1. *Estimate the slope* $\widehat{C}_{\text{slope}}$ *of* $-\widehat{\mathscr{R}}_n(\widehat{s}_m)$ *as a function of* $\operatorname{pen}_0(m)$ *for all* $m \in \mathscr{M}$ *with* $\mathscr{C}_m$ *"large enough", for instance by (robust) linear regression.*
2. *Select* $\widehat{m}_{\text{Alg.6}} \in \operatorname{argmin}_{m\in\mathscr{M}} \left\{ \widehat{\mathscr{R}}_n(\widehat{s}_m) + \widehat{C}_{\text{slope}} \operatorname{pen}_1(m) \right\}$.

`Output:` $\widehat{m}_{\text{Alg.6}}$.

What remains now is to identify natural candidates for being a minimal or an optimal penalty in the general setting.

3.6. *Optimal and minimal penalties*

In the general setting, the unbiased risk estimation heuristics [Akaike, 1970, Stein, 1981] suggests the following optimal (deterministic) penalty

$$\operatorname{pen}^{\text{gal}}_{\text{opt}}(m) := \mathbb{E}\left[\mathscr{R}(\widehat{s}_m) - \widehat{\mathscr{R}}_n(\widehat{s}_m)\right] \tag{51}$$

which generalizes formula (4). If $\mathscr{R}(\widehat{s}_m) - \widehat{\mathscr{R}}_n(\widehat{s}_m)$ is concentrated around its expectation uniformly over $m \in \mathscr{M}$ —which excludes too large collections $\mathscr{M}$—, one can prove an oracle inequality for the penalty (51), or for any penalty which differs from Eq. (51) by an additive term independent from m, as in Eq. (9) and (48).

Building a minimal penalty in the general setting is more difficult. For every $m \in \mathscr{M}$, let $\mathscr{C}_m$ be some "complexity measure" associated with $\widehat{s}_m$, that is, we assume that the empirical risk $\widehat{\mathscr{R}}_n(\widehat{s}_m)$ (or its expectation) is (approximately) a decreasing function of $\mathscr{C}_m$. Algorithm 6 suggests that the minimal penalty is a quantity which exactly compensates this decreasing trend, such as

$$\operatorname{pen}^{\text{gal}}_{\text{min},0}(m) := -\mathbb{E}\left[\widehat{\mathscr{R}}_n(\widehat{s}_m)\right]. \tag{52}$$

Nevertheless, in most cases including least-squares and linear estimators, $\operatorname{pen}^{\text{gal}}_{\text{min},0}$ is unknown, even up to a multiplicative factor, so that we need another candidate for being a minimal penalty.

For every $m \in \mathscr{M}$, let $s^\star_m \in \mathbb{S}$ be such that $\mathscr{R}(s^\star_m) - \inf_{t\in\mathbb{S}} \mathscr{R}(t)$ decreases to zero as $\mathscr{C}_m \to \infty$. As argued below, a natural choice for the minimal penalty is

$$\operatorname{pen}^{\text{gal}}_{\text{min}}(m) := \mathbb{E}\left[\widehat{\mathscr{R}}_n(s^\star_m) - \widehat{\mathscr{R}}_n(\widehat{s}_m)\right]. \tag{53}$$

Indeed, for every $C \geqslant 0$ let

$$\operatorname{crit}^{\text{gal}}_C(m) := \mathbb{E}\left[\widehat{\mathscr{R}}_n(\widehat{s}_m)\right] + C \operatorname{pen}^{\text{gal}}_{\text{min}}(m) = C\mathscr{R}(s^\star_m) + (1-C)\mathbb{E}\left[\widehat{\mathscr{R}}_n(\widehat{s}_m)\right] \tag{54}$$

so that $m_{\min}^{\mathrm{gal}\star}(C) \in \operatorname{argmin}_{m\in\mathscr{M}}\{\mathrm{crit}_C^{\mathrm{gal}}(m)\}$ is a proxy for

$$\widehat{m}_{\min}^{\mathrm{gal}}(C) \in \operatorname*{argmin}_{m\in\mathscr{M}}\left\{\widehat{\mathscr{R}}_n(\widehat{s}_m) + C\,\mathrm{pen}_{\min}^{\mathrm{gal}}(m)\right\}.$$

Let us assume for simplicity that $\widehat{\mathscr{R}}_n(\widehat{s}_m)$ is a decreasing function of $\mathscr{C}_m$. Then, when $C < 1$, $\mathrm{crit}_C^{\mathrm{gal}}(m)$ is a decreasing function of $\mathscr{C}_m$, so that $\mathscr{C}_{m_{\min}^{\mathrm{gal}\star}(C)} \approx \max_{m\in\mathscr{M}} \mathscr{C}_m$ which corresponds to overfitting. On the contrary, when $C > 1$, $(1-C)\mathbb{E}[\widehat{\mathscr{R}}_n(\widehat{s}_m)]$ is an increasing function of $\mathscr{C}_m$ while $C\mathscr{R}(s_m^\star)$ is approximately constant for $\mathscr{C}_m$ large enough, so that $\mathscr{C}_{m_{\min}^{\mathrm{gal}\star}(C)} \ll \max_{m\in\mathscr{M}} \mathscr{C}_m$. Therefore, if concentration inequalities show that $\widehat{m}_{\min}^{\mathrm{gal}}(C)$ behaves likes $m_{\min}^{\mathrm{gal}\star}(C)$, $\mathrm{pen}_{\min}^{\mathrm{gal}}$ is a minimal penalty.

Let us emphasize that for making use of the fact that $\mathrm{pen}_{\min}^{\mathrm{gal}}$ is a minimal penalty, we must assume that $\mathrm{pen}_{\min}^{\mathrm{gal}} \approx C^\star\,\mathrm{pen}_0$ and $\mathrm{pen}_{\mathrm{opt}}^{\mathrm{gal}} \approx C^\star\,\mathrm{pen}_1$ for some unknown $C^\star > 0$ and some *known* penalty shapes pen_0 and pen_1. Remark that we could generalize this assumption to $\mathrm{pen}_{\min}^{\mathrm{gal}} \approx C^\star\,\mathrm{pen}_0$ and $\mathrm{pen}_{\mathrm{opt}}^{\mathrm{gal}} \approx f(C^\star)\,\mathrm{pen}_1$ for some known function f, but such a generalization has not been proved useful yet.

Remark 2. *When $\widehat{s}_m \in \operatorname{argmin}_{t\in S_m} \widehat{\mathscr{R}}_n(t)$ is an empirical risk minimizer over some model $S_m \subset \mathbb{S}$, a natural choice is $s_m^\star \in \operatorname{argmin}_{t\in S_m} \mathscr{R}(t)$, so that $\mathscr{R}(s_m^\star) - \inf_{t\in\mathbb{S}} \mathscr{R}(t)$ is the approximation error. For linear estimators, the decomposition (46) of the risk suggests to take $s_m^\star = A_m F$. In the general case, choosing $s_m^\star$ might be more difficult. By analogy, we call $\mathscr{R}(s_m^\star) - \inf_{t\in\mathbb{S}} \mathscr{R}(t)$ the approximation error associated with $\widehat{s}_m$ in the general case.*

3.7. *Historical remarks*

Algorithms The slope-heuristics algorithm for calibrating penalties was first proposed in the Gaussian least-squares regression setting of Section 2 with a penalty proportional to the dimension [Birgé and Massart, 2001], as in Algorithms 1–2. Then, it was generalized to a penalty function of the dimension [Birgé and Massart, 2007], and to a general penalty shape [Massart, 2005, Blanchard and Massart, 2006, Massart, 2007], as in Algorithm 3.

The first implementations of the slope heuristics were done directly with Algorithm 3 (or its "slope" version) with $\mathscr{C}_m = D_m$, instead of Algorithms 1–2, since they were outside the setting of Section 2: maximum-likelihood estimators [Letué, 2000, Section A.4], and change-point detection [Lebarbier, 2002, Chapter 4].

The proposition of using a general complexity measure $\mathscr{C}_m$ instead of a dimension D_m (as in Algorithms 3, 5–6) was first made in density estimation [Lerasle, 2009], with the suggestion of estimating $\mathscr{C}_m$ by resampling if necessary.

The failure of Algorithm 3 for linear estimators in regression was noticed by [Arlot and Bach, 2009], where Algorithm 4 was proposed and theoretically justified. The general Algorithm 5 has only been formalized by [Arlot, 2011, Section 2.5], while its "slope estimation" version (Algorithm 6) is new, even if the equivalence between "jump" and "slope" algorithms is not. Up to now, the general formulation of Algorithms 5–6 has only been proved useful in the case of linear estimators in regression [Arlot and Bach, 2009, Arlot and Bach, 2011] and in density estimation [Magalhães, 2015, Lerasle et al., 2016], with different shapes for pen_0 and pen_1. It

can also be useful in a few other settings where pen_1 is proportional to pen_0 but the ratio between optimal and minimal penalty might be different from 2, for selecting among a rich collection of models or estimators. For instance, for pruning a decision tree, [Bar-Hen et al., 2018] uses a "maximal plateau method" that is equivalent to Algorithm 5 with $\mathrm{pen}_1 = \mathrm{pen}_0$, hence selecting the estimator "just after" the maximal jump. The name comes from the fact that exchanging the x-axis and the y-axis in Figure 2 left transforms jumps into plateaus.

Theory The first (partial) theoretical result proved outside the setting of Section 2 was for maximum-likelihood estimators (histograms) in density estimation, assuming that the true density $s^\star$ is the uniform density over $[0,1]$ [Castellan, 1999]. Other theoretical results outside the setting of Section 2 are reviewed in Section 4.

The first theoretical result proved for Algorithm 3 with a penalty shape $\mathrm{pen}_1(m)$ *not* function of a dimension D_m was obtained in heteroscedastic least-squares regression [Arlot and Massart, 2009], where the penalty shape can be estimated by resampling [Arlot, 2009].

The general heuristics "$\mathrm{pen}_{\mathrm{opt}} \approx 2\,\mathrm{pen}_{\mathrm{min}}$" underlying Algorithm 3 was formulated by [Blanchard and Massart, 2006, Section 2] and [Massart, 2007, Section 8.5.2], together with a heuristic argument for suggesting

$$p_2(m) := \widehat{\mathscr{R}}_n(s_m^\star) - \widehat{\mathscr{R}}_n(\widehat{s}_m)$$

as a minimal penalty, when $\widehat{s}_m$ is an empirical risk minimizer and $s_m^\star$ is defined according to Remark 2. In these papers, $p_2(m)$ is called $\widehat{v}_m$ since it can be interpretated as a variance. Here, the general minimal penalty that we propose is $\mathrm{pen}_{\mathrm{min}}^{\mathrm{gal}}(m) = \mathbb{E}[p_2(m)]$, as in [Arlot, 2007, Chapter 3] for instance. Another formulation of the heuristics behind Algorithm 3 is "$p_1(m) \approx p_2(m)$", where

$$p_1(m) := \mathscr{R}(\widehat{s}_m) - \mathscr{R}(s_m^\star)\,,$$

as for instance written in a binary classification framework by [Zwald, 2005, Section 6.4.3], together with "$p_2(m) \propto D_m$ for D_m large enough".

4. Theoretical results in the literature

This section collects all theoretical results that are directly related to minimal-penalty algorithms, to the best of our knowledge. First, the proof of Algorithm 5 is split into several subproblems (Section 4.1). Then, we present full proofs of Algorithm 5 (Section 4.2) and partial results (Sections 4.3–4.7). Note that some related results outside the setting of Section 3 are reported in conclusion (Sections 8.3.6 and 8.5).

In this section, all partial or full proofs of Algorithm 5 that we present define $\widehat{C}_{\mathrm{jump}}$ as $\widehat{C}_{\mathrm{thr.}}$ or $\widehat{C}_{\mathrm{window}}$ for some well-chosen T_n or η. For the sake of simplicity, we do not discuss anymore the exact definition chosen for $\widehat{C}_{\mathrm{jump}}$, until we tackle this question in Section 7.1.

4.1. General approach for proving Algorithm 5

Following Theorem 1 and its proof, let us suggest a general approach towards a theoretical justification of Algorithm 5, that we split into several subproblems.

(α) **The minimal and optimal penalty are known up to some common multiplicative factor:** Guess pen_0, pen_1, and some complexity measure $(\mathscr{C}_m)_{m\in\mathscr{M}}$, such that for some (unknown) $C^\star > 0$, $C^\star\,\mathrm{pen}_0$ is a minimal penalty and $C^\star\,\mathrm{pen}_1$ is an optimal penalty.

(β) $C^\star\,\mathrm{pen}_0$ **is actually a minimal penalty:** $\eta_n^-, \eta_n^+ > 0$ exist such that, on a large-probability event,

$$\forall C < \left(1-\eta_n^-\right) C^\star, \qquad \mathscr{C}_{\widehat{m}^{(0)}_{\min}(C)} \geqslant \mathscr{C}_{\mathrm{overfit}} \propto \max_{m\in\mathscr{M}} \mathscr{C}_m \tag{β^-}$$

$$\forall C > \left(1+\eta_n^+\right) C^\star, \qquad \mathscr{C}_{\widehat{m}^{(0)}_{\min}(C)} \leqslant \mathscr{C}_{\mathrm{small}} \ll \max_{m\in\mathscr{M}} \mathscr{C}_m \tag{β^+}$$

$$\text{where} \qquad \forall C \geqslant 0, \quad \widehat{m}^{(0)}_{\min}(C) \in \operatorname*{argmin}_{m\in\mathscr{M}} \left\{ \widehat{\mathscr{R}}_n(\widehat{s}_m) + C\,\mathrm{pen}_0(m) \right\}.$$

The above statements about $\mathscr{C}_{\widehat{m}^{(0)}_{\min}(C)}$ are vague on purpose, since the range of $(\mathscr{C}_m)_{m\in\mathscr{M}}$ are not specified. When $\mathscr{C}_m$ is the dimension D_m of some model, one can specify $\mathscr{C}_{\mathrm{overfit}}$ and $\mathscr{C}_{\mathrm{small}}$ similarly to Eq. (14) and (16), respectively.

(γ) $C^\star\,\mathrm{pen}_1$ **is actually an optimal penalty:** on a large-probability event, for every $C \in ((1-\eta)C^\star, (1+\eta)C^\star)$ with $\eta > 0$ small enough,

$$\forall \widehat{m}^{(1)}_{\mathrm{opt}}(C) \in \operatorname*{argmin}_{m\in\mathscr{M}} \left\{ \widehat{\mathscr{R}}_n(\widehat{s}_m) + C\,\mathrm{pen}_1(m) \right\},$$

$$\mathscr{R}\left(\widehat{s}_{\widehat{m}^{(1)}_{\mathrm{opt}}(C)}\right) - \mathscr{R}(s^\star) \leqslant \left(1+\varepsilon_n(\eta)\right) \inf_{m\in\mathscr{M}} \left\{ \mathscr{R}(\widehat{s}_m) - \mathscr{R}(s^\star) \right\} + R_n(\eta) \tag{γ}$$

where $\lim_{\eta\to 0, n\to+\infty} \varepsilon_n(\eta) = 0$ and $R_n(\eta)$ is negligible in front of the oracle risk $\inf_{m\in\mathscr{M}}\{\mathscr{R}(\widehat{s}_m) - \mathscr{R}(s^\star)\}$.

As in [Arlot, 2007], we use in this section the following notation:

$$p_1(m) := \mathscr{R}(\widehat{s}_m) - \mathscr{R}(s_m^\star) \tag{55}$$

$$p_2(m) := \widehat{\mathscr{R}}_n(s_m^\star) - \widehat{\mathscr{R}}_n(\widehat{s}_m) \tag{56}$$

$$\delta(m) := \mathscr{R}(s_m^\star) - \widehat{\mathscr{R}}_n(s_m^\star). \tag{57}$$

In particular, with the notation of Section 3.6,

$$\mathrm{pen}^{\mathrm{gal}}_{\mathrm{opt}}(m) = \mathbb{E}\left[p_1(m) + \delta(m) + p_2(m)\right] \qquad \text{and} \qquad \mathrm{pen}^{\mathrm{gal}}_{\min}(m) = \mathbb{E}\left[p_2(m)\right].$$

4.2. Full proofs of Algorithm 5

Few settings exist where a full proof of Algorithm 5 is available, that is, a proof that (β^-), (β^+), and (γ) hold true on a large-probability event for some known pen_0, pen_1, $(\mathscr{C}_m)_{m\in\mathscr{M}}$ and some (unknown) $C^\star$. In this paper, $\mathscr{M}$ is always assumed to be finite with $\mathrm{card}(\mathscr{M}) \leqslant L_1 n^{L_2}$ for some $L_1, L_2 > 0$, except in Section 4.7.

We first collect results assuming that $\mathrm{pen}_1 = 2\,\mathrm{pen}_0$, so that Algorithm 5 reduces to Algorithm 3. Without explicit mention of the contrary, for all results reviewed in the list below, the noise is assumed independent and identically distributed, $\widehat{s}_m \in \operatorname{argmin}_{t\in S_m} \widehat{\mathscr{R}}_n(t)$ is an empirical risk minimizer, so we take $s_m^\star \in \operatorname{argmin}_{t\in S_m} \mathscr{R}(t)$ for defining $p_2(m)$, and the complexity used is $\mathscr{C}_m = D_m$ the dimension of S_m. Full proofs of Algorithm 5 exist in the following settings:

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

— *Regression on a fixed design, homoscedastic (sub-)Gaussian noise, least-squares risk and estimators*: [Birgé and Massart, 2007] or Theorem 1 (and Remark 1 for the sub-Gaussian case), with $\mathrm{pen}_0(m) = D_m/n$ and $C^\star = \sigma^2$ the (constant) noise-level. Note that $p_1(m) = p_2(m) = n^{-1}\|\Pi_m\varepsilon\|^2$ in this setting.

— Regression on a *random design*, *heteroscedastic noise* (not necessarily Gaussian), least-squares risk, with various least-squares estimators: regressograms with moment assumptions on the noise [Arlot and Massart, 2009], piecewise polynomials with bounded noise [Saumard, 2013] (with key concentration results for p_1 and p_2 proved in [Saumard, 2012]), or more general models satisfying a "strongly-localized basis" assumption with bounded noise [Saumard, 2010a, Navarro and Saumard, 2017]. Contrary to the previous setting, $\mathbb{E}[p_1(m)] \approx \mathbb{E}[p_2(m)]$ holds true only for most models and for n large enough.
The penalty shape $\mathrm{pen}_0(m) = \mathbb{E}[p_2(m)]$ is unknown in general and $C^\star = 1$. For regressograms, the results remain true when $\mathrm{pen}_0(m)$ is a resampling-based estimation of $\mathbb{E}[p_2(m)]$, see [Arlot, 2009]. For piecewise polynomials, the same holds when $\mathrm{pen}_0(m)$ is a hold-out estimation of $\mathbb{E}[p_2(m)]$, see [Saumard, 2013]. For strongly localized bases, the (approximate) closed-form formula for $\mathbb{E}[p_1(m)]$ and $\mathbb{E}[p_2(m)]$ provided by [Navarro and Saumard, 2017, Theorem 6.3] might be used for estimating $\mathrm{pen}_0(m)$ without resampling; another option is V-fold penalization [Navarro and Saumard, 2017, Section 5].

— *Density estimation*, least-squares risk and estimators: i.i.d. [Lerasle, 2012] or *mixing data* [Lerasle, 2011]. The penalty shape $\mathrm{pen}_0(m) = \mathbb{E}[p_2(m)]$ is approximately known for some specific models (regular histograms), in general it can be estimated by resampling as previously. In this setting, the complexity $\mathscr{C}_m$ can either be the dimension of S_m or the resampling-based estimator of $\mathbb{E}[p_2(m)]$ itself. Note that in least-squares density estimation, $p_1(m) = p_2(m)$ almost surely.

— Density estimation, *Kullback risk and maximum-likelihood estimators*, histogram models [Saumard, 2010c]. This result is the first one obtained without the least-squares risk. The penalty shape $\mathrm{pen}_0(m) = D_m/(2n)$ is known, $C^\star = 1$, and the optimal penalty is AIC. A partial result, for the uniform density over $[0,1]$ only, has previously been proved by [Castellan, 1999].

— *Specification probabilities in general random fields* (that is, graphical models), least-squares or Kullback risks, estimators that are empirical distributions conditionally to the values observed on a subset m of the field [Lerasle and Takahashi, 2016]. The shape of the penalty and the complexity $\mathrm{pen}_0(m) = \mathscr{C}_m = p_2(m)$ are unknown. The authors suggest to use instead the shape of a theoretical upper bound on $\mathbb{E}[p_2(m)]$, dropping off pessimistic constants, with convincing experimental results.

The above results for least-squares regression on a random design, least-squares density estimation (i.i.d. case), and maximum-likelihood density estimation can all be recovered (sometimes up to minor differences) as a corollary of a general result which holds for all "regular estimators" [Saumard, 2010b, Chapters 7–8].

Full proofs of Algorithm 5 (or a slight modification of it) also exist in two settings where $\mathrm{pen}_1 \neq 2\,\mathrm{pen}_0$ in general:

— Regression on a fixed design, independent and identically distributed (homoscedastic) Gaussian noise, least-squares risk, *linear estimators*: [Arlot and Bach, 2009, Arlot and

Bach, 2011] prove that Algorithm 4 works, while Algorithm 3 fails in general, as detailed in Sections 3.2–3.3.

— Density estimation, independent and identically distributed data, least-squares risk, *linear estimators* (for instance, Parzen density estimators and weighted least-squares estimators): [Lerasle et al., 2016] —and its preliminary version in [Magalhães, 2015, Chapter 2]— define some theoretical quantities $\mathrm{pen}_0(m) \approx \mathbb{E}[p_2(m)]$ and $\mathscr{C}_m \approx \mathbb{E}[p_1(m)]$ —easy to estimate in general, and known for several examples such as Parzen density estimators— such that

$$\widehat{m}(C) \in \operatorname*{argmin}_{m\in\mathscr{M}} \left\{ \widehat{\mathscr{R}}_n(\widehat{s}_m) + \mathrm{pen}_0(m) + C\mathscr{C}_m \right\}$$

overfits for $C < 0$ and satisfies an oracle inequality for all $C > 0$, first-order optimal when $C = C^\star = 1$. In other words, this almost proves that Algorithm 5 works with $\mathrm{pen}_0(m) = \mathbb{E}[p_2(m)]$, $\mathrm{pen}_1(m) = \mathbb{E}[p_1(m)] + \mathbb{E}[p_2(m)]$, $\mathscr{C}_m = \mathbb{E}[p_1(m)]$, and $C^\star = 1$. This result implies the one of [Lerasle, 2012] for least-squares estimators. A noticeable fact in this framework is that $\mathrm{pen}_{\min}$ —and sometimes even $\mathrm{pen}_{\mathrm{opt}}$!— can be negative, making the terminology "minimal" penalty questionable [Lerasle et al., 2016, sections 4.3 and 5] . Nevertheless, for most usual estimators, $\mathrm{pen}_{\min}$ and $\mathrm{pen}_{\mathrm{opt}}$ are always positive. Theoretical results for choosing among Parzen density estimators with slightly different minimal-penalty algorithms —closer to Goldenshluger-Lepski's method— are discussed in Section 8.5.

Remark 3 (Minimal penalties with $p_2(m)$)**.** *In several papers mentioned above, theoretical results validate Algorithm 5 with* $\mathrm{pen}_0(m) = \mathscr{C}_m = p_2(m)$ *or* $\mathbb{E}[p_2(m)]$, $\mathrm{pen}_1(m) = 2\,\mathrm{pen}_0(m)$, *and* $C^\star = 1$. *Such results might seem useless since (i)* pen_0 *is unknown, and (ii)* $C^\star$ *is known, that is, the exact opposite of our initial motivation. Nevertheless, (i) can be solved by using a resampling-based estimator* $\widehat{p}_2^W(m)$ *of* $\mathbb{E}[p_2(m)]$ *for* $\mathrm{pen}_0(m)$ *and* $\mathscr{C}_m$, *see [Lerasle, 2012] for instance. Then, a classical problem of resampling is to find the constant* C_W *such that* $C_W\mathbb{E}\left[\widehat{p}_2^W(m)\right] \approx \mathbb{E}[p_2(m)]$ *for all* $m \in \mathscr{M}$. *When it exists,* C_W *usually depends on the resampling scheme* W, *the sample size, and the particular setting considered [Arlot, 2009]. As a consequence, we recover a setting where* $\mathrm{pen}_0(m) = \mathscr{C}_m = \widehat{p}_2^W(m)$ *is known and* $C^\star = C_W$ *is unknown, for which Algorithm 5 can be useful. Note that taking* $\mathscr{C}_m = \widehat{p}_2^W(m)$ *instead of* $C_W\widehat{p}_2^W(m)$ *doesn't matter since the complexity jump is independent from the rescaling by* C_W.
According to simulation experiments, combining resampling penalties with the slope heuristics can be useful —in the least-squares density estimation framework [Lerasle, 2010]— or not —for context-tree estimation [Garivier and Lerasle, 2011]—, compared to resampling penalties multiplied by a deterministic constant C *that derives from asymptotic theoretical results and does not depend on any unknown quantity in these two settings.*

4.3. Partial proofs: uncertainty on the optimal penalty

An optimal oracle inequality like (γ) has not been proved in many frameworks, and it is quite difficult to obtain a leading constant $1 + \varepsilon_n(\eta) = 1 + \mathrm{o}(1)$ while keeping the remainder term negligible in front of the oracle risk. A much more usual result in the model-selection literature is the following weakened version of (γ): on a large-probability event, for every $C \in$

$((1-\eta)C^\star, (1+\eta)C^\star)$ with $\eta > 0$ small enough,

$$\mathscr{R}\left(\widehat{s}_{\widehat{m}^{(1)}_{\mathrm{opt}}(C)}\right) - \mathscr{R}(s^\star) \leqslant K_n(\eta) \inf_{m\in\mathscr{M}} \left\{\mathscr{R}(\widehat{s}_m) - \mathscr{R}(s^\star)\right\} + R_n(\eta) \qquad (\widetilde{\gamma})$$

for some $K_n(\eta), R_n(\eta) < \infty$. Note that (γ) with $\varepsilon_n = 0$ and $R_n \geqslant \inf_{m\in\mathscr{M}}\{\mathscr{R}(\widehat{s}_m) - \mathscr{R}(s^\star)\}$ should be understood as $(\widetilde{\gamma})$ with $K_n \geqslant 2$. Similarly, a classical way to write an oracle-type inequality is

$$\mathscr{R}\left(\widehat{s}_{\widehat{m}^{(1)}_{\mathrm{opt}}(C)}\right) - \mathscr{R}(s^\star) \leqslant \inf_{m\in\mathscr{M}} \left\{\mathscr{R}(\widehat{s}_m) - \mathscr{R}(s^\star) + R_n(m)\right\}. \qquad (\widetilde{\gamma}')$$

When $R_n(m)$ is comparable to $\mathscr{R}(\widehat{s}_m) - \mathscr{R}(s^\star)$, or even larger, (for instance, $R_n(m) \geqslant \mathrm{pen}(m)$), $(\widetilde{\gamma}')$ should be understood as $(\widetilde{\gamma})$ with $K_n \geqslant 2$.

Proving only $(\widetilde{\gamma})$ instead of (γ) is a significant limitation: $(\widetilde{\gamma})$ does not show that $C^\star \mathrm{pen}_1$ is an optimal penalty if we cannot prove that $K_n(\eta)$ is first-order optimal, which is very difficult to prove unless $K_n(\eta) = 1 + \mathrm{o}(1)$ as in (γ). As a consequence, in such cases, $C^\star \mathrm{pen}_1$ might not be optimal, and the optimal penalty might be $C' \mathrm{pen}_1$ with $C^\star \neq C'$, or even have a completely different shape than pen_1. For instance, in the setting of Section 3.4, the optimal penalty is $2\sigma^2 \mathrm{tr}(A_m)/n$, but taking $2\sigma^2(\mathrm{tr}(A_m) + \mathrm{tr}(A_m^\top A_m))/n$ as a penalty, we could have an oracle inequality $(\widetilde{\gamma})$ with a penalty having a suboptimal shape.

Results Nevertheless, proving (β) and $(\widetilde{\gamma})$ still shows that Algorithm 5 provides a data-driven estimator satisfying an oracle inequality. Such a result exist for context tree estimation with the Kullback risk, ϕ-mixing processes, and maximum-likelihood estimators [Garivier and Lerasle, 2011], with $\mathrm{pen}_0(m) = \mathscr{C}_m = p_2(m)$ and $\mathrm{pen}_1(m) = 2\mathrm{pen}_0(m)$. Simulation experiments suggest that $p_2(m)$ can be replaced by a BIC-type penalty or a resampling-based estimator of $\mathbb{E}[p_2(m)]$, see Remark 3. What is missing to get a proof of (γ) is a tight concentration inequality for $\delta(m) - \delta(m')$, that is, to have Eq. (60) satisfied with $\varepsilon_\delta = \mathrm{o}(1)$ as required in Proposition 2, see Section 5.2.2. Simulation experiments suggest that $\mathrm{pen}_1 = 2\mathrm{pen}_0$ is indeed an optimal choice here.

4.4. Minimal penalty in terms of risk: (β')

Another way to define a minimal penalty is in terms of the risk of $\widehat{s}_{\widehat{m}^{(0)}_{\min}(C)}$, which is theoretically interesting but does not prove the presence of a complexity jump as expected by Algorithm 5:

$$\forall C < (1-\eta_n^-)C^\star, \mathscr{R}\left(\widehat{s}_{\widehat{m}^{(0)}_{\min}(C)}\right) - \mathscr{R}(s^\star) \geqslant \kappa \max_{m\in\mathscr{M}} \left\{\mathscr{R}(\widehat{s}_m) - \mathscr{R}(s^\star)\right\} \qquad (\beta'^-)$$

$$\forall C > (1+\eta_n^+)C^\star, \mathscr{R}\left(\widehat{s}_{\widehat{m}^{(0)}_{\min}(C)}\right) - \mathscr{R}(s^\star) \leqslant K\left(\frac{C}{C^\star}\right) \inf_{m\in\mathscr{M}} \left\{\mathscr{R}(\widehat{s}_m) - \mathscr{R}(s^\star)\right\} + R_n\left(\frac{C}{C^\star}\right) \qquad (\beta'^+)$$

where $\kappa > 0$ is an absolute constant, and for every $x > 1$, $K(x) \in [1,\infty)$, and generally $R_n(x) \ll \inf_{m\in\mathscr{M}}\{\mathscr{R}(\widehat{s}_m) - \mathscr{R}(s^\star)\}$.

When (β) is replaced by (β'), the justification of Algorithm 5 is far from being complete, since there might be no complexity jump as required in the definition of $\widehat{C}_{\mathrm{window}}$. Nevertheless,

once (β') is proved, one can reasonably conjecture that (β) holds true under similar assumptions, provided that $\mathscr{C}_m$ is well chosen. Moreover, (β) and (β') are closely related if

$$\forall x > 0\,, \qquad \inf_{m\in\mathscr{M}\,/\,\mathscr{C}_m\geqslant x} \left\{\mathscr{R}(\widehat{s}_m) - \mathscr{R}(s^\star)\right\} \geqslant g(x) > 0 \tag{58}$$

for some increasing function g. Indeed, (β^-) implies (β'^-) with

$$\kappa = \frac{g(\mathscr{C}_{\text{overfit}})}{\max_{m\in\mathscr{M}}\left\{\mathscr{R}(\widehat{s}_m) - \mathscr{R}(s^\star)\right\}}\,,$$

and (β'^+) implies (β^+) with

$$\mathscr{C}_{\text{small}} = g^{-1}\left(K\left(\frac{C}{C^\star}\right)\inf_{m\in\mathscr{M}}\left\{\mathscr{R}(\widehat{s}_m) - \mathscr{R}(s^\star)\right\}\right) + R_n\left(\frac{C}{C^\star}\right).$$

Note that Eq. (58) holds true with $g(x) = x/\alpha$ if $\mathscr{C}_m \approx \alpha p_1(m)$; for instance, for least-squares estimators and risk, $\alpha = \sigma^2/n$ since $\mathscr{C}_m = D_m$ and $p_1(m) \approx \sigma^2 D_m/n$. Let us remark finally that the proof of (β'^-) usually relies on a proof of (β^-), sometimes hidden by technical details.

Results To the best of our knowledge, a full proof of (β') currently exists only in settings where (β) is proved to hold, except one result that we report in Section 4.7. Some partial proofs of (β') are reviewed in the next subsections.

4.5. Partial proofs: uncertainty on the minimal penalty

A weaker result than (β) can be proved about the complexity jump: for some $C_1^\star < C_2^\star$ (that remain distinct even when $n \to +\infty$),

$$\forall C < C_1^\star\,, \quad (\beta^-) \text{ holds true,} \qquad \text{and} \qquad \forall C > C_2^\star\,, \quad (\beta^+) \text{ holds true.} \tag{$\widetilde{\beta}$}$$

In other words, $C_1^\star\,\text{pen}_0$ is a too small penalty, while $C_2^\star\,\text{pen}_0$ is sufficiently large.

From the theoretical point of view, proving $(\widetilde{\beta})$ instead of (β) is a serious limitation: for reasons similar to the ones explained in Section 4.3 for $(\widetilde{\gamma})$, it can happen that pen_0 is not the shape of a minimal penalty. For instance, in the setting of Section 3.3, $(\widetilde{\beta})$ holds true with $\text{pen}_0(m) = \text{tr}(A_m)$ although this quantity is not always proportional to the minimal penalty, as shown by Figure 3.

Nevertheless, from the practical point of view, one can still derive from $(\widetilde{\beta})$ a way to get from data some $\widehat{C} \in [C_1^\star, C_2^\star]$, for instance by taking a large η in the definition of $\widehat{C}_{\text{window}}$. If $(C_2^\star/C_1^\star)$ is not too large and if $(\widetilde{\gamma})$ holds true for some $C^\star \in [C_1^\star, C_2^\star]$, this leads to an estimator satisfying an oracle inequality.

Results One full proof of $(\widetilde{\beta})$ and $(\widetilde{\gamma})$ is available for prediction in a Gaussian graphical model via neighborhood selection, with conditional least-squares risk and estimators, $\text{pen}_0(m) = \mathscr{C}_m = D_m$, and $\text{pen}_1 \propto \text{pen}_0$ [Verzelen, 2010]; the proof of $(\widetilde{\beta})$ assumes in addition that the graph is a square lattice. Simulation experiments suggest that there is indeed a jump around $C^\star$ and that Algorithm 5 works well.

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

4.6. *Partial proofs: for some specific $s^\star$ only*

The weakest partial proofs of (β) are the ones only valid for some particular $s^\star$, often $s^\star = 0$. Then, although $C^\star \operatorname{pen}_0$ is a minimal penalty for this particular $s^\star$, the general shape of the minimal penalty can differ from pen_0. For instance, in the Lasso case, an empirical study shows that the shape of $\mathbb{E}[p_2(m)]$ depends on $s^\star$ and on some other features of the distribution of the data [Connault, 2011].

Nevertheless, such weak results still are a good way to guess pen_0 for a practical use of Algorithm 5, and they can be a first step towards a full theoretical justification.

Results Such partial proofs exist in the case of multiplicative penalties, an apparently different setting that can still be cast into the framework of Algorithm 5. The principle, as exposed in [Baraud et al., 2009] for least-squares regression, is to replace the penalized criterion (3) by the *product* of the empirical risk by some penalty term, that is, choosing

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}_n(\widehat{s}_m) \left(1 + \frac{\operatorname{pen}^{\text{mult}}(m)}{n - D_m} \right) \right\} . \tag{59}$$

This can actually be seen as an additive penalization method as in Eq. (3), with the penalty

$$\operatorname{pen}(m) := \widehat{\mathcal{R}}_n(\widehat{s}_m) \frac{\operatorname{pen}^{\text{mult}}(m)}{n - D_m} .$$

So, choosing a multiplicative factor in front of $\operatorname{pen}^{\text{mult}}(m)$ is equivalent to choosing a multiplicative factor in front of an additive penalty of a particular form. For fixed-design regression with least-squares risk and estimators, [Baraud et al., 2009] proves that (β^-) holds true if $s^\star = 0$, while (β'^+) and $(\widetilde{\gamma})$ hold true in general, with $C^\star \operatorname{pen}_0^{\text{mult}}(m) = D_m$ and $\operatorname{pen}_1^{\text{mult}} \propto \operatorname{pen}_0^{\text{mult}}$.

In addition, in the setting of multivariate regression on a fixed design with the least-squares risk and low-rank least-squares estimators, $(\widetilde{\gamma})$ and (β'^+) are proved in a general case, while (β^-) is proved only for $s^\star = 0$ [Giraud, 2011]; remark that (β^+) can certainly be proved in a general case, although its proof is not written in [Giraud, 2011]. Note also that these results are valid both for additive penalties and for multiplicative penalties as in [Baraud et al., 2009].

4.7. *Partial proofs: richer collections of models*

Throughout the paper, we assume (at least implicitly) that $\mathcal{M}$ is not too large, that is, $\operatorname{card}(\mathcal{M})$ grows at most polynomially with the sample size n, or $\mathcal{M}$ can be well approximated by such a polynomial set of estimators —e.g., kernel ridge regression with one continuous parameter λ [Arlot and Bach, 2011]. Nevertheless, the case where $\mathcal{M}$ is larger deserves attention, and we review in this subsection the partial results about minimal penalties in such settings. Note that each of them suffers from some of the limitations emphasized in Sections 4.3–4.6.

Let us consider the fixed-design regression setting, with least-squares risk and estimators on finite-dimensional vector spaces S_m. Assuming as in [Birgé and Massart, 2007] that the penalty is a function of the dimension, the selected estimator

$$\widehat{s}_{\widehat{m}} \qquad \text{with} \qquad \widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}_n(\widehat{s}_m) + \operatorname{pen}(D_m) \right\}$$

can be rewritten as $\widehat{s}'_{\widehat{D}}$ where

$$\forall D \in \mathbb{N}, \quad \widehat{s}'_D \in \operatorname*{argmin}_{t \in S'_D} \left\{ \widehat{\mathscr{R}}_n(t) \right\}, \quad S'_D := \bigcup_{\substack{m \in \mathscr{M} \\ D_m = D}} S_m, \quad \text{and} \quad \widehat{D} \in \operatorname*{argmin}_{D \in \mathbb{N}} \left\{ \widehat{\mathscr{R}}_n(\widehat{s}'_D) + \operatorname{pen}(D) \right\}.$$

Then, discarding all models of dimension $D > n$, $\widehat{s}_{\widehat{m}} = \widehat{s}'_{\widehat{D}}$ is a penalized empirical risk minimizer over a collection $(S'_D)_{0 \leqslant D \leqslant n}$ of cardinality at most $n+1$. The difference with the initial formulation is that the models S'_D are not vector spaces (in general), and the complexity of S'_D strongly depends on $f(D) := \operatorname{card}\{m \in \mathscr{M} \,/\, D_m = D\}$. Three cases can be distinguished, following [Birgé and Massart, 2007]:

(i) $\mathscr{M}$ is "small" or "polynomial" when $f(D) \leqslant CD^{\omega}$ for some $C, \omega > 0$. Then, $\operatorname{card}(\mathscr{M})$ grows polynomially with n (since models of dimension $D_m > n$ can safely be discarded), and the complexity of S'_D is essentially the same as the one of a D-dimensional vector space.

(ii) $\mathscr{M}$ is "large" or "exponential" when $f(D)$ grows much faster —typically of order $\binom{n}{D}$—, which implies in particular that $\operatorname{card}(\mathscr{M})$ grows exponentially with n. Then, S'_D is much more complex than a D-dimensional vector space. A typical example is (full) variable selection among $p \geqslant n$ variables, for which $f(D) = \binom{p}{D}$.

(iii) $\mathscr{M}$ is "moderate" in the intermediate situation, when $D^{-1} \log f(D)$ stays bounded away from 0 and ∞ for $n \gg D \gg 1$.

The current subsection focuses on cases (ii) and (iii); all other results mentioned in this article correspond to case (i).

Results for case (ii): large number of models

In fixed-design regression with least-squares risks and estimators, (β') and ($\widetilde{\gamma}$) are proved by [Birgé and Massart, 2007] for the (full) variable-selection problem with orthonormal variables, assuming that the noise is Gaussian, and defining

$$\operatorname{pen}_0(m) = \frac{D_m}{n}\left[1 + 2\log\left(\frac{p}{D_m}\right)\right], \qquad C^\star = \sigma^2,$$

and $\operatorname{pen}_1(m) = C \operatorname{pen}_0$ with any $C > 1$. Note that (β^-) can be derived from the proof of [Birgé and Massart, 2007, Proposition 2], but it is not written in [Birgé and Massart, 2007].

Similar theoretical results are proved by [Sorba, 2017, Chapter 8] for Gaussian variable selection with $p = n$ and a more general collection of models, that can be smaller than full variable selection but still exponentially large. Formally, [Sorba, 2017, Section 8.1] assumes that $\mathscr{M}$ satisfies a "completion rule", which holds for instance for the collection of regressograms over a partition whose cells are hyperrectangles of $\mathbb{R}^d$. Then, (β'^-) and ($\widetilde{\gamma}$) hold true, with $\operatorname{pen}_0(m)$ proportional to $\frac{D_m}{n}\log(\frac{en}{D_m})$ —showing that a $\log(n)$ factor is still necessary here— and (β^-) is proved when the target signal is null. Compared to [Birgé and Massart, 2007], a gap of a multiplicative constant remains between minimal and sufficient penalties.

[Sorba, 2017, Chapter 9] proves similar theoretical results about histogram selection for density estimation by penalized log-likelihood, with the Kullback risk, for any large collection of

subpartitions of a regular partition of $[0,1]$ into $N+1$ pieces, assuming $N \leqslant n/(\log(n))^2$. A sufficient penalty $\propto \frac{D_m}{n}\log(N)$ satisfies $(\widetilde{\gamma})$. If the target density is uniform over $[0,1]$, (β'^{-}) holds true with a minimal penalty level of the same order of magnitude. As a consequence, when $\log(N) \sim \log(n)$, this proves that a $\log(n)$ factor must be added to the penalty compared to the case of a polynomial collection $\mathscr{M}$.

Two partial results are available with multiplicative penalties, which are introduced in Section 4.6. In the same setting as [Birgé and Massart, 2007], (β^{-}) —assuming $s^\star = 0$ and a specific "exponential" collection $\mathscr{M}$ with $f(D) \approx \binom{n}{D}$—, (β'^{+}), and $(\widetilde{\gamma})$ are proved by [Baraud et al., 2009], with $\mathrm{pen}_0^{\mathrm{mult}}(m) = 2D_m\log(n)$. For estimation of a Gaussian graph —that is, in a Gaussian graphical model, predict the value at each vertex of the graph given its neighbors, by linear regression—, with least-squares risk and estimators, (β^{-}) —assuming $s^\star = 0$ and $\mathscr{M}$ contains some specific "exponential" collection with $f(D) \approx \binom{p}{D}$ for some $p > n$—, (β'^{+}), and $(\widetilde{\gamma})$ are proved by [Giraud, 2008], with $\mathrm{pen}_0^{\mathrm{mult}}(m) = 2D_m\log(p)$. In both papers [Baraud et al., 2009, Giraud, 2008], $(\widetilde{\gamma})$ holds with $\mathrm{pen}_1^{\mathrm{mult}} = C\,\mathrm{pen}_0^{\mathrm{mult}}$ for any $C > 1$.

Finally, several other arguments can be found for the necessity of a penalty larger than

$$\sigma^2\frac{D_m}{n}\left[1+\log\left(\frac{n}{D_m}\right)\right]$$

—up to a numerical constant— for change-point detection, which is an instance of variable selection with $p = n-1$. Minimax lower bounds [Durot et al., 2009, Theorem 2] and general oracle inequalities [Birgé and Massart, 2007] prove that for the true model $m^\star$,

$$\mathrm{pen}(m^\star) \geqslant \kappa\sigma^2\frac{D_{m^\star}}{n}\left(1+\log\left(\frac{n}{D_{m^\star}}\right)\right)$$

is necessary for some constant $\kappa > 0$. [Abramovich et al., 2006, Section 1.9] provides several other reasons why the optimal penalty should be close to $2\sigma^2\frac{D_m}{n}\log(\frac{n}{D_m})$.

Results for case (iii): moderate number of models

In fixed-design regression with least-squares risks and estimators, $(\widetilde{\gamma})$ holds true in general, and $(\widetilde{\beta}')$ is proved assuming $s^\star = 0$ and that all models of the same dimension D are orthogonal [Birgé and Massart, 2007, Proposition 3], with

$$\mathrm{pen}_0(m) = \lambda\frac{D_m}{n}\left[1+2\sqrt{f(D_m)}+2f(D_m)\right], \qquad f(D) = a+\frac{b\log(D+1)}{D}, \qquad \text{and} \qquad C^\star = \sigma^2$$

under some condition on the constants λ, a, b. By $(\widetilde{\beta}')$, we mean $(\widetilde{\beta})$ with a jump in the risk instead of the complexity; here, the gap in $(\widetilde{\beta}')$ is $C_2^\star/C_1^\star = 6/5$.

Results of the same flavor can be found in [Sorba, 2017, Chapter 6] about a Gaussian linear process and a b-ary tree partition collection —with no assumption on $s^\star$ for proving (β^{-}) and (β'^{-})—, and in [Sorba, 2017, Chapter 10] for a toy problem close to [Birgé and Massart, 2007, Proposition 3]. All these results show that "intermediate" collections of models can require a penalty strictly larger than the minimal penalty $\frac{\sigma^2 D_m}{n}$ of "polynomial" collections.

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

5. Towards new theoretical results on minimal penalties

We now describe some strategies for proving that Algorithm 5 works in other settings. This section is a bit more abstract and technical than the rest of the paper, so it can be skipped at first reading. As in Section 4.1, whose notations are used throughout the section, we consider separately subproblems (α), (β), and (γ).

5.1. *Hints for (α): how to guess* pen_0, pen_1, *and* $\mathscr{C}_m$*?*

Using the notation defined by Eq. (55), (56), and (57), Section 3.6 suggests that $p_1(m)+p_2(m)$ or its expectation $\mathrm{pen}^{\mathrm{gal}}_{\mathrm{opt}}(m)$ should be an optimal penalty, and $p_2(m)$ or its expectation $\mathrm{pen}^{\mathrm{gal}}_{\mathrm{min}}(m)$ should be a minimal penalty.

In both cases, computing (approximately) $\mathbb{E}[p_i(m)]$, $i=1,2$, or deriving an asymptotic expansion of $p_i(m)$, $i=1,2$, at least for $\mathscr{C}_m$ large enough, can lead to formulas for $\mathrm{pen}_0(m)$ and $\mathrm{pen}_1(m)$. For instance, for fixed-design regression with the least-squares risk, exact formulas for $\mathbb{E}[p_i(m)]$ lead to Algorithm 1 for least-squares estimators, and to Algorithm 4 for linear estimators. The main difficulty here is to have no unknown quantity inside pen_0 or pen_1.

For general estimators in the fixed-design regression setting, an exact formula for $\mathrm{pen}^{\mathrm{gal}}_{\mathrm{opt}}(m)$ is given by covariance penalties [Efron, 2004], which can be expressed using the degrees of freedom when the noise is Gaussian and the loss is quadratic. For maximum-likelihood estimators and risk, a partial asymptotic solution is given by the formula of the AIC criterion [Akaike, 1973], which derives from some version of the Wilks phenomenon (see also Section 5.2). In both cases, only a formula for pen_1 is available, and pen_0 remains unknown, even if one can sometimes conjecture that $\mathrm{pen}_0 = \mathrm{pen}_1/2$.

For random-design regression with the quadratic risk, [Navarro and Saumard, 2017, Theorem 6.3] provides an (approximate) closed-form formula for $\mathbb{E}[p_1(m)]$ and $\mathbb{E}[p_2(m)]$ —by proving that $p_1(m)$ and $p_2(m)$ concentrate around some deterministic quantity, that is not necessarily equal to their expectation—, that is not directly useful because it depends on the unknown distribution of the (X_i,Y_i).

Another option is to define $\mathrm{pen}_0(m)$, resp. $\mathrm{pen}_1(m)$, as some resampling-based estimator of $\mathbb{E}[p_2(m)]$, resp. $\mathbb{E}[p_2(m)+p_1(m)]$, and to use Algorithm 5 for estimating the common (unknown) multiplicative factor $C^\star$ such that

$$C^\star\mathbb{E}\big[\mathrm{pen}_0(m)\big] \approx \mathbb{E}\big[p_2(m)\big] \qquad \text{and} \qquad C^\star\mathbb{E}\big[\mathrm{pen}_1(m)\big] \approx \mathbb{E}\big[p_1(m)+p_2(m)\big]\,,$$

see Remark 3. In addition to the papers mentioned in Section 4.2, let us mention here that a concentration result for the resampling estimate of $\mathbb{E}[p_2(m)]$ is proved by [Arlot, 2007, Chapter 7], for empirical risk minimizers and a general bounded risk.

If no natural quantity arises as a complexity measure $\mathscr{C}_m$, such as the number of parameters in regression, $\mathbb{E}\left[p_2(m)\right]$ (or a resampling-based estimator of it) can be a good guess for $\mathscr{C}_m$, see [Lerasle, 2012] and Remark 3.

5.2. *Hints for (β): how to prove that $C^\star\,\mathrm{pen}_0$ is a minimal penalty?*

Following the results mentioned in the previous subsections, two general approaches can be used for proving (β), assuming either that $C^\star\,\mathrm{pen}_0(m) = \mathbb{E}\left[p_2(m)\right]$ as in Theorem 1, or that $\mathrm{pen}_0(m) =$

$\mathscr{C}_m = p_2(m)$ as in [Lerasle and Takahashi, 2016]. This section details these two approaches for proving (β^-) and (β^+), before focusing on the concentration inequalities they both require. Recall that

$$\forall C \geqslant 0, \qquad \widehat{m}^{(0)}_{\min}(C) \in \operatorname*{argmin}_{m\in\mathscr{M}} \left\{ \widehat{\mathscr{R}}_n(\widehat{s}_m) + C\,\mathrm{pen}_0(m) \right\}.$$

5.2.1. Below the minimal penalty: (β^-)

Following the proof of Theorem 1 When $C^\star\,\mathrm{pen}_0(m) = \mathbb{E}[p_2(m)]$, similarly to the proof of Eq. (14) in Theorem 1, one can prove (β^-) by showing that for some well-chosen $m_1 \in \mathscr{M}$,

$$\forall C < C^\star, \qquad \inf_{m\in\mathscr{M}\,/\,\mathscr{C}_m < \mathscr{C}_{\text{overfit}}} G_C(m) > G_C(m_1)$$

$$\text{where} \quad G_C(m) := \widehat{\mathscr{R}}_n(\widehat{s}_m) + C\,\mathrm{pen}_0(m) = \mathscr{R}(s^\star_m) - \delta(m) - p_2(m) + \frac{C}{C^\star}\mathbb{E}\big[p_2(m)\big]$$

is the quantity minimized by $\widehat{m}^{(0)}_{\min}(C)$. Then, in addition to the arguments sketched in Section 3.5, we only need here tight concentration inequalities for $\delta(m) - \delta(m_1)$ and for $p_2(m)$, see Section 5.2.3. A natural choice for m_1 is a minimizer of the approximation error $\mathscr{R}(s^\star_m) - \mathscr{R}(s^\star)$ over $m \in \mathscr{M}$.

Generalizing the strategy of [Lerasle and Takahashi, 2016, Garivier and Lerasle, 2011] When $\mathrm{pen}_0(m) = \mathscr{C}_m \propto p_2(m)$, the approach used by [Lerasle and Takahashi, 2016, Garivier and Lerasle, 2011] can be summarized into the following proposition.

Proposition 1. *Let us consider the general framework of Section 3.1 and use the notation of Section 4.1. Let $\varepsilon_\delta \in [0,1]$, $\varepsilon'_\delta \geqslant 0$, and assume that for every $m \in \mathscr{M}$, $\mathrm{pen}_0(m) = p_2(m)$ and*

$$\forall m, m' \in \mathscr{M}, \qquad \big|\delta(m) - \delta(m')\big| \leqslant \varepsilon_\delta \big[\mathscr{R}(s^\star_m) - \mathscr{R}(s^\star) + \mathscr{R}(s^\star_{m'}) - \mathscr{R}(s^\star)\big] + \varepsilon'_\delta\,. \tag{60}$$

Then, for every $C \in [0,1)$,

$$p_2\big(\widehat{m}^{(0)}_{\min}(C)\big) \geqslant \sup_{m\in\mathscr{M}} \left\{ p_2(m) - \frac{2}{1-C}\big[\mathscr{R}(s^\star_m) - \mathscr{R}(s^\star)\big] \right\} - \frac{\varepsilon'_\delta}{1-C}\,, \tag{61}$$

and if $\mathscr{R}(s^\star_{m_1}) = \mathscr{R}(s^\star)$ for some $m_1 \in \mathscr{M}$ with $p_2(m_1) > 0$, for any $\alpha \in (0,1)$,

$$\forall C \leqslant 1 - \eta_\alpha\,, \quad p_2\big(\widehat{m}^{(0)}_{\min}(C)\big) \geqslant (1-\alpha)p_2(m_1) \quad \textit{with} \quad \eta_\alpha = \frac{\varepsilon'_\delta}{\alpha p_2(m_1)}\,. \tag{62}$$

Assume in addition that the data $\xi_1, \ldots, \xi_n \in \mathscr{X}$ are i.i.d. and some contrast function $\gamma : \Xi \times \mathbb{S} \to \mathbb{R}$ and constants $A, L > 0$ exist such that

$$\forall t \in \mathbb{S}, \quad \widehat{\mathscr{R}}_n(t) = \frac{1}{n}\sum_{i=1}^n \gamma(\xi_i, t) \quad \textit{and} \quad \mathscr{R}(t) = \mathbb{E}\big[\widehat{\mathscr{R}}_n(t)\big] = \mathbb{E}\big[\gamma(\xi_1, t)\big]\,, \tag{63}$$

$$\forall t \in \mathbb{S}, \quad \big|\gamma(\xi_1, t)\big| \leqslant A \quad \textit{a.s.}\,, \tag{64}$$

$$\textit{and} \qquad \forall m \in \mathscr{M}\,, \quad \mathrm{var}\big(\gamma(\xi_1, s^\star_m) - \gamma(\xi_1, s^\star)\big) \leqslant L\big[\mathscr{R}(s^\star_m) - \mathscr{R}(s^\star)\big]\,. \tag{65}$$

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

Then, for every $x \geqslant 0$, with probability at least $1 - 2\operatorname{card}(\mathscr{M})\mathrm{e}^{-x}$, for any $\theta > 0$, Eq. (60) holds true with

$$\varepsilon_\delta = \theta \quad \text{and} \quad \varepsilon'_\delta = \left(\frac{L}{\theta} + \frac{4A}{3}\right)\frac{x}{n}.$$

Proposition 1 is proved in Appendix A.1. Eq. (62) proves that (β^-) holds true if $p_2(m_1)$ is close to $\sup_{m\in\mathscr{M}} p_2(m)$, which is a reasonable assumption. For instance, in the setting of Section 2, m_1 is given by assumption (**HId**) leading to $\widehat{s}_{m_1} = Y$, hence

$$p_2(m_1) = \frac{1}{n}\left(\|Y - F_{m_1}\|^2 - \|Y - \widehat{F}_{m_1}\|^2\right) = \frac{1}{n}\|\varepsilon\|^2 \approx \sigma^2 .$$

The proof of Proposition 1 also works when assuming only that $p_2(m) \geqslant 0$ for every $m \in \mathscr{M}$ and

$$(1 - \varepsilon_0)p_2(m) \leqslant \operatorname{pen}_0(m) \leqslant (1 + \varepsilon_0)p_2(m),$$

which can be used when $\operatorname{pen}_0(m)$ is a resampling estimate of $\mathbb{E}[p_2(m)]$ for instance. Then, we loose a factor in the "rate" η_α of estimation of $C^\star$ in Eq. (62), see Appendix A.1.

Proposition 1 is new —apart from the fact that its proof relies heavily on the proof technique proposed by [Lerasle and Takahashi, 2016]— but rather abstract in its general form. Under the additional conditions (63)–(65), it can be used for minimum-contrast estimators

$$\widehat{s}_m \in \operatorname*{argmin}_{t\in S_m}\left\{\widehat{\mathscr{R}}_n(t)\right\}$$

with a bounded contrast γ, so that one automatically has $p_2(m) \geqslant 0$ and Eq. (63)–(64) as requested. Then, Eq. (65) is a classical assumption [Massart and Nédélec, 2006] which holds for bounded regression with the least-squares contrast —with $L = A = 8M^2$ if data are bounded by M, according to [Arlot and Massart, 2009]—, and for binary classification with the 0–1 loss under the margin condition [Mammen and Tsybakov, 1999, Massart and Nédélec, 2006]. Let us emphasize that Algorithm 5 has never been justified for binary classification with the 0–1 loss up to now, so Proposition 1 is of significant interest even if it only provides a partial justification —with (β^-) only, in a rather abstract form.

5.2.2. *Above the minimal penalty:* (β^+)

Before detailing two approaches for proving (β^+), let us recall that if an oracle inequality like (β'^+) is available, a simple way to prove (β^+) is to use the connection from (β'^+) to (β^+) explained in Section 4.4.

Following the proof of Theorem 1 When $C^\star \operatorname{pen}_0(m) = \mathbb{E}[p_2(m)]$, following the proof of Eq. (16) in Theorem 1, (β^+) can be proved by showing that for some well-chosen $m_2 \in \mathscr{M}$,

$$\forall C > C^\star, \qquad \inf_{m\in\mathscr{M}\,/\,\mathscr{C}_m > \mathscr{C}_{\text{small}}} G_C(m) > G_C(m_2),$$

which requires concentration inequalities for $\delta(m) - \delta(m_2)$ and for $p_2(m)$, see Section 5.2.3. A natural choice is

$$m_2 \in \operatorname*{argmin}_{m\in\mathscr{M}}\left\{\mathbb{E}\left[\mathscr{R}(\widehat{s}_m) - \mathscr{R}(s^\star)\right]\right\}.$$

Generalizing the strategy of [Lerasle and Takahashi, 2016, Garivier and Lerasle, 2011] When $\mathrm{pen}_0(m) = \mathscr{C}_m \propto p_2(m)$, the approach used by [Lerasle and Takahashi, 2016, Garivier and Lerasle, 2011] can be summarized into the following proposition.

Proposition 2. *Let us consider the general framework of Section 3.1 and use the notation of Section 4.1. Let $\varepsilon_\delta, \varepsilon_p \in [0,1)$, $\varepsilon'_\delta > 0$, assume that Eq. (60) holds true and that for every $m \in \mathscr{M}$, $\mathrm{pen}_0(m) = p_2(m)$ and*

$$\bigl|p_1(m) - p_2(m)\bigr| \leqslant \varepsilon_p p_1(m)\,. \tag{66}$$

Then, for every $C > 1$, we have

$$\mathscr{R}\Bigl(\widehat{s}_{\widehat{m}^{(0)}_{\min}(C)}\Bigr) - \mathscr{R}(s^\star) \leqslant K(C) \inf_{m \in \mathscr{M}} \bigl\{\mathscr{R}(\widehat{s}_m) - \mathscr{R}(s^\star)\bigr\} + K'(C) \tag{67}$$

$$\text{and} \qquad p_2\Bigl(\widehat{m}^{(0)}_{\min}(C)\Bigr) \leqslant K(C)(1+\varepsilon_p) \inf_{m \in \mathscr{M}} \bigl\{\mathscr{R}(\widehat{s}_m) - \mathscr{R}(s^\star)\bigr\} + K'(C)(1+\varepsilon_p)\,, \tag{68}$$

$$\text{where} \qquad K(C) := \frac{\max\bigl\{(C-1)(1+\varepsilon_p)\,,\, 1+\varepsilon_\delta\bigr\}}{\min\bigl\{(C-1)(1-\varepsilon_p)\,,\, 1-\varepsilon_\delta\bigr\}}$$

$$\text{and} \qquad K'(C) := \frac{\varepsilon'_\delta}{\min\bigl\{(C-1)(1-\varepsilon_p)\,,\, 1-\varepsilon_\delta\bigr\}}\,.$$

Proposition 2 is proved in Appendix A.2.

Now, assume that Eq. (60) and (66) hold on a large-probability event. Then, taking $C = 1+\eta$ with $\eta > 0$, we get $K(C) \leqslant \mathscr{O}(1)/\eta$ if η is small enough, and Eq. (67) implies (β'^{+}). If in addition $p_2(m_1)$ stays bounded away from zero as n tends to infinity, Eq. (62) implies (β^{-}). If moreover the oracle risk tends to zero and $\varepsilon'_\delta = \mathrm{o}(1)$, then, Eq. (68) implies (β^{+}). Assuming also that $\max\{\varepsilon_\delta, \varepsilon_p\} = \mathrm{o}(1)$, then $K(2) = 1 + \mathrm{o}(1)$ hence Eq. (67) with $C = 2$ implies a first-order optimal oracle inequality (γ) with $\mathrm{pen}_1 = 2\,\mathrm{pen}_0$.

The conditions of Proposition 2 can be relaxed. First, $\mathrm{pen}_0(m) = p_2(m)$ can be replaced by $(1-\varepsilon_0)p_2(m) \leqslant \mathrm{pen}_0(m) \leqslant (1+\varepsilon_0)p_2(m)$ for some $\varepsilon_0 \geqslant 0$ with $C(1-\varepsilon_0) > 1$, which can be used when $\mathrm{pen}_0(m)$ is a resampling estimate of $\mathbb{E}[p_2(m)]$ for instance. Second, Eq. (66) can be replaced by $\forall m \in \mathscr{M}$, $-\varepsilon'_p + \varepsilon_p^- p_1(m) \leqslant p_2(m) \leqslant \varepsilon_p^+ p_1(m) + \varepsilon'_p$ for some $\varepsilon_p^-, \varepsilon_p^+ > 0$ and $\varepsilon'_p \geqslant 0$. Then, $K(C)$ and $K'(C)$ are slightly enlarged, as well as the bound in Eq. (68), see Appendix A.2. In particular, Proposition 2 can justify (γ) with $\mathrm{pen}_1 = (1+\alpha^{-1})\,\mathrm{pen}_0$ provided that both ε_p^- and ε_p^+ converge to $\alpha > 0$, and ε'_p is small enough.

Assumption (66) is strong and we do not expect that it can be proved as generally as assumption (60) in Proposition 1. Nevertheless, it holds when $p_1(m)$ and $p_2(m)$ both concentrate around $\mathbb{E}[p_1(m)] \approx \mathbb{E}[p_2(m)]$ for every $m \in \mathscr{M}$, and it can be satisfied in other cases. For instance, for least-squares estimators in least-squares fixed-design regression (Section 2) or in least-squares density estimation [Lerasle, 2012], $p_1(m) = p_2(m)$ almost surely. Bounding $|p_1(m) - p_2(m)|$ also turns out to be easier to get than a concentration inequality for p_1 and p_2 separately in some settings where $p_1(m) \neq p_2(m)$ in general [Garivier and Lerasle, 2011].

Whatever the proof technique used —through (β'^{+}), as in the proof of Theorem 1 or as in [Lerasle and Takahashi, 2016]—, proving that the upper bound on the complexity in (β^{+}) is much smaller than the lower bound in (β^{-}) is done by assuming that the oracle risk tends to zero

as $n \to \infty$, or at least that some estimator of "not too large" complexity has a small approximation error (in Theorem 1). We conjecture that such an assumption is unavoidable in general.

5.2.3. *Concentration inequalities*

The proof techniques summarized in Sections 5.2.1–5.2.2 require some concentration inequalities for $\delta(m) - \delta(m')$, for $p_2(m)$ and for $|p_1(m) - p_2(m)|/p_1(m)$.

Concentration of $\delta(m) - \delta(m')$ As explained in the proof of Proposition 1, $\delta(m) - \delta(m')$ is a sum of independent and identically distributed random variables, so it can be concentrated with Bernstein's inequality [Boucheron et al., 2013, Theorem 2.10], leading to a result like Eq. (60) if a boundedness assumption (64) and some margin-type condition (65) are satisfied.

Concentration of $p_2(m)$ The problem is much harder for $p_2(m)$. It can be seen as proving a non-asymptotic version of the Wilks phenomenon [Wilks, 1938] in a non-parametric setting with model misspecification [Boucheron and Massart, 2011], which makes this problem interesting beyond minimal-penalty algorithms. In addition to the settings mentioned in Section 4.2, concentration results for p_2 are available in two cases.
For bounded-contrast minimizers, a concentration inequality is proved in a general setting including bounded regression and classification with Vapnik-Chervonenkis classes [Boucheron and Massart, 2011]. This result can be used for proving that Algorithm 5 works with regressogram estimators [Arlot and Massart, 2009].
For maximum-likelihood estimators, in a parametric setting [Spokoiny, 2012], in a semiparametric setting [Andresen and Spokoiny, 2014] and in a nonparametric setting with a quadratic penalty [Spokoiny, 2017], p_2 is close to some quadratic form with high probability, and this quadratic form itself satisfies some concentration properties. Nevertheless, these results have not been used yet for proving Algorithm 5 works.
Note also that a concentration inequality for $p_2(m)$, with histogram (maximum-likelihood) density estimators and the Kullback risk, have been obtained by [Saumard and Navarro, 2018b], improving previous results by [Saumard, 2010c].

Proof of Eq. (66) Apart from the specific approaches mentioned in Section 5.2.2, we do not know any result for bounding directly $|p_1(m) - p_2(m)|/p_1(m)$ as required in Eq. (66).

5.3. *Hints for (γ): how to prove that $C^\star \operatorname{pen}_1$ is an optimal penalty?*

When $C^\star \operatorname{pen}_1(m) = \mathbb{E}[p_1(m) + p_2(m)]$, oracle inequalities ($\gamma$) or ($\widetilde{\gamma}$) usually rely on some concentration inequality for the ideal penalty $\mathscr{R}(\widehat{s}_m) - \widehat{\mathscr{R}}_n(\widehat{s}_m) = p_1(m) + \delta(m) + p_2(m)$, as in step 3 of the proof of Theorem 1. One actually needs only concentration for

$$\left[\mathscr{R}(\widehat{s}_m) - \widehat{\mathscr{R}}_n(\widehat{s}_m)\right] - \left[\mathscr{R}(\widehat{s}_{m'}) - \widehat{\mathscr{R}}_n(\widehat{s}_{m'})\right]$$

for all $m, m' \in \mathscr{M}$. Concentration results for $\delta(m) - \delta(m')$ and for $p_2(m)$ are reviewed in Section 5.2 since they are usually required for proving (β).

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

Concentration of the excess risk $p_1(m)$ What remains is to concentrate $p_1(m)$ —or equivalently $\mathscr{R}(\widehat{s}_m)$ or $\mathscr{R}(\widehat{s}_m) - \mathscr{R}(s^\star)$— around its expectation, a difficult problem that has not been solved except in a few settings: the ones for which a full proof of the slope heuristics exists —see Section 4.2—, and the ones listed below.

Several papers recently tackled the case of fixed-design linear regression with the least-squares risk, when $\widehat{s}_m$ minimizes a (penalized) least-squares criterion over a convex set, assuming that the penalty Ω is convex. Concentration inequalities for

$$\sqrt{\mathscr{R}(\widehat{s}_m) - \mathscr{R}(s^\star)} \qquad \text{or} \qquad \sqrt{\mathscr{R}(\widehat{s}_m) - \mathscr{R}(s^\star) + \Omega(\widehat{s}_m)}$$

are available under different assumptions on the noise (Gaussian or not) and on Ω [Chatterjee, 2014, Bellec, 2017, Bellec and Tsybakov, 2017, Muro and Geer, 2018]. They apply to various examples such as the Lasso (in its constrained formulation) and isotonic regression [Chatterjee, 2014], the Lasso and the group Lasso (in their usual regularization formulation) [Bellec, 2017, Bellec and Tsybakov, 2017], splines and total variation regularization [Muro and Geer, 2018]. When Ω is a semi-norm, [Bellec, 2018] proves upper and lower bounds on $\mathbb{E}[\mathscr{R}(\widehat{s}_m)]$. A related paper on this topic is [Chen et al., 2017].

For general losses, high-probability upper and lower bounds on $\mathscr{R}(\widehat{s}_m) - \mathscr{R}(s^\star)$ —sometimes plus a regularization term $\Omega(\widehat{s}_m)$— are proved by [Bartlett and Mendelson, 2006] for general empirical minimizers —with a rather abstract result—, by [Saumard, 2010b] for "regular" estimators and losses, and by [van de Geer and Wainwright, 2017] for regularized empirical risk minimizers —with precise applications provided for "linear losses" such as linearized least-squares regression, maximum-likelihood estimators on an exponential model, and log-linear regression. Note that the general approaches of [Bartlett and Mendelson, 2006, Saumard, 2010b, van de Geer and Wainwright, 2017] are closely related; [Chatterjee, 2014, Bellec, 2017, Bellec and Tsybakov, 2017, Muro and Geer, 2018], which are mentioned above for linear regression, use a similar technique that is exposed clearly by [Bellec, 2017, Section 2] for instance.

[Saumard, 2017] proves a concentration inequality for the quadratic risk $\mathscr{R}(\widehat{s}_m)$ of a least-squares estimator over a convex set, in the heteroscedastic random-design regression setting; this result requires to handle specifically the quadratic part of the empirical process, which cannot be concentrated tightly with the general approach of [van de Geer and Wainwright, 2017] for instance. Note also that the result of [Saumard, 2017] applies to more general models than the ones of [Navarro and Saumard, 2017] for which a full proof of the slope heuristics exist.

For histogram (maximum-likelihood) estimators and the Kullback risk in density estimation, in addition to [Saumard, 2010c] that Section 4.2 mentions, concentration inequalities for $\mathscr{R}(\widehat{s}_m)$ have been obtained by [Castellan, 1999], and [Saumard and Navarro, 2018b] has recently improved them.

Let us also recall that Proposition 2 in Section 5.2.2 provides an alternative approach for proving (γ) when $\mathrm{pen}_1(m) \propto p_2(m)$.

6. Related procedures

Minimal-penalty calibration algorithms are primarily made for model/estimator selection, but in the fixed-design regression setting (Algorithms 1 and 4) they also provide an estimator $\widehat{C}_{\mathrm{jump}}$ of

the noise variance σ^2. This section compares minimal penalties to its main alternatives for both tasks, starting by residual-(co)variance estimation.

6.1. Residual-variance estimation

Let us consider the fixed-design regression setting of Sections 2 and 3.3–3.4 and their notation. An example of interest is when

$$\forall i \in \{1,\dots,n\}\,, \quad F_i = f(x_i) \quad \text{for some smooth } f \text{ and some design points } x_i \in \mathbb{R}^d. \qquad (\star)$$

Literature on residual-variance estimation Many estimators exist for the residual variance σ^2 in nonparametric regression. An exhaustive list is beyond the scope of this paper; for more references, we refer to [Hall et al., 1990, Dette et al., 1998, Spokoiny, 2002, Müller et al., 2003, Liitiäinen et al., 2009], and to [Chatterjee, 2015, Reid et al., 2016, Giacobino et al., 2017] for the high-dimensional variable-selection case. A related problem is noise-variance estimation in heteroscedastic regression, see [Brown and Levine, 2007, Gendre, 2008] and references therein.

This section focuses on minimal-penalty based estimators and on estimators that are quadratic forms of the data vector $Y \in \mathbb{R}^n$, that is,

$$\widehat{\sigma}^2_B := \frac{\langle Y, BY\rangle}{\operatorname{tr}(B)} \qquad (69)$$

where B is some $n \times n$ symmetric matrix. Eq. (69) actually covers several, if not all, classical variance estimators —in particular the ones suggested in the context of model/estimator selection with C_p or C_L— which allows their common non-asymptotic analysis [Dette et al., 1998].

Residual-based estimators The most classical variance estimators are based upon the residuals —through the empirical risk— on some model S_{m_0}:

$$\widehat{\sigma}^2_{m_0} := \frac{1}{n - D_{m_0}} \left\| Y - \widehat{F}_{m_0} \right\|^2. \qquad (70)$$

Remark that $\widehat{\sigma}^2_{m_0} = \widehat{\sigma}^2_B$ with $B = I_n - \Pi_{m_0}$. When σ^2 must be estimated in the formula of the C_p penalty, the classical suggestions are of the form $\widehat{\sigma}^2_{m_0}$, see [Mallows, 1973, Efron, 1986, Baraud, 2000].

The bias of $\widehat{\sigma}^2_{m_0}$ as an estimator of σ^2 can be derived from Eq. (8):

$$\mathbb{E}\left[\widehat{\sigma}^2_{m_0}\right] - \sigma^2 = \frac{1}{n - D_{m_0}} \left\| (I_n - \Pi_{m_0}) F \right\|^2. \qquad (71)$$

If $F \in S_{m_0}$, then $\widehat{\sigma}^2_{m_0}$ is unbiased, and otherwise it suffers some upward bias, depending on the approximation error and on the dimension of S_{m_0}. Proposition 4 in Appendix A.4 provides a general formula for the variance and MSE of $\widehat{\sigma}^2_{m_0}$. For instance, assuming for simplicity that the noise is Gaussian,

$$\mathbb{E}\left[\left(\widehat{\sigma}^2_{m_0} - \sigma^2\right)^2\right] = \frac{2\sigma^4}{n - D_{m_0}} + \frac{4\sigma^2 \left\| (I_n - \Pi_{m_0}) F \right\|^2}{(n - D_{m_0})^2} + \frac{\left\| (I_n - \Pi_{m_0}) F \right\|^4}{(n - D_{m_0})^2}. \qquad (72)$$

Choosing the model S_{m_0} without prior knowledge is a difficult question. For minimizing the MSE, one must trade off terms depending on $1/(n-D_{m_0})$ and on the approximation error that vary differently as functions of S_{m_0}, which can be as difficult as the model-selection problem. When some unbiased model S_{m_0} is known with $D_{m_0}=\mathrm{o}(n)$, taking it for the estimation of σ^2 is a reasonable choice. This matches the suggestion of [Mallows, 1973, Efron, 1986] in the variable-selection setting with a full model of dimension $p=\mathrm{o}(n)$. In setting ($\star$), when contiguous x_i are close enough, a natural choice for S_{m_0} is the linear span of $(e_{2i}+e_{2i-1})_{1\leqslant i\leqslant n/2}$ where $(e_1,\dots,e_n)$ denotes the canonical basis of $\mathbb{R}^n$. Then, assuming n is even for simplicity,

$$\widehat{\sigma}^2_{m_0}=\frac{1}{n}\sum_{i=1}^{n/2}(Y_{2i}-Y_{2i-1})^2 \tag{73}$$

is a consistent estimator of σ^2 if f is uniformly continuous and $\max_{1\leqslant i\leqslant n}\|x_i-x_{i+1}\|=\mathrm{o}(1)$.

Note that for high-dimensional variable selection, several residual-based estimators on a data-driven model m_0 —for instance, chosen by cross-validation— are available, but few theoretical guarantees are known for them [Reid et al., 2016, Giacobino et al., 2017].

Variance estimation with minimal penalties The problem of choosing m_0 for $\widehat{\sigma}^2_{m_0}$ is solved (bypassed in fact) by using the slope heuristics. Let us state non-asymptotic risk bounds for $\widehat{C}_{\text{thr.}}$ and $\widehat{C}_{\text{window}}$ as estimators of σ^2, that derive from Theorem 1 and its proof.

Proposition 3. *In the framework described in Section 2.1, assume that $\mathscr{M}$ is finite, contains at least one model of dimension at most $c_n\in[0,n/3)$, and that* (**HId**) *and* (**HG**) *hold true —see Section 2.5. Let $\widehat{C}_{\text{thr.}}(T_n)$ be defined by Eq. (20) with $T_n\in(c_n,n)$, and $\widehat{C}_{\text{window}}(\eta)$ be defined by Eq. (19) with $\eta>0$. For any $x\geqslant 0$ and $T\in(c_n,n)$, let us define*

$$C_1(x;T):=\sigma^2\left(1-\frac{4\sqrt{\frac{x}{n}}+6\frac{x}{n}}{1-\frac{T}{n}}\right)$$

$$\text{and}\qquad C_2(x;T;c_n):=\sigma^2\left[1+\frac{4n}{T-c_n}\left(\sqrt{\frac{x}{n}}+\frac{2x}{n}\right)\right]+\frac{2n}{T-c_n}\mathscr{B}(c_n)$$

$$\text{where}\qquad \mathscr{B}(c_n):=\inf_{m\in\mathscr{M}\,/\,D_m\leqslant c_n}\left\{\frac{1}{n}\left\|(I_n-\Pi_m)F\right\|^2\right\}.$$

Then, an event Ω_x of probability at least $1-4\operatorname{card}(\mathscr{M})\mathrm{e}^{-x}$ exists on which

$$\frac{C_1\left(x;\frac{2n}{3}\right)}{1+\eta}\leqslant\widehat{C}_{\text{window}}(\eta)\leqslant C_2\left(x;\frac{n}{3}\right)(1+\eta)\qquad\text{if }\eta>\sqrt{\frac{C_2\left(x;\frac{n}{3}\right)}{C_1\left(x;\frac{2n}{3}\right)}}-1\text{ and }x\in\left[0,\frac{n}{180}\right], \tag{74}$$

$$\text{and}\qquad C_1(x;T_n)\leqslant\widehat{C}_{\text{thr.}}(T_n)\leqslant C_2(x;T_n;c_n)\,. \tag{75}$$

If we assume in addition that $c_n\leqslant T_n/2$, then,

$$\begin{aligned}\mathbb{E}\left[\left(\widehat{C}_{\text{thr.}}-\sigma^2\right)^2\right]&\leqslant 739\max\left\{\left(1-\frac{T_n}{n}\right)^{-2},\left(\frac{T_n}{2n}\right)^{-2}\right\}\\&\quad\times\left[\left(\mathscr{B}\left(\frac{T_n}{2}\right)\right)^2+\frac{\sigma^4\log\left[4\operatorname{card}(\mathscr{M})\right]}{n}+\sigma^4\left(\frac{\log\left[4\operatorname{card}(\mathscr{M})\right]}{n}\right)^2\right].\end{aligned} \tag{76}$$

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

Proposition 3 is proved in Appendix A.3. Note that the constant 739 in Eq. (76) can be strongly reduced under mild additional assumptions, see Appendix A.3. Proposition 3 can also be extended to $(\phi^2\sigma^2)$-sub-Gaussian noise, at the price of replacing x by $L\phi^2 x$ in Eq. (74)–(75) and $\log[4\,\mathrm{card}(\mathscr{M})]$ by $L\phi^2\log[4\,\mathrm{card}(\mathscr{M})]$ in Eq. (76), where L is a numerical constant; see Remark 1 in Section 2.5.

If $\mathscr{B}(c_n)\to 0$ as $n\to+\infty$ —which is a mild assumption—, taking $x=2\log(n)+\log(4\,\mathrm{card}(\mathscr{M}))$ in Eq. (74)–(75) shows that $\widehat{C}_{\text{thr.}}$ and $\widehat{C}_{\text{window}}$ estimate consistently σ^2, with deviation bounds of order

$$\mathscr{B}(c_n)+\sigma^2\sqrt{\frac{\log(n)+\log\bigl[\mathrm{card}(\mathscr{M})\bigr]}{n}}\,,$$

provided that

$$T_n=\rho n\,,\quad \rho\in(0,1)\,,\qquad\text{and}\qquad \sigma^2\eta\propto\mathscr{B}(c_n)+\sqrt{\frac{\log(n)+\log\bigl[\mathrm{card}(\mathscr{M})\bigr]}{n}}\,.$$

These deviation bounds for $\widehat{C}_{\text{thr.}}$ and $\widehat{C}_{\text{window}}$ can be interpreted as an oracle inequality, since they coincide with the best possible risk of $\widehat{\sigma}^2_{m_0}$ with $D_{m_0}\leqslant c_n$, without any prior knowledge except the choice of $(S_m)_{m\in\mathscr{M}}$. Indeed, Eq. (72) shows that when $D_{m_0}\leqslant c_n\leqslant\rho n$ with $\rho<1$, up to constants depending on ρ only,

$$\sqrt{\mathbb{E}\bigl[(\widehat{\sigma}^2_{m_0}-\sigma^2)^2\bigr]}\gtrsim\frac{1}{n}\bigl\|(I_n-\Pi_{m_0})F\bigr\|^2+\frac{\sigma^2}{\sqrt{n}}\geqslant\mathscr{B}(c_n)+\frac{\sigma^2}{\sqrt{n}}\,.$$

The bound (76) on the mean squared error (MSE) of $\widehat{C}_{\text{thr.}}$ can be compared to the minimax optimal rate $\propto n^{-\min\{1,8/d\}}$ for the MSE in setting ($\star$) when f has a bounded second-order derivative [Spokoiny, 2002]. Assuming that $\log[\mathrm{card}(\mathscr{M})]=\mathrm{o}(n^{1/9})$, that the approximation error $\mathscr{B}(c_n)$ is minimax optimal hence of order $c_n^{-4/d}$, and that $c_n\leqslant\rho n$ with $\rho<1$, we get that $\widehat{C}_{\text{thr.}}$ is optimal when $d\geqslant 9$ up to constants and within a factor $\log(\mathrm{card}(\mathscr{M}))$ of the minimax risk when $d\leqslant 8$. Similar risk bounds can easily be obtained from Proposition 3 under different assumptions on the signal. For instance, in setting ($\star$) with f that is α-Hölderian for some $\alpha>0$, an approximation error of order $D_m^{-2\alpha/d}$ can be obtained with local polynomials of maximal degree $\lfloor\alpha\rfloor$. We conjecture that these residual-variance estimation bounds are minimax-optimal up to logarithmic factors provided that $(S_m)_{m\in\mathscr{M}}$ has a cardinality at most polynomial in n and achieves the minimax approximation error bounds. These consequences of Proposition 3 have the flavor of adaptive risk bounds derived from oracle inequalities [Birgé and Massart, 1997], which is new for residual-variance estimation to the best of our knowledge. Therefore, the additional $\log(\mathrm{card}(\mathscr{M}))$ factor —coming from the union bound over $m\in\mathscr{M}$ and typically of order $\log(n)$— seems a mild price for the versatility of $\widehat{C}_{\text{thr.}}$ and $\widehat{C}_{\text{window}}$.

Residual-based estimators vs. the slope heuristics In addition to the risk bounds comparison above, we can compare the definition of $\widehat{\sigma}^2_{m_0}$ with the one of $\widehat{C}_{\text{slope}}$ in Algorithm 2. On the one hand, $\widehat{\sigma}^2_{m_0}$ estimates the asymptotic slope of $-\|Y-\widehat{F}_m\|^2$ as a function of D_m from two points: m_0, and m_1 such that $S_{m_1}=\mathbb{R}^n$. On the other hand, Algorithm 2 makes a (robust) linear regression over, say, all m such that $D_m\in[n/2,n]$. An illustration is provided by Figure 10 in the Appendix.

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

This confirms that minimal penalties —here, the slope heuristics— allow to avoid the choice of a single $m_0 \in \mathscr{M}$ by making use of the full collection $(S_m)_{m\in\mathscr{M}}$ for estimating the variance. Intuitively, this difference makes Algorithm 2 more stable and less dependent on some strong assumption on S_{m_0}. The numerical experiments of Figure 6b in Section 7.1 and Figure 10 in the Appendix indeed show that when S_{m_0} happens to be a bad model, $\widehat{\sigma}^2_{m_0}$ can suffer from a large error, whereas $\widehat{C}_{\text{thr.}}$ is much more robust.

Residuals of linear estimators Some residual-based estimators have also been proposed with other linear estimators $\widehat{F}_{m_0} = A_{m_0} Y$ of F, that is, defined as $\widehat{\sigma}^2_B$ in Eq. (69) with $B = (I_n - A_{m_0})^\top (I_n - A_{m_0})$. For instance, A_{m_0} can correspond to some Nadaraya-Watson fit (a.k.a. kernel-based estimator) [Hall and Marron, 1990] or to spline smoothing [Carter and Eagleson, 1992]; see [Dette et al., 1998] for more references. In setting ($\star$), the challenging case $d > 1$ can be tackled with k-nearest neighbors [Liitiäinen et al., 2010] or a local linear fit of f [Spokoiny, 2002]. All these estimators suffer from the same drawback as $\widehat{\sigma}^2_{m_0}$, that is, they rely on the choice of a single matrix A_{m_0}, hence requiring to specify the regularization parameter m_0. On the contrary, the minimal-penalty approach of Algorithm 4 avoids this choice in a principled way.

Difference-based estimators Difference-based estimators are an important family of residual-variance estimators, which are designed for setting ($\star$) when $\|x_i - x_{i+1}\| = \mathrm{o}(1)$ and f is smooth, often assuming $d = 1$. The first example has been proposed by Rice [Rice, 1984],

$$\widehat{\sigma}^2_{\text{Rice}} := \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (Y_{i+1} - Y_i)^2 \,,$$

which is close to the residual-based estimator defined by Eq. (73). More generally, difference-sequence estimators of order $m \geqslant 1$ are defined by

$$\widehat{\sigma}^2_{(d_0,\dots,d_m)} = \frac{1}{n-m} \sum_{i=1}^{n-m} \left(\sum_{j=0}^{m} d_j Y_{i+j} \right)^2 \quad \text{where} \quad \sum_{j=0}^{m} d_j = 0 \quad \text{and} \quad \sum_{j=0}^{m} d_j^2 = 1 \,.$$

The only admissible sequence $(d_j)_{j=0,\dots,m}$ for $m = 1$ leads to $\widehat{\sigma}^2_{\text{Rice}}$. For a general order m, when $x_i \in \mathbb{R}$, the optimal sequence in terms of MSE does not depend on f asymptotically and can be computed explicitly [Hall et al., 1990], although the picture can be quite different in a non asymptotic setting [Dette et al., 1998].

Assuming $x_i \in \mathbb{R}$, difference-based estimators of order $m \geqslant 1$ are suboptimal by a constant factor $1 + 1/(2m)$ in terms of MSE for normal data [Hall et al., 1990], while for instance the residual-based estimator of [Hall and Marron, 1990] attains the optimal rate $\mathrm{var}(\varepsilon_1^2)/n$. This issue is corrected for instance with covariate matching [Müller et al., 2003, Du and Schick, 2009], which in the case of order one consists in replacing $\widehat{\sigma}^2_{\text{Rice}}$ by

$$\frac{1}{2n(n-1)} \sum_{i \neq j} W_{i,j} (Y_i - Y_j)^2$$

with some well-chosen weights $W_{i,j} \geqslant 0$ [Müller et al., 2003]. Another variant of difference-based estimators which is asymptotically optimal is studied in [Tong et al., 2013].

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

Choosing m and the sequence $(d_j)_{j=0,\dots,m}$ is also a difficult problem with no prior knowledge [Dette et al., 1998]. But the main drawback of such estimators is that, when f is not continuous, they can be severely biased in an unpredictable way. An empirical method for detecting whether the bias is small enough is proposed by [Buckley and Eagleson, 1989] and might be useful.

6.2. *Estimation of the residual covariance matrix*

Assume now that several regression problems such as (1) must be solved simultaneously, a framework known as 'multi-task regression', 'multivariate regression', 'multiple linear regression', and 'seemingly unrelated regression', see [Solnon et al., 2012, Solnon, 2013] for references. One observes $Y^j = F^j + \varepsilon^j \in \mathbb{R}^n$ for $j = 1,\dots,p$, assuming the noise vectors $\mathscr{E}_i := (\varepsilon_i^j)_{j=1,\dots,p} \in \mathbb{R}^p$ are independent and identically distributed, with zero mean and covariance matrix $\Sigma \in \mathscr{M}_p(\mathbb{R})$. Then, a natural extension of the residual-variance estimation problem is the estimation of Σ with as few prior knowledge on the F^j as possible, which is often required for the multi-task problem of estimating $(F^j)_{j=1,\dots,p}$. For instance, [Solnon et al., 2012] makes use of the prior knowledge that the F^j are close, in combination with kernel ridge regression, and selects regularization parameters with a penalty generalizing C_L which depends on the full matrix Σ.

An estimator $\widehat{\Sigma}$ of Σ based upon minimal penalties is proposed by [Solnon et al., 2012]. It satisfies $(1-\eta)\Sigma \preceq \widehat{\Sigma} \preceq (1-\eta)\Sigma$ with large probability, with $\eta \propto p\sqrt{\log(n)/n}c(\Sigma)^2$ where $c(\Sigma)$ is the condition number of Σ. The construction of $\widehat{\Sigma}$ goes as follows:

(i) For $j \in \{1,\dots,p\}$, apply Algorithm 4 to the one-dimensional regression problem $Y^j = F^j + \varepsilon^j$, and store $\widehat{a}_j = \widehat{C}_{\text{jump}}$ which estimates $a_j := \Sigma_{j,j}$.

(ii) For $i \neq j \in \{1,\dots,p\}$, apply Algorithm 4 to the one-dimensional regression problem $(Y^i + Y^j) = (F^i + F^j) + (\varepsilon^i + \varepsilon^j)$, and store $\widehat{a}_{i,j} = \widehat{C}_{\text{jump}}$ which estimates $a_{i,j} := \Sigma_{i,i} + \Sigma_{j,j} + 2\Sigma_{i,j}$.

(iii) Denote by J the linear map on $\mathbb{R}^{p(p+1)/2}$ that sends $((a_j)_{1\leqslant j\leqslant p}, (a_{i,j})_{1\leqslant i\neq j\leqslant p})$ to Σ, and define $\widehat{\Sigma} = J((\widehat{a}_j)_{1\leqslant j\leqslant p}, (\widehat{a}_{i,j})_{1\leqslant i\neq j\leqslant p})$.

This construction can actually be generalized to any other one-dimensional residual-variance estimator $\widehat{\sigma}^2$ that satisfies $(1-\eta_0)\sigma^2 \leqslant \widehat{\sigma}^2 \leqslant (1+\eta_0)\sigma^2$ with large probability, leading to a similar result with $\eta \propto c(\Sigma)p\eta_0$. The remarkable property of the minimal-penalty-based estimator $\widehat{\Sigma}$ of [Solnon et al., 2012] is that it suffices to assume that an estimator of small complexity has a small approximation error —with slightly stronger constraints compared to Theorem 1, see the exact assumptions of [Arlot and Bach, 2011]— for each one-dimensional problem Y^j to get this assumption automatically satisfied for all the $Y^i + Y^j$, $i \neq j$.

6.3. *Model/estimator-selection procedures based on C_p/C_L*

Let us go back to the model/estimator-selection problem in the fixed-design regression setting. For selecting among linear estimators with the least-squares risk, a popular penalization approach is Mallows' C_L [Mallows, 1973], as described in Section 3.3:

$$\widehat{m}_{C_L} \in \operatorname*{argmin}_{m\in\mathscr{M}} \left\{ \frac{1}{n}\left\| \widehat{F}_m - Y \right\|^2 + \operatorname{pen}_{C_L}(\sigma^2, m) \right\} \quad \text{with} \quad \operatorname{pen}_{C_L}(\sigma^2, m) := \frac{2\sigma^2 \operatorname{tr}(A_m)}{n}. \tag{77}$$

In the particular case of projection estimators, $\operatorname{tr}(A_m) = D_m$ the dimension of the corresponding model, and C_L reduces to C_p which is described in Section 2.2. Both C_p and C_L penalties assume

that the noise-level σ^2 is known, so in general it must be replaced by some data-driven estimation of it. Minimal penalties provide an estimator of σ^2 specially built for this estimator-selection task (Algorithms 1 and 4), which can be plugged into Eq. (77) and for which theoretical guarantees can be proved, see Theorem 1 and [Arlot and Bach, 2011]. This subsection reviews some classical ways to estimate σ^2 inside Eq. (77), as well as other estimator-selection procedures that are closely related, in the framework of Section 3.3 (linear estimator selection) and with its notation.

Fixed variance estimator A first option is to replace σ^2 by some fixed estimator $\widehat{\sigma}^2$ of this quantity, for instance chosen among the estimators described in Section 6.1. The most classical choice is to take a residual-based estimator $\widehat{\sigma}^2_{m_0}$ as defined by Eq. (70), for some $m_0 \in \mathscr{M}$ [Mallows, 1973, Efron, 1986, Baraud, 2000]. For projection estimators, this option is often named "C_p" and called "$C_p(\mathscr{L}_0, \mathscr{L})$" by [Efron, 1986, Table 4]. As discussed in Section 6.1, choosing m_0 can then be as difficult as the original estimator-selection problem.

When using the penalty $\mathrm{pen}_{C_L}(\widehat{\sigma}^2, m)$, theoretical guarantees can be obtained as soon as one can prove that $\widehat{\sigma}^2$ is close to σ^2 with large probability, in combination with Theorem 1 or its analogous for linear estimators. For instance, [Baraud, 2000, Thm 6.1] does it for projection estimators with $\widehat{\sigma}^2_{m_0}$ such that $D_{m_0} = n/2$; the approximation error of m_0 then appears as an additive term in the right-hand side of the oracle inequality.

Nevertheless, even if such guarantees imply some asymptotic-optimality result —provided that the approximation error of m_0 tends to zero—, they might not help for choosing the best possible estimator $\widehat{\sigma}^2$ in terms of estimator selection, since these only are *upper bounds*. Indeed, the best bound is obtained when $\widehat{\sigma}^2$ is well concentrated around σ^2, but it is known that overpenalizing a bit —that is, taking $\widehat{\sigma}^2$ slightly larger than σ^2— empirically improves the estimator-selection performance [Arlot, 2009, Section 6.3.2]. Residual-based estimators do overpenalize, because of the approximation error of m_0, but the overpenalization factor is unknown in practice and cannot be controlled without strong assumptions on the target F; the consequences of a bad choice of m_0 with $\widehat{\sigma}^2_{m_0}$ are illustrated in the numerical experiments of Section 7.1, see Figures 6–7.

On the contrary, minimal-penalty algorithms are more than a simple "plug in" of an estimator of σ^2 —independent of the estimator-selection problem— inside pen_{C_L}. As detailed in Section 8.4, minimal-penalty algorithms seem to overpenalize slightly, by design, but a formal proof of this phenomenon remains an open problem.

Variance estimator depending on m Another approach is to plug into Eq. (77) a different variance estimator for each $m \in \mathscr{M}$, by considering the residuals on the model m for which the penalty is computed. In other words, assuming that $D_m < n$ for all $m \in \mathscr{M}$, $\widehat{m}$ is chosen by penalization with the penalty $\mathrm{pen}_{C_L}(\widehat{\sigma}^2_m, m)$. Let us consider projection estimators for simplicity. Then $\widehat{m}$ minimizes over $m \in \mathscr{M}$ the criterion

$$\mathrm{crit}_{\mathrm{FPE}}(m) := \frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2 + \frac{2\widehat{\sigma}^2_m D_m}{n} = \frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2 \left(1 + \frac{2D_m}{n - D_m}\right) \tag{78}$$

which has been proposed by Akaike [Akaike, 1969, Akaike, 1970] under the name FPE —final prediction error— and is called "naive C_p" or "$C_p(\mathscr{L}_0, \mathscr{L}_0)$" by [Efron, 1986, Table 4]. Actually, the FPE criterion (78) actually is a particular case of the multiplicative penalties defined by

Eq. (59) in Section 4.6. More references and non-asymptotic oracle inequalities satisfied by such multiplicative penalties can be found in [Baraud et al., 2009], where it is explained in particular how the FPE criterion (78) should be enlarged, depending on the size of the collection $\mathscr{M}$. The main drawback of multiplicative penalties is probably that they need to deal carefully with the largest models. For instance, for FPE, the results of [Baraud et al., 2009, Theorem 1] assume that $D_m \leqslant 0.39(n+2)-1$ for all $m \in \mathscr{M}$, an assumption that can be changed into $D_m \leqslant \rho n$ for some $\rho < 1$ only when considering a modified multiplicative penalty [Baraud et al., 2009, Corollary 1].

Generalized cross-validation For choosing the regularization parameter of some smoothing methods, Wahba [Wahba, 1977] proposed the criterion called "generalized cross-validation" (GCV) defined as a rotationally invariant form of the cross-validation estimate, that is,

$$\mathrm{crit}_{\mathrm{GCV}}(m) := \frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2 \left(\frac{1}{n}\mathrm{tr}(I_n - A_m)\right)^{-2}. \tag{79}$$

GCV can also be seen as a reweighted cross-validation estimate, which takes into account the asymmetry of the design [Craven and Wahba, 1978]. Nevertheless, as remarked by [Efron, 1986, Remark W]: "Despite its name, GCV is (nearly) a member of the C_p family of estimates". Indeed, considering projection estimators only,

$$\mathrm{crit}_{\mathrm{GCV}}(m) = \frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2 \left(\frac{n}{n-D_m}\right)^2 \approx \frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2 \frac{n+D_m}{n-D_m} = \mathrm{crit}_{\mathrm{FPE}}(m) \tag{80}$$

where the approximation holds true when $D_m \ll n$. Theoretical guarantees for GCV are available in various settings [Li, 1985, Li, 1987, Cao and Golubev, 2006], with the same limitation as the ones of FPE and other multiplicative penalties. For instance, [Cao and Golubev, 2006, Theorem 2] considers a truncated version of GCV where all $m \in \mathscr{M}$ such that $\mathrm{tr}(A_m) > \sqrt{n}$ are discarded, and some examples exist where GCV is not asymptotically optimal [Li, 1986]. Let us finally mention that an empirical comparison of GCV and minimal penalties (Algorithm 4) is done by [Arlot and Bach, 2009, Arlot and Bach, 2011] for several kinds of linear estimators, showing that either Algorithm 4 clearly outperforms GCV or the two methods perform similarly, depending on the setting.

6.4. L-curve, corner, and elbow heuristics

The minimal-penalty algorithms, in particular Algorithms 1–3, can be related to some "L-curve", "corner" or "elbow" heuristics, which are often used for choosing hyperparameters in the statistics and machine learning communities.

The L-curve is defined as a plot of the size of the residuals versus the size or the estimator complexity. Using the notation of Section 3, when the goal is to select an estimator among $(\widehat{s}_m)_{m\in\mathscr{M}}$, the L-curve can be defined as $(\mathscr{C}_m, \widehat{\mathscr{R}}_n(\widehat{s}_m))_{m\in\mathscr{M}}$. For instance, the right part of Figure 2 shows an L-curve (the black dots); Figure 11 in the Appendix provides another illustration. The practical use of the L-curve —a name given by [Hansen and O'Leary, 1993]— has been suggested by [Miller, 1970] and [Lawson and Hanson, 1974, Chapters 25–26]. Some precise heuristic choice of a regularization parameter —often called "the L-curve method"— has first been proposed by [Hansen and O'Leary, 1993, Hansen, 1992] for some inverse problem with Tikhonov regularization. The main idea is that the L-curve has three main parts:

(i) a straight part where the residuals $\widehat{\mathscr{R}}_n(\widehat{s}_m)$ decrease fastly while $\mathscr{C}_m$ is almost constant, where the regularization is too strong,
(ii) a flat part where $\mathscr{C}_m$ increases much while the residuals $\widehat{\mathscr{R}}_n(\widehat{s}_m)$ decrease slowly, where some overfitting occurs, and
(iii) in between, a "*corner*" or "*elbow*", where the regularization parameter is of the correct order.

Therefore, the L-curve is L-shaped —hence its name— and the L-curve method suggests to choose m corresponding to some point at the corner of the 'L'.

Several definitions of the corner can be proposed, as well as several measures of "size" and "complexity" can be considered when plotting the L-curve [Hansen and O'Leary, 1993]. The most common choice is to define the corner as the location where the L-curve has a maximal curvature [Hansen and O'Leary, 1993] —hence the name "maximum-curvature criterion" [Grodzevich and Wolkowicz, 2009] often used for this heuristics— and to look at the L-curve in log-log scale [Hansen and O'Leary, 1993], although several variants exist [Regińska, 1996]. Additional practical problems need also to be solved, especially when $\mathscr{M}$ is discrete (how to define the curvature of a finite set of points?) and when some computational issues arise, for instance because computing every single point of the L-curve is expensive [Hansen and O'Leary, 1993, Castellanos et al., 2002, Hansen et al., 2007, Heng et al., 2010].

Empirical or theoretical studies of L-curve algorithms are available mostly for inverse problems with Tikhonov regularization [Hansen, 1992], truncated SVD [Hansen and O'Leary, 1993, Reichel and Rodriguez, 2013], or conditional gradient regularization [Castellanos et al., 2002], showing reasonably good empirical performances. Several of the papers mentioned in this subsection show that L-curve algorithms compare favorably to generalized cross-validation (GCV) on simulated examples; for instance, [Hansen, 1992] show some similarity between GCV and L-curve algorithms, and report a tendency of GCV to overregularize. Nevertheless, the L-curve method is proved to be not consistent in several families of realistic examples [Engl and Grever, 1994, Vogel, 1996, Hanke, 1996], when the noise tends to zero or when the sample size tends to infinity. According to [Hanke, 1996], the reason for this inconsistency is that the corner seems to correspond to the minimal level of regularization —the minimal penalty, with the words of the present survey— more than to the optimal level; hence choosing m at the corner leads to some overfitting.

The L-curve can also be used similarly in unsupervised learning for choosing the number of clusters, where it is defined as a plot of the within-cluster dispersion as a function of the number of clusters. Indeed, as written by [Tibshirani et al., 2001], "Statistical folklore has it that the location of such an 'elbow' indicates the appropriate number of clusters". Various methods actually use a similar idea [Tibshirani et al., 2001, Sugar and James, 2003, Catherine and Vincent, 2017], although they are not straightforward applications of the method of [Hansen and O'Leary, 1993]. Remark that procedures choosing the number of clusters using the slope heuristics show good experimental results, according to [Maugis and Michel, 2011a, Bontemps and Toussile, 2013] and [Baudry, 2009, Section 4.4].

Comparison with minimal-penalty algorithms Let us start with their common points. Both corner/elbow heuristics and minimal-penalty algorithms are based on the L-curve: directly in Al-

gorithm 2, indirectly in Algorithm 1 since $(D_{\widehat{m}(C)})_{C\geqslant 0}$ can be seen as a reparametrization of the convex hull of the L-curve. Both rely on the idea of detecting a sharp variation of an observable quantity (the curvature / the selected dimension in $D_{\widehat{m}(C)}$) in some region of interest (the optimal value of regularization parameters / the minimal value of the constant in front of the penalty). For both methods, a visual check (of the presence of an elbow / a jump) is possible, and strongly encouraged [Hansen and O'Leary, 1993]. The strength of the connection between elbow heuristics and minimal penalties is emphasized in the following three works. For choosing the constant in front of the penalty for change-point detection, [Lavielle, 2005, Remark 2] suggests an algorithm close to (but slightly different from) the slope heuristics [Lebarbier, 2005], which can be formulated as a maximal-curvature criterion on the L-curve. For Hawkes-process estimation via model selection, when the model collection is large, [Reynaud-Bouret and Schbath, 2010] remarks that the true model generally corresponds to a sharp angle of the L-curve, hence propose an algorithm between the slope and elbow heuristics, which consists in choosing $\widehat{m}(\widehat{C})$ with $\mathrm{pen}(m) = \mathscr{C}_m$ and $\widehat{C}$ equal to the opposite of the slope of the segment joining the first and the last point of the L-curve. For choosing the bandwidth of a Gaussian kernel for quantile estimation with one-class support vector machines, [Vert, 2006, Section 6.2.2] points out an "elbow effect" and locates the elbow region with a "maximal jump" procedure similar to Algorithm 1.

Nevertheless, several important differences must be pointed out between the two approaches. First, the elbow heuristics tries to localize directly the optimal m whereas the slope heuristics localizes it in two steps: first, the minimal penalty, then, the optimal one. Second, the assumptions made on the shape of the L-curve are different. The L-curve must be *exactly* L-shaped for elbow heuristics, since otherwise the curvature can be large far from the "true" elbow. On the contrary, the slope heuristics assumes a linear behavior of the empirical risk as a function of D_m (or $\mathscr{C}_m$) only for large models (Algorithms 2 and 6), and makes an even milder assumption with its 'jump' formulation (Algorithms 1 and 5; see Theorem 1). Third, theoretical grounds are much stronger for the slope heuristics (with strong optimality results like Theorem 1 in several settings) than for elbow heuristics which is even proved inconsistent in some realistic cases [Engl and Grever, 1994, Vogel, 1996, Hanke, 1996].

Overall, we consider the slope heuristics and its generalization (minimal penalties) as a simple and principled way to localize an elbow on the L-curve (when there is one), and to make use of it for optimal model/estimator selection. In particular, a natural answer to the problem of choosing the scale at which the L-curve should be considered on the x-axis is given by Section 3.6: it should be (the shape of) the minimal penalty.

6.5. Scree test and related methods

For choosing the number of factors in factor analysis, or the number of components in principal components analysis, some classical methods can be related to minimal-penalty and L-curve algorithms.

Scree test The most popular one —named the scree test— has been proposed by Cattell for factor analysis [Cattell, 1966, Cattell and Vogelmann, 1977]. It is based upon the "scree plot", that is, a plot of the eigenvalues versus their rank (in decreasing order), which can be seen as an L-curve for factor analysis. Cattell's key remark [Cattell, 1966] is that the scree plot ends with

a linear part —a scree—, and that the beginning of the linear part corresponds to the "correct" number of factors. Overall, the scree test chooses a number of factors equal to the rank of the starting point of the linear part at the end of the scree plot.

This idea is close to the "slope" formulation of minimal-penalty algorithms (Algorithms 2 and 6). By analogy, we can say that the starting point of the linear part in the scree plot is a "minimal regularization level" —an upper bound on the number of factors that should be kept at the end. This fits well the goal of the initial paper [Cattell, 1966], which is not to find the exact true number of factors —a quantity which might be impossible to define formally—, but only to keep a number of factors which explain 95% to 99% of the "substantive variance". Nevertheless, the scree test seems to be often used for estimating the "true number of factors" itself [Jackson, 1993].

Similarly to minimal-penalty and L-curve algorithms, making use of the scree test requires to overcome practical issues: Cattell remarks that "even a test as simple as this requires the acquisition of *some* art in administering it" [Cattell, 1966]. For instance, it seems important to normalize the data [Cattell, 1966], and sometimes the scree plot ends with two or three linear parts —then, one should cut at the beginning of the first linear part [Cattell, 1966].

Variants Several variants of the scree test exist. The number of factors can for instance be given by the intersection of the scree plot with some reference curve, which corresponds to some "average scree plot" obtained with "pure noise", as proposed by Horn [Horn, 1965, Horn and Engstrom, 1979] and by Frontier's broken stick method [Frontier, 1976].

Another variant exists for estimating the intrinsic dimension of some data set, inside a classification procedure [Bouveyron et al., 2015b]. Given a decreasing sequence of eigenvalues $(\lambda_{(j)})_{1 \leqslant j \leqslant n}$, the estimated intrisic dimension is the smallest j such that $\lambda_{(j)} - \lambda_{(j+1)} \leqslant T$ for some threshold T, which can be chosen by cross-validation in [Bouveyron et al., 2015b]. The underlying assumption is that the scree plot is L-shaped, and that the point where the discrete derivative goes below T corresponds to the "elbow", or to the beginning of the linear part.

Results All these methods are only validated by numerical experiments [Cattell and Vogelmann, 1977]. For instance, for principal component analysis, [Jackson, 1993] concludes that the broken stick method is one of the two best methods for choosing the number of components. The scree test tends to overestimate by one the number of components according to [Jackson, 1993], which is consistent with our remark above that it corresponds to a "minimal regularization level" —not an optimal one.

Note however that some theoretical results are proved for the closely related problem of low-rank matrix recovery from noisy data by hard-thresholding of singular values. In an asymptotic framework, when the goal is to minimize the asymptotic mean squared error in some specific asymptotic regime, [Gavish and Donoho, 2014] shows that the optimal hard threshold can be written $\lambda_\star(m/n)\sqrt{n}\sigma$ when the matrix to recover is of size $m \times n$. Interestingly, the "minimal threshold", which corresponds to the largest singular value obtained from pure noise, is asymptotically equal to $(1+\sqrt{m/n})\sqrt{n}\sigma$. In the framework of [Gavish and Donoho, 2014], the scree test would correspond to using the "minimal hard threshold", and it seems indeed reasonable to use it for estimating the rank (the number of factors). On the contrary, when the goal is to

minimize some quadratic error, the optimal threshold is a bit larger: [Gavish and Donoho, 2014, Figure 4] shows that

$$\forall \beta \in (0,1], \qquad \lambda_\star(\beta) > 1 + \sqrt{\beta}\,.$$

6.6. *Thresholding under the null*

A related approach, for choosing the threshold λ of thresholding estimators, starts by considering the minimal value $\widehat{\lambda}_{\min}$ of the threshold such that the estimator is equal to zero. Under the null hypothesis —that is, when the true signal is zero—, $\widehat{\lambda}_{\min}$ corresponds to the *minimal* thresholding level, and any good threshold must be larger than $\widehat{\lambda}_{\min}$ under the null. For instance, in the setting of the previous paragraph —that is, singular-values hard thresholding—, $\widehat{\lambda}_{\min}$ is of order $(1+\sqrt{m/n})\sqrt{n}\sigma$ and [Gavish and Donoho, 2014] provides an explicit formula for the optimal threshold, which can be written $c(m/n)\widehat{\lambda}_{\min}$ for some $c(m/n) > 1$.

In a general setting, [Giacobino et al., 2017] defines the quantile universal threshold (QUT) λ^{QUT} as the $1-\alpha$ quantile of $\widehat{\lambda}_{\min}$ under the null hypothesis for some $\alpha \in (0,1)$. It turns out that QUT corresponds to the universal threshold proposed for wavelet thresholding [Donoho et al., 1995], and it can be used much more generally, beyond hard or soft thresholding. For instance, for choosing the regularization parameter λ of the Lasso and related procedures in high-dimensional regression, good performance can be obtained with $\lambda = c\lambda^{\mathrm{QUT}}$ for some $c > 1$ [Giacobino et al., 2017, Section 4.2].

QUT and Algorithm 1 have common points: they both start by identifying a minimal value for the parameter of interest (λ or C), then multiply it by a constant factor to get an optimal value of the parameter. Their main difference lies in the definition of the minimal parameter value: it is obtained from data under the null-hypothesis for QUT, hence requiring to know —or at least to approximate— the null-hypothesis distribution, while Algorithm 1 defines it directly from the data, whatever their distribution.

Note that Section 7.5 details a procedure by [Rozenholc, 2012], which is a variant of minimal-penalty algorithms for change-point detection, and can also be seen as a null-hypothesis based calibration procedure, hence similar to QUT.

6.7. *Other model/estimator-selection procedures*

Many other model or estimator-selection procedures exist and are studied in the literature. A detailed account on these is far beyond the scope of this survey. This subsection only mentions a few of them, that are of interest in relation with minimal-penalty algorithms.

Unknown variance First, in addition to the procedures based upon C_p and C_L that are listed in Section 6.3, some procedures are specially built for dealing with the problem of not knowing the noise variance in regression, which is also what minimal-penalty algorithms do in the regression case. We refer to [Giraud et al., 2012] for a detailed survey on high-dimensional variable-selection methods when the variance is unknown. See also [Baraud et al., 2014] for a general estimator-selection procedure based upon similar ideas.

Cross-validation and resampling An important family of general-purpose estimator-selection procedures is cross-validation [Arlot and Celisse, 2010], and more generally all resampling-based selection procedures —e.g., resampling-based penalties, see [Arlot, 2009] and references therein. Comparing them to minimal-penalty algorithms is interesting at least in two distinct situations.

First, when Algorithm 5 works with pen_0 and pen_1 known but $C^\star$ is unknown —for instance, for linear estimators in regression with the least-squares risk— resampling-based procedures are natural competitors, that can be used either for choosing the constant in front of pen_1, or directly for the initial estimator-selection problem. Then, minimal-penalty algorithms have a clear advantage over resampling, because of their much smaller computational cost (see Section 7.2), while they have comparable or better statistical performance according to both theoretical and experimental results, as shown for instance by [Arlot and Bach, 2011].

Second, when Algorithm 5 works with some unknown pen_0 and/or pen_1, an option mentioned in Remark 3 is to estimate them by resampling. Then, the computational cost of Algorithm 5 is comparable to that of cross-validation and other resampling strategies applied to the initial estimator-selection problem. In such cases, the interest of using minimal penalties is the precise non-asymptotic calibration of the constant in front of the resampling-based penalty, which is not guaranteed when using the theoretical value for this constant, since it is often based upon asymptotic considerations. In addition, the conjecture detailed in Section 8.4 suggests another reason for combining resampling and minimal penalties in such frameworks.

7. Some practical remarks

This section discusses several practical questions about the use of minimal-penalty algorithms. A more detailed study of some of them can be found in [Baudry et al., 2012].

7.1. Several definitions for $\widehat{C}$

The minimal-penalty estimator $\widehat{C}$ of the constant that should be put in front of the penalty pen_1 can be defined in several ways, which leads to the practical issue of choosing one among these definitions. Two main approaches are proposed in the previous sections.

Jump approach First, $\widehat{C}$ can be defined as the position $\widehat{C}_{\mathrm{jump}}$ of "the unique large jump" of $C \mapsto \mathscr{C}_{\widehat{m}^{(0)}_{\min}(C)}$, as in Algorithms 1, 3, 4, and 5. Section 2.5 suggests two ways to formally define $\widehat{C}_{\mathrm{jump}}$: choosing the maximal jump $\widehat{C}_{\mathrm{window}}(\eta)$ over a geometric window $[C/(1+\eta), C(1+\eta)]$, and choosing the value $\widehat{C}_{\mathrm{thr.}}(T_n)$ of C for which $\mathscr{C}_{\widehat{m}^{(0)}_{\min}(C)}$ goes under some threshold T_n. Another natural option is to choose the position of the maximal jump

$$\widehat{C}_{\mathrm{maxj.}} \in \operatorname*{argmax}_{C \geqslant 0} \left\{ \mathscr{C}_{\widehat{m}^{(0)}_{\min}(C^-)} - \mathscr{C}_{\widehat{m}^{(0)}_{\min}(C^+)} \right\},$$

that is, taking $\lim_{\eta \to 0} \widehat{C}_{\mathrm{window}}(\eta)$.

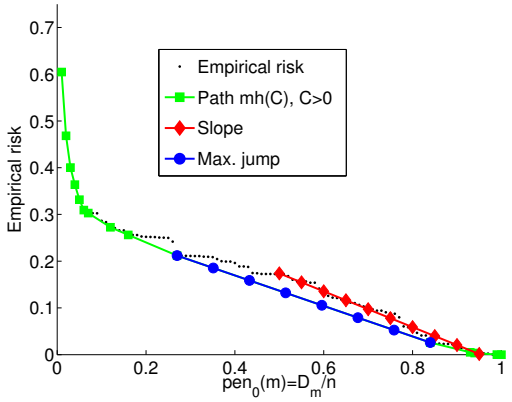


FIGURE 4. *Connection between Algorithms 1 and 2: Plot of* $D_m \mapsto n^{-1}\|Y - \widehat{F}_m\|^2$ *and visualization of* $\widehat{C}_{\text{slope}}$. *'Easy' setting, see Appendix D for details.*

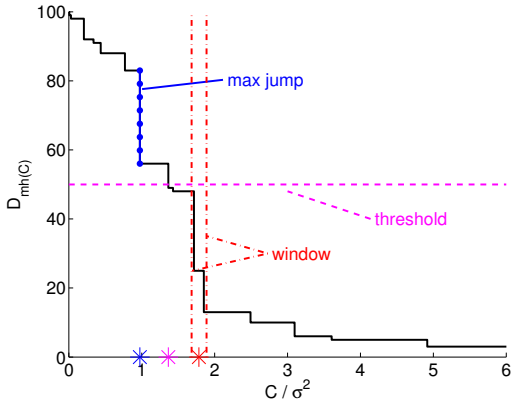


FIGURE 5. *Plot of* $C \mapsto D_{\widehat{m}(C)}$ *and visualization of the three versions of Algorithm 1 on the same sample. 'Easy' setting, see Appendix D for details. The sample chosen here is not typical at all (see Table 1) but illustrates well the differences between* $\widehat{C}_{\text{max j.}}$, $\widehat{C}_{\text{window}}$, *and* $\widehat{C}_{\text{thr.}}$.

Slope approach Second, $\widehat{C}$ can be defined as $\widehat{C}_{\text{slope}}$, the opposite of the estimated value of the slope of the empirical risk as a function of pen_0, as in Algorithms 2 and 6. This approach can be formalized in several ways, using ordinary or robust linear regression, either over a fixed range $[p_{\min}, p_{\max}]$ of values of pen_0, or with the method $\widehat{m}_{\text{CAPUSHE}}$ proposed by [Baudry et al., 2012, Section 4.2], which is based upon a stability study of the selected estimator and depends on some parameter $pct \in (0,1]$.

Note that $\widehat{C}_{\text{window}}$, $\widehat{C}_{\text{thr.}}$, and $\widehat{C}_{\text{slope}}$ all depend on some hyperparameter (η, T_n, $p_{\min}$ and $p_{\max}$, pct). We refer to Appendix D.2 for more details on each definition of $\widehat{C}$ considered in this section.

Theoretical comparison Let us first compare theoretically the various definitions of $\widehat{C}$. For $\widehat{C}_{\text{jump}}$, when there is a single large jump in $C \mapsto \mathscr{C}_{\widehat{m}^{(0)}_{\min}(C)}$, as illustrated by Figure 2, reasonable choices for T_n and η make $\widehat{C}_{\text{window}}$ very close to $\widehat{C}_{\text{max j.}} = \widehat{C}_{\text{thr.}}$. On the contrary, when the phase transition around the minimal penalty yields several jumps of medium size in $C \mapsto \mathscr{C}_{\widehat{m}^{(0)}_{\min}(C)}$, as in Figure 5 for instance, $\widehat{C}_{\text{max j.}}$, $\widehat{C}_{\text{thr.}}$, and $\widehat{C}_{\text{window}}$ can take quite different values and lead to selecting different models. Theoretical guarantees such as Theorem 1 do not exclude such a situation, even asymptotically, so they only apply to $\widehat{C}_{\text{thr.}}$ and $\widehat{C}_{\text{window}}$ with T_n and η of the correct order of magnitude.

For the slope approach, no theoretical guarantee is available, but the linear behavior of $\widehat{\mathscr{R}}_n(\widehat{s}_m)$ as a function of $\text{pen}_0(m)$ is supported theoretically from expectation computations, as detailed in Sections 2.2–2.3 and 3.3–3.4.

The maximal jump and threshold definitions with $T_n = \overline{\mathscr{C}} := (\max_m \mathscr{C}_m + \min_m \mathscr{C}_m)/2$ coincide when the largest jump is of size at least $(\max_m \mathscr{C}_m - \min_m \mathscr{C}_m)/2$. This condition always holds true if no $m \in \mathscr{M}$ has a complexity $\mathscr{C}_m \in (\overline{\mathscr{C}}; \max_m \mathscr{C}_m)$, which often occurs for computational

reasons, since estimators with complexity $\mathscr{C}_m > \overline{\mathscr{C}}$ usually are hard to compute and known to be suboptimal.

The jump and slope approaches can seem quite different at first sight, but they actually are the two sides of the same coin. Section 2 shows that reasoning from the same computations, Eq. (7)–(8), can lead to a heuristic justification of both approaches. Another argument enlightens the similarity of the jump and slope approaches. By Proposition 5 and its proof in Appendix B.1, the path $(\widehat{m}^{(0)}_{\min}(C))_{C>0}$ is piecewise constant, $\widehat{m}^{(0)}_{\min}(C) = m_i$ for $C \in [C_i, C_{i+1})$, and the sequences $(m_i)_{0 \leqslant i \leqslant i_{\max}}$ and $(C_i)_{0 \leqslant i \leqslant i_{\max}}$ can be visualized on the L-curve $(\mathrm{pen}_0(m), \widehat{\mathscr{R}}_n(\widehat{s}_m))_{m \in \mathscr{M}}$: the angles of the lower convex envelope of the L-curve exactly correspond to the m_i, $0 \leqslant i \leqslant i_{\max}$, and

$$C_i = \frac{\widehat{\mathscr{R}}_n(\widehat{s}_{m_i}) - \widehat{\mathscr{R}}_n(\widehat{s}_{m_{i-1}})}{\mathrm{pen}_0(m_{i-1}) - \mathrm{pen}_0(m_i)}$$

is the opposite of the slope of the segment joining m_{i-1} to m_i on the L-curve. So, $\widehat{C}_{\mathrm{max\,j.}}$ can be visualized on the L-curve, as illustrated by Figure 4. Given the L-curve (black dots), draw its (piecewise linear) lower convex envelope (green squares), localize the widest segment —in terms of values of $\mathscr{C}_m$, which is usually proportional to $\mathrm{pen}_0(m)$—: its slope is $-\widehat{C}_{\mathrm{max\,j.}}$. Then, one clearly see why $\widehat{C}_{\mathrm{max\,j.}}$ is often close to $\widehat{C}_{\mathrm{slope}}$ in the setting of Figure 4: for a random point cloud with a linear trend of slope $\approx -C^\star$ for large abscissa values, estimating its slope by linear regression is almost equivalent to looking at the slope of the longest segment of its lower convex envelope. Note that $\widehat{C}_{\mathrm{thr.}}$ can be visualized on the L-curve similarly to $\widehat{C}_{\mathrm{max\,j.}}$.

This direct comparison emphasizes the respective drawbacks of $\widehat{C}_{\mathrm{max\,j.}}$ and $\widehat{C}_{\mathrm{slope}}$. When the amplitude of the largest jump is small, $\widehat{C}_{\mathrm{max\,j.}}$ is not a reliable estimation of $C^\star$, see Figure 9b in Appendix C.

When some large models have a significantly positive approximation error, as in the 'hard' setting described in Appendix D —see the right of Figure 10— they pollute the slope estimation and make $\widehat{C}_{\mathrm{slope}}$ biased, unless only a few such models are present and robust regression is used. In the latter case, since $\widehat{C}_{\mathrm{jump}} \in \{\widehat{C}_{\mathrm{max\,j.}}, \widehat{C}_{\mathrm{window}}, \widehat{C}_{\mathrm{thr.}}\}$ only depends on the *lower convex envelope* of the L-curve, even a large number of "polluting" models will not influence $\widehat{C}_{\mathrm{jump}}$ at all, making it more robust.

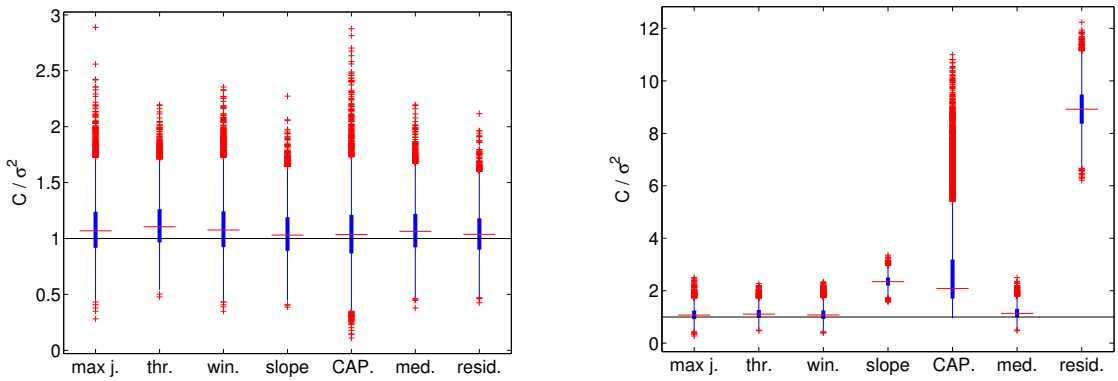
This difference between $\widehat{C}_{\mathrm{jump}}$ and $\widehat{C}_{\mathrm{slope}}$ also appears in the assumptions made for their theoretical and heuristic justifications. In Section 2, $\widehat{C}_{\mathrm{slope}}$ requires the approximation error to be almost constant over *all* large models, whereas $\widehat{C}_{\mathrm{jump}}$ only assumes that *two* models exist with a small approximation error, one of moderate complexity and one of large complexity.

Experimental comparison In addition to the above theoretical comparison, we report the results of new simulation experiments for variable selection in least-squares regression. We consider two settings: in the 'easy' setting, the order between variables is known, while in the 'hard' setting, two possible orders (the correct one and its converse) are considered alternatively, making half of the models very bad. All details about simulation experiments (data generation, model collection, and exact implementation for each definition of $\widehat{C}$) are given in Appendix D.

First, since the beginning of this section outlines strong theoretical connections between the different definitions of $\widehat{C}$, a natural question is: how different are the models finally selected,

Configuration	All equal	Exactly 4 equal	At least 3 equal	All different	$\widehat{C}_{\mathrm{max\,j.}} = \widehat{C}_{\mathrm{thr.}}$	Max, thr, and win different
Frequency ('easy')	0.524	0.238	0.967	$< 10^{-3}$	0.777	0.009
Frequency ('hard')	0.134	0.118	0.894	$< 10^{-3}$	0.769	0.008

TABLE 1. *Frequency of various configurations for the models $\widehat{m}$ selected by Algorithm 1 with $\widehat{C}_{\mathrm{max\,j.}}$ ('max'), $\widehat{C}_{\mathrm{thr.}}$ ('thr'), $\widehat{C}_{\mathrm{window}}$ ('win'), by Algorithm 2 ($\widehat{C}_{\mathrm{slope}}$) and by $\widehat{m}_{\mathrm{CAPUSHE}}$. 'Easy' and 'hard' settings, see Appendix D for details.*

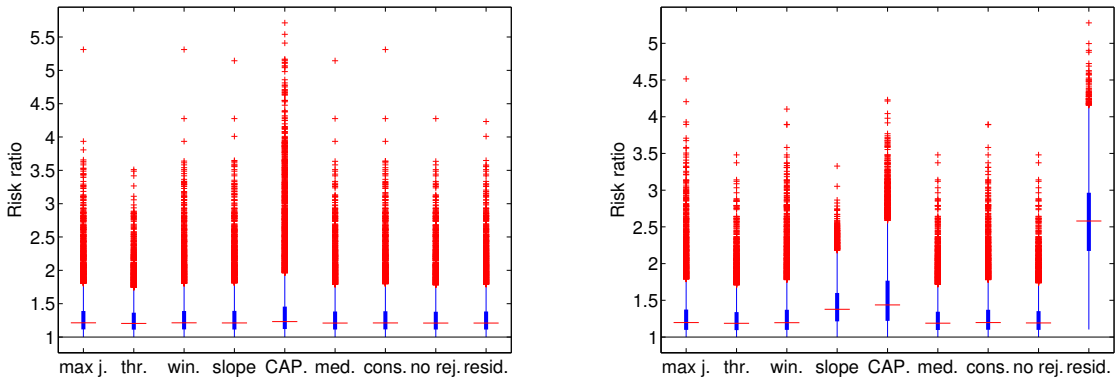


(a) 'Easy' setting. (b) 'Hard' setting.

FIGURE 6. *Distribution over 10000 independent samples of $\widehat{C}/\sigma^2$ for seven estimators $\widehat{C}$ of σ^2: $\widehat{C}_{\mathrm{max\,j.}}$ ('max j.'), $\widehat{C}_{\mathrm{thr.}}$ ('thr.'), $\widehat{C}_{\mathrm{window}}$ ('win.'), $\widehat{C}_{\mathrm{slope}}$ in Algorithm 2 ('slope'), $\widehat{C}_{\mathrm{CAPUSHE}}$ ('CAP.'), the median of $\{\widehat{C}_{\mathrm{max\,j.}}, \widehat{C}_{\mathrm{thr.}}, \widehat{C}_{\mathrm{window}}, \widehat{C}_{\mathrm{slope}}, \widehat{C}_{\mathrm{CAPUSHE}}\}$ ('med.'), and $\widehat{\sigma}^2_{m_0}$ defined by Eq. (70) ('resid.'). See Appendix D for details.*

depending on the definition taken for $\widehat{C}$? Table 1 shows that they all coincide most of the time in the 'easy' setting (with a clear single large jump, as for the sample of Figures 2 and 4), and they globally agree 90% of the time or more in both settings. The probability of a total disagreement is very small (less than 0.1%) even if it sometimes occurs, as illustrated by Figure 5, where $\widehat{C}_{\mathrm{max\,j.}}$, $\widehat{C}_{\mathrm{thr.}}$, and $\widehat{C}_{\mathrm{window}}$ respectively lead to selecting $\widehat{m} = 14$, 11, and 7; Figure 9b in Appendix C shows a similar configuration. Similar conclusions are obtained by [Arlot and Massart, 2009, Section 3.3] about $\widehat{C}_{\mathrm{max\,j.}}$ and $\widehat{C}_{\mathrm{thr.}}$.

Second, since our experiments consider projection estimators in least-squares regression, all minimal-penalty based $\widehat{C}$ estimate σ^2, so they can be compared to $\widehat{\sigma}^2_{m_0}$ —defined by Eq. (70)— as estimators of the residual variance σ^2. Results are provided in Figure 6, as well as Tables 2–3 in the Appendix, where several values of parameters on which some of the $\widehat{C}$ depend are compared. In the 'easy' setting (Figure 6a), all methods behave similarly and as expected from theoretical arguments: the distribution of $\widehat{C}$ is asymmetric around σ^2, with smaller deviations below σ^2 than above σ^2, as in the bounds of Proposition 3. Such an asymmetry is a good property in terms of model-selection performance, as suggested by Figure 8 in Section 8.4 for instance. The order of magnitude of the deviations of $\widehat{C}_{\mathrm{jump}}/\sigma^2$ from Proposition 3 is $\mathscr{B}(c_n)/\sigma^2 + \sqrt{\log(n)/n}$ with $c_n = n/3$ (for $\widehat{C}_{\mathrm{window}}$) or $T_n/2$ (for $\widehat{C}_{\mathrm{thr.}}$); in our experiments, with $n = 100$ and $T_n = n/2$, we get $\mathscr{B}(c_n)/\sigma^2 \in [0.04, 0.08]$, and $\sqrt{\log(n)/n} \approx 0.2$, so the constants appearing in Proposition 3 here



(a) 'Easy' setting. (b) 'Hard' setting.

FIGURE 7. *Distribution over* 10 000 *independent samples of* $\|\widehat{F}_{\widehat{m}} - F\|^2 / \inf_{m \in \mathscr{M}} \|\widehat{F}_m - F\|^2$ *for* $\widehat{m} = \widehat{m}(2\widehat{C})$ *with* $\widehat{C}$ *among the seven estimators* $\widehat{C}$ *compared in Figure 6, and for* $\widehat{m}$ *obtained by majority vote among* $\{\widehat{m}(2\widehat{C}_{\max \text{j.}}), \widehat{m}(2\widehat{C}_{\text{thr.}}), \widehat{m}(2\widehat{C}_{\text{window}}), \widehat{m}(2\widehat{C}_{\text{slope}}), \widehat{m}_{\text{CAPUSHE}}\}$ *with* $\widehat{m}(2\widehat{C}_{\text{window}})$ *as a default choice ('cons.') or considering only the samples on which such a majority exists ('no rej.'). See Appendix D for details.*

are pessimistic.

The most variable $\widehat{C}$ clearly is $\widehat{C}_{\text{CAPUSHE}}$, but to be completely fair, we must notice that the procedure proposed by [Baudry et al., 2012] only outputs a selected model $\widehat{m}_{\text{CAPUSHE}}$ and we make an arbitrary choice for defining some $\widehat{C}_{\text{CAPUSHE}}$ from the definition of $\widehat{m}_{\text{CAPUSHE}}$ (see Appendix D.2). Among other definitions of $\widehat{C}$, $\widehat{C}_{\max \text{j.}}$, and $\widehat{C}_{\text{window}}$ are slightly more variable than the others but the difference is mild.

Interesting differences occur in the 'hard' setting, which is designed as a case example for difficult situations for $\widehat{C}_{\text{slope}}$, $\widehat{C}_{\text{CAPUSHE}}$, and $\widehat{\sigma}^2_{m_0}$. As expected, $\widehat{C}_{\text{slope}}$ completely fails because of the wide amplitude of the approximation error among large models, and $\widehat{\sigma}^2_{m_0}$ behaves totally differently depending on the parity of m_0: $\widehat{\sigma}^2_{m_0}$ is worse than $\widehat{C}_{\text{slope}}$ when S_{m_0} is one of the 'bad' models, while it works well when S_{m_0} is one of the 'good' models (see Figure 10 and Table 3 in the Appendix). This failure of $\widehat{C}_{\text{slope}}$ and $\widehat{C}_{\text{CAPUSHE}}$ —when $(S_m)_{m \in \mathscr{M}}$ is the union of subcollections having different approximation properties— is also reported by [Baudry, 2009, Section 4.5], [Devijver et al., 2015, Figure 4], and [Devijver, 2017b, Figures 3–4] in realistic settings. The nested algorithm presented in Section 7.3 might be a way to fix this issue, even if it has not been tested yet in such situations.

More generally, depending on the setting, choosing the parameter for one definition of $\widehat{C}$ can be a big practical issue. For instance, Tables 2–3 in the Appendix show that $\widehat{C}_{\text{thr.}}$ is sensitive to the choice of T_n. Even if $T_n = n/2$ works well for the 'easy' and 'hard' settings, it is certainly not a universally good choice, and changing F, n or σ^2 could easily make it fail compared to other definitions of $\widehat{C}$. Similarly, the performance of $\widehat{C}_{\text{slope}}$ strongly depends on the parameters $p_{\min}, p_{\max}$ and choosing them from data is not an easy task, a problem also reported in the change-point detection setting [Lebarbier, 2002, Chapter 4]. A reasonable option is given by $\widehat{m}_{\text{CAPUSHE}}$ [Baudry et al., 2012, Section 4.2], and it works reasonably well in the 'easy' setting, but it fails in the 'hard' setting as expected.

Third, the model-selection performance of all these procedures is assessed by Figure 7 and by

Tables 2–3 in the Appendix.

At first order, the conclusions are similar to the ones obtained for estimating σ^2. All definitions of $\widehat{C}$ work well in the 'easy' setting. In the 'hard' setting, $\widehat{\sigma}^2_{m_0}$ completely fails, while $\widehat{m}_{\text{CAPUSHE}}$ and $\widehat{C}_{\text{slope}}$ do slightly worse than the other formulations of the slope heuristics algorithm.

The detailed comparison of the procedures that work well is a bit different: the model-selection performance (risk ratios) are not ordered exactly as the mean-squared errors in Tables 2–3. The main reason is that risk estimation —which reduces to estimating σ^2 in our setting— is different from model selection [Breiman and Spector, 1992]. Figure 8 in Section 8.4 shows at least one reason for this difference: overpenalizing slightly, that is, overestimating σ^2 a bit, improves the model-selection performance. According to Figure 8, the best overpenalization factor is 1.12 in the 'easy' setting. For instance, Table 2 shows that taking $D_{m_0} = n/10$ for $\widehat{\sigma}^2_{m_0}$ leads to better model-selection performance than $D_0 = n/2$ in the 'easy' setting, even if $D_0 = n/2$ yields a much better estimator of σ^2.

Note however that for a given bias (as an estimator of σ^2), the best model-selection performance is obtained when the variance is the smallest: compare for instance $\widehat{C}_{\text{slope}}$ with $D_0 = n/2$ and CAPUSHE in the 'easy' setting (Table 2).

Let us finally mention that previous simulation experiments in various settings have compared some of the definitions of $\widehat{C}$. In short, almost all of them report that $\widehat{C}_{\text{max j.}}$ is less reliable —because of the event on which there is not a single large jump [Maugis, 2008, Fig 8.11] [Baudry et al., 2012, Section 5], which happens more or less often— compared to $\widehat{C}_{\text{thr.}}$ [Arlot and Bach, 2011, Solnon et al., 2012], $\widehat{C}_{\text{window}}$ [Bontemps and Toussile, 2013], and $\widehat{C}_{\text{slope}}$ or $\widehat{C}_{\text{CAPUSHE}}$ [Maugis and Michel, 2011a, Baudry et al., 2012, Connault, 2011] [Baudry, 2009, Section 3.2] [Roche, 2014, Tab 2.1]. Only [Devijver and Gallopin, 2018] reports similar performances for $\widehat{C}_{\text{max j.}}$ and $\widehat{C}_{\text{CAPUSHE}}$. Nevertheless, $\widehat{C}_{\text{max j.}}$ remains useful for confirming the choice made with another definition of $\widehat{C}$ [Baudry et al., 2012, Connault, 2011], with a visual check that there is a single large jump. The slope approach can also fail for reasons detailed previously in this subsection [Baudry, 2009, Devijver et al., 2015, Devijver, 2017b]. [Lebarbier, 2005] even shows that $\widehat{C}_{\text{max j.}}$ and $\widehat{C}_{\text{slope}}$ can both fail, which motivates a modified algorithm —called "calibrated method"— for change-point detection; note that [Arlot et al., 2012] fixes this precise failure by using the slope heuristics with a penalty shape depending on two constants, as detailed in Section 7.4.

Conclusion on the choice of $\widehat{C}$ First, it is not surprising to have to choose among several definitions of $\widehat{C}$ or to choose some hyperparameter such as η, T_n or $p_{\min}$, because of no free lunch theorems: no fully automatic estimation procedure can work uniformly well over all statistical problems [Devroye et al., 1996, Chapter 7]. An expert advice is always necessary at some point. For minimal-penalty algorithms, our suggests join the ones of [Baudry et al., 2012, Connault, 2011]: never use a single definition of $\widehat{C}$ in a blind way, either by considering several definitions for $\widehat{C}$ or by checking visually that there is a clear complexity jump and/or that the L-curve exhibits a clear linear trend on the data. When computing all values of $(\widehat{\mathscr{R}}_n(\widehat{s}_m), \text{pen}_0(m), \text{pen}_1(m), \mathscr{C}_m)_{m \in \mathscr{M}}$ is too expensive, one should also take into account the computational cost of the procedure, as discussed in Section 7.2.

We propose the following (semi-automatic) approach for using several definitions $\widehat{C}_{\text{max j.}}, \widehat{C}_{\text{thr.}},$

$\widehat{C}_{\text{window}}$, $\widehat{C}_{\text{slope}}$, and $\widehat{C}_{\text{CAPUSHE}}$ of $\widehat{C}$ simultaneously. If the goal is to estimate σ^2, take their median. If the goal is estimator selection, compute the five corresponding estimator choices $\widehat{m}$, and make a majority vote: if at least three over five coincide, take their common value, otherwise, output a warning and ask the user to look at the complexity jump and the L-curve. When the five methods disagree, using a completely different approach remains a good option, for instance, cross-validation. The results of using this strategy (in a fully automatic way since our experiments need to be reproducible) are reported in Figures 6–7 as well as Tables 2–3 in the Appendix, showing good performances in all settings.

Finally, the above comparison also points out several risky choices for $\widehat{C}$ (in addition to $\widehat{\sigma}^2_{m_0}$): $\widehat{C}_{\text{max j.}}$ without checking that there is indeed a single large jump, $\widehat{C}_{\text{thr.}}$ with a bad choice for T_n, the "naive" version $\widehat{C}_{\text{slope}}$ of the slope approach, and $\widehat{C}_{\text{slope}}$ or $\widehat{C}_{\text{CAPUSHE}}$ when selecting among a union of subcollection of estimators that may have different approximation properties.

7.2. Algorithmic cost

When all empirical risks can be computed Let us assume that the values of the empirical risk $\widehat{\mathscr{R}}_n(\widehat{s}_m)$, the minimal and optimal penalty shapes $\text{pen}_0(m)$ and $\text{pen}_1(m)$, and the complexity $\mathscr{C}_m$ for all $m \in \mathscr{M}$ are stored in memory. Then, the computational complexity of minimal-penalty algorithms is the following.

For Algorithm 5, computing the full path $(\widehat{m}^{(0)}_{\text{min}}(C))_{C\geqslant 0}$ requires at most $\mathscr{O}((\text{card}(\mathscr{M}))^2)$ operations —as shown in Appendix B.1— and much less in practice. Indeed, denoting by $i_{\max}+2 \leqslant \text{card}(\mathscr{M})$ the cardinality of this path, it can be computed with $\mathscr{O}(i_{\max}\,\text{card}(\mathscr{M}))$ operations.

Furthermore, depending on the definition of $\widehat{C}_{\text{jump}}$, it might not be necessary to compute the full path. For instance, with the threshold approach, using the notation of Appendix B.1, if $i(T_n)$ is such that $\widehat{C}_{\text{thr.}} = C_{i(T_n)}$, only $\mathscr{O}(i(T_n)\,\text{card}(\mathscr{M}))$ operations are necessary, and usually $i(T_n) \ll i_{\max} \ll \text{card}(\mathscr{M})$.

Computing $\widehat{C}_{\text{window}}$ as defined in Algorithm 5 might seem costly at first sight. Appendix B.2 shows that given the path $(\widehat{m}^{(0)}_{\text{min}}(C))_{C\geqslant 0}$, of cardinality $i_{\max}+2$, computing $\widehat{C}_{\text{window}}$ can be done with at most $\mathscr{O}(i_{\max}\log(i_{\max}))$ operations.

Finally, step 3 of Algorithm 5 requires at most $\mathscr{O}(\text{card}(\mathscr{M}))$ operations. Overall, Algorithm 5 always has a complexity $\mathscr{O}(i_{\max}\,\text{card}(\mathscr{M})) \leqslant \mathscr{O}((\text{card}(\mathscr{M}))^2)$.

For Algorithm 6, step 1 is a (robust) linear regression —hence of computational cost $\mathscr{O}(\text{card}(\mathscr{M}))$— and step 2 can be done with $\mathscr{O}(\text{card}(\mathscr{M}))$ operations. Note that $\widehat{m}_{\text{CAPUSHE}}$ has a larger computational cost since it requires to run $\mathscr{O}(\text{card}(\mathscr{M}))$ times Algorithm 6, hence a total cost of $\mathscr{O}(\text{card}(\mathscr{M})^2)$.

When computing all empirical risks is not tractable In general, most of the computational complexity of computing $\widehat{m}_{\text{Alg.5}}$ or $\widehat{m}_{\text{Alg.6}}$ corresponds to computing $\widehat{\mathscr{R}}_n(\widehat{s}_m)$ for *all* $m \in \mathscr{M}$. For instance, for density estimation with Gaussian mixture models [Maugis and Michel, 2011a], performing maximum-likelihood estimation in several large models involves a large computational cost, while we know that all corresponding estimators are always bad. Can we remove from the collection $(\widehat{s}_m)_{m\in\mathscr{M}}$ most estimators with $\mathscr{C}_m$ "large", without degrading too much the performance of Algorithms 5–6?

For the jump approach, two estimators having a small approximation error are needed to get a jump, as with the assumptions of Theorem 1: one of large complexity, one much less complex. If we are not sure of which estimators have a small enough approximation error, considering more than two of them can be helpful; otherwise, this does not hurt —and we conjecture that this slightly decreases the variance of $\widehat{C}_{\text{jump}}$—, without being mandatory.

For the slope approach, the picture is different. Having only two estimators with a small approximation error implies making a linear regression over the corresponding two points, which is very close to the residual-based estimator $\widehat{\sigma}^2_{m_0}$ defined by Eq. (70), as shown by Figure 10. Therefore, $\widehat{C}_{\text{slope}}$ with only a few large-complexity estimators faces the risk that some of them have a large approximation error, to which it will be quite sensitive, unlike $\widehat{C}_{\text{jump}}$ (see Figure 10b). Using a robust regression in $\widehat{C}_{\text{slope}}$ decreases the risk but does not exclude it totally, as shown by the poor results of $\widehat{m}_{\text{CAPUSHE}}$ in our experiments in the 'hard' setting in Section 7.1.

A more reliable strategy for the slope approach is to consider only estimators of complexity up to $\mathscr{C}_{\max}$, and to carefully check that $\mathscr{C}_{\max}$ is large enough by visualizing the linear relation between the empirical risk and pen_0. This can be done easily with the CAPUSHE package [Baudry et al., 2012]. Experiments in [Baudry et al., 2012, Section 5] show as expected that $\widehat{C}_{\text{slope}}$ is better —more stable— when $\mathscr{C}_{\max}$ is large enough. Similarly, for change-point detection, [Lebarbier, 2002, Chapter 4] and [Lebarbier, 2005, Section 4.2] study the influence of such a bound $\mathscr{C}_{\max}$ on $\widehat{C}_{\max \text{j}.}$ and propose a heuristic method —called "calibrated"— for choosing $\mathscr{C}_{\max}$ from data.

Note finally that in some frameworks, well-chosen large-complexity estimators are easy to compute. For instance, in fixed-design regression, the estimator equal to the original data always has an empirical risk and an approximation error equal to zero —see assumption (**HId**) in Theorem 1.

7.3. Nested minimal-penalty algorithm

In a framework where $\mathscr{M}$ is a cartesian product $\mathscr{M}_1 \times \mathscr{M}_2$, Devijver, Gallopin and Perthame [Devijver et al., 2017] propose a "nested slope heuristics" algorithm, that we here generalize straightforwardly to Algorithms 5–6. The idea is to choose $\widehat{m} = (\widehat{m}_1, \widehat{m}_2) \in \mathscr{M}_1 \times \mathscr{M}_2$ in two steps. First, for every $m_1 \in \mathscr{M}_1$, select one estimator among $(\widehat{s}_{(m_1,m_2)})_{m_2 \in \mathscr{M}_2}$ with a minimal-penalty algorithm; the selected index is denoted by $\widehat{m}_2(m_1)$. Then, select one estimator among $\left(\widehat{s}_{(m_1,\widehat{m}_2(m_1))}\right)_{m_1 \in \mathscr{M}_1}$ with a minimal-penalty algorithm. The numerical experiments of [Devijver et al., 2017] on some transcriptomic data analysis problem show that such a nested algorithm can work, for choosing a number m_1 of clusters (of individuals) and a partitioning m_2 of the features (the genes) used for inferring a cluster-dependent gene regulatory network.

7.4. Estimation of several unknown constants in the penalty

When the optimal penalty involves several unknown constants, that is,

$$\forall m \in \mathscr{M}\,, \qquad \text{pen}_{\text{opt}}(m) = C_1^\star \,\text{pen}_1^{(1)}(m) + \cdots + C_k^\star \,\text{pen}_1^{(k)}(m) \tag{81}$$

for some known $\text{pen}_1^{(1)}, \ldots, \text{pen}_1^{(k)}$, the slope approach can be generalized, using linear regression for estimating simultaneously $C_1^\star, \ldots, C_k^\star$. The idea has first been proposed with Algorithm 2

by Lebarbier [Lebarbier, 2002, Section 4.3.2] in the case of change-point detection, where the optimal penalty depends on $k = 2$ constants.

It has since been used —with good numerical performances— in several settings: change-point detection [Arlot et al., 2012], joint variable selection and clustering via Gaussian mixture models [Meynet and Maugis-Rabusseau, 2012], principal curves estimation [Biau and Fischer, 2012], and unsupervised segmentation of spectral images via piecewise-constant Gaussian mixture models [Cohen and Le Pennec, 2014].

Nevertheless, no theoretical guarantees are currently available for such an algorithm. In addition to the practical issues already mentioned for the slope approach, this procedure is difficult to apply when there is not a single natural complexity measure $\mathscr{C}_m$ but several of them —$\mathrm{pen}_1^{(1)}(m), \ldots, \mathrm{pen}_1^{(k)}(m)$ can be k complexity measures—, which have to be combined wisely for defining what are the "complex enough" $m \in \mathscr{M}$ over which the (robust) linear regression should be done. Another major difficulty is when $\mathrm{card}(\mathscr{M})$ or n are not large enough to allow a good estimation of several constants $C_1^\star, \ldots, C_k^\star$ simultaneously.

Another option is to make use of a simplified penalty shape —depending on a single multiplicative constant—, even when we know that it differs from the optimal shape given by Eq. (81). Several articles make use of such a simplified penalty, instead of trying to calibrate $k = 2$ constants, with satisfactory numerical results: for density estimation/clustering with Gaussian mixture models [Maugis and Michel, 2011a] —see also [Michel, 2008, App. C.2]— or multinomial mixture models [Derman and Le Pennec, 2017], for choosing a simplicial complex in the computational geometry field [Caillerie and Michel, 2011], and for selecting jointly the rank and a set of variables in a high-dimensional finite mixture regression model [Devijver, 2017a].

A numerical comparison between simplified penalty shape and calibration of two constants is done in a few other papers, with various conclusions: favorable to the simplified shape [Lebarbier, 2002, Section 4.3.2, for change-point detection with Gaussian noise], similar for both methods [Devijver and Gallopin, 2018, for inference of an high-dimensional Gaussian graphical model], or favorable to the calibration of two constants —for change-point detection with Laplace noise and a simplified shape derived from experiments with Gaussian noise [Lebarbier, 2002, Section 4.6.4], and for curve clustering [Meynet and Maugis-Rabusseau, 2012, Figure 5]. As a conclusion, choosing between these two strategies should be done carefully, depending on the framework.

7.5. Variants for change-point detection

For change-point detection seen as a model-selection problem, three approaches have been proposed, which are not exactly of the form of Algorithm 5, but still are closely related to minimal penalties. The first is the "calibrated method" [Lebarbier, 2005, Section 4.2] mentioned in Section 7.2.

Second, [Lavielle, 2005, Section 2.3] remarks that $\widehat{C}_{\mathrm{max\,j.}}$ often leads to underestimating the number of changes. Then, using the notation of Appendix B.1, it is proposed instead to define $\widehat{m}$ as the largest $m_i = \widehat{m}(C_i)$ corresponding to a jump whose height $C_i - C_{i-1}$ is much larger than the one of the largest subsequent jump, that is, $\max_{j>i}\{C_j - C_{j-1}\}$. This approach might be closer to an elbow heuristics (see Section 6.4) than to a minimal-penalty algorithm.

Statistical Base Jumping Third, an unpublished idea by Rozenholc [Rozenholc, 2012] is the following. Assume that change-point detection is cast as a model-selection problem in fixed-design regression, so that we can use the notation of Section 2. Take some D 'large' but, say, smaller than $n/2$, for instance $D = n/\log(n)$ or $D = \sqrt{n}$. Compute $\widehat{F}_D$ the empirical risk minimizer over the set of piecewise-constant signals with D pieces (which is a union of $\binom{n-1}{D-1}$ vector spaces of dimension D). Consider the residual vector $\widetilde{Y} = Y - \widehat{F}_D$ and apply the penalization approach to this pseudo-data, that is, compute

$$\forall C > 0\,, \quad \widetilde{m}(C) \in \operatorname*{argmin}_{m \in \mathscr{M}_n^{\mathrm{chpt}}} \left\{ \frac{1}{n} \left\| \widetilde{Y} - \Pi_m \widetilde{Y} \right\|^2 + C \operatorname{pen}_0(m) \right\} \tag{82}$$

where the model collection $\mathscr{M}_n^{\mathrm{chpt}}$ and the penalty shape pen_0 are the ones adapted to change-point detection, see [Comte and Rozenholc, 2004, Lavielle, 2005, Lebarbier, 2005]. Define $\widehat{C}_{\mathrm{SBJ}}$ as the minimal value of $C > 0$ such that $D_{\widetilde{m}(C)} = 1$, and finally select

$$\widehat{m}_{\mathrm{SBJ}} \in \operatorname*{argmin}_{m \in \mathscr{M}_n^{\mathrm{chpt}}} \left\{ \frac{1}{n} \|Y - \Pi_m Y\|^2 + \frac{2\widehat{C}_{\mathrm{SBJ}}}{1 + \frac{D}{n}} \operatorname{pen}_0(m) \right\}.$$

Note that $\widehat{F}_D$, $(\widetilde{m}(C))_{C \geqslant 0}$ and $\widehat{m}_{\mathrm{SBJ}}$ can all be computed efficiently, in particular using dynamic programming. The heuristics behind this method is that if D is large enough to catch all true change-points of F in $\widehat{F}_D$, then $\widetilde{Y}$ does not contain any signal anymore, and $\widehat{C}_{\mathrm{SBJ}} \operatorname{pen}_0$ is the minimal penalization level needed to recover with Eq. (82) the unique model of dimension one (constant signal). The factor $1 + \frac{D}{n}$ dividing $\widehat{C}_{\mathrm{SBJ}}$ corrects for the variance of the pseudo-sample $\widetilde{Y}$. Unpublished experiments [Rozenholc, 2012] suggest that $\widehat{m}_{\mathrm{SBJ}}$ provides very good segmentations, much better than with the original slope heuristics —that is, Algorithm 1 or 2, as done in [Lebarbier, 2005] for instance.

7.6. Other uses of minimal penalties

Let us finish this section by mentioning two other uses of a minimal-penalty algorithm in the literature.

Choice of a penalty shape Algorithm 5 can be used for choosing among several penalty shapes, by detecting bad ones, which are the ones that do not lead to a clear dimension jump, as illustrated by [Baudry et al., 2012, Section 3 of supplementary material] in the setting of [Caillerie and Michel, 2011].

Minimal-penalty assisted experiments For estimating a good deterministic constant $\widetilde{C}^\star$ to be put in front of the penalty, [Chagny, 2013] computes on 100 samples the constant $\widehat{C}$ chosen by a slope heuristics algorithm, and defines $\widetilde{C}^\star$ as the maximal value of $\widehat{C}$ observed over the 100 samples. The main interest of this approach is to require less computations than the standard one —which would be to compute, for every C in a grid, the average over the 100 samples of the risk of the estimator selected with the penalty $C \operatorname{pen}_1$, and then to take $\widetilde{C}^\star$ that minimize the average risks over C in the grid—, even if the value $\widetilde{C}^\star$ might not be the optimal one.

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

8. Conclusion, conjectures, and open problems

As a conclusion of this survey, we sketch what is known theoretically for minimal-penalty algorithms, as well as several conjectures and open problems of high interest. Let us recall that Section 5 provides some hints for tackling many of these conjectures and open problems.

8.1. Settings and losses where minimal-penalty algorithms apply

Let us sketch the set of frameworks for which minimal-penalty algorithms are theoretically justified, at least partially (see Section 4).

The typical situation is a polynomial collection of minimum-contrast estimators (also called empirical risk minimizers) with a regular contrast [Saumard, 2010b] (for instance, the least-squares contrast), the corresponding risk (expected value of the contrast), and (almost) i.i.d. data. Then, all obtained results show that $\mathrm{pen}_{\mathrm{opt}} \approx 2\,\mathrm{pen}_{\mathrm{min}}$. In other terms, the slope heuristics holds true in several frameworks "close to" choosing among a polynomial collection of projection estimators with the least-squares risk and i.i.d. data.

The general result by Saumard [Saumard, 2010b, Chapters 7–8] for regular contrasts suggests that the slope heuristics is probably valid in all frameworks that are close enough to this ideal situation —e.g., least-squares estimators in regression or (conditional) density estimation with the least-squares risk—, under appropriate assumptions.

Two results, in regression [Arlot and Bach, 2009] and in density estimation [Lerasle et al., 2016], show that linear estimators can be considered instead of minimum-contrast estimators, at the price of changing the slope heuristics (Section 2) into a minimal-penalty heuristics (Section 3). A similar extension can probably be done in other settings for estimators that are (close to) be linear functions of (part of) the data.

More generally, using the notation introduced in Section 4.1, it is probably often true that $\mathbb{E}[p_2(m)]$ is a minimal penalty and $\mathbb{E}[p_1(m)+p_2(m)]$ an optimal penalty. But unless these expectations are (approximately) known up to a multiplicative constant, applying such results requires to estimate $\mathbb{E}[p_2(m)]$ and $\mathbb{E}[p_1(m)+p_2(m)]$ by resampling (see Remark 3), and we then loose a nice feature of Algorithm 5 which is its small computational cost compared to cross-validation.

8.2. Unavoidable assumptions

Even in settings for which a full proof of a minimal-penalty algorithm is known, a natural question to ask is which assumptions are unavoidable for this algorithm to work.

Based upon existing proofs (in particular the one of Theorem 1, which is typical), we conjecture that at least three assumptions are (almost) needed.

First, a "complex" estimator should be present in the collection, similarly to (**HId**). Note that such an estimator can often be added on purpose to a predefined collection.

Second, one "less complex" but "good" estimator should also be present in the collection, similarly to what Theorem 1 assumes implicitly. The exact definition of "good" can depend on the context. In general, being consistent should suffices; note that assuming that the oracle estimator is consistent is a mild assumption for estimator selection, since otherwise the problem

is not much interesting. In the setting of Theorem 1, it suffices to have a model with a small approximation error, even if the corresponding estimator is not consistent.

Third, it seems reasonable to make some mild moment assumption on the data so that the key quantities p_1 and p_2 concentrate around their expectations, at least when using deterministic penalty shapes. Nevertheless, a Gaussian assumption such as (**HG**) is not necessary [Arlot and Massart, 2009, Saumard, 2013]. Independence of data is not necessary either [Lerasle, 2011, Garivier and Lerasle, 2011]. Risk bounds could be obtained under much weaker moment assumptions —for instance, when the noise only has a finite moment of order two—, for empirical risk minimizers [Mendelson, 2018] or for robust estimators [Audibert and Catoni, 2011, e.g.,]. Nevertheless, we are not aware of any theoretical result on minimal penalties in such a setting.

8.3. Other settings, losses, estimators

Numerical experiments show that minimal-penalty algorithms can be used fruitfully in many other settings such as supervised classification [Zwald, 2005], model-based clustering [Baudry, 2015, Maugis and Michel, 2011a], high-dimensional inference [Devijver and Gallopin, 2018], change-point detection [Lebarbier, 2005, Bardet et al., 2012], topological data analysis [Caillerie and Michel, 2011], functional linear models [Roche, 2014] or Hawkes-process intensity estimation [Reynaud-Bouret and Schbath, 2010], with applications in various domains such as biology —genomics [Akakpo, 2011, Reynaud-Bouret and Schbath, 2010], transcriptomics [Rau et al., 2015, Devijver and Gallopin, 2018], quantitative trait prediction from genomic data [Devijver et al., 2017], population genetics [Bontemps and Toussile, 2013]—, energy —electricity consumption prediction [Devijver et al., 2015], oil production modelization [Michel, 2008, Chapter 6]—, hyperspectral image segmentation [Cohen and Le Pennec, 2014], text analysis [Derman and Le Pennec, 2017], and bike sharing systems [Bouveyron et al., 2015a, Godichon-Baggioni et al., 2019]. Combining these numerical experiments —especially the ones showing that the empirical risk is indeed close to a linear function of some known pen_0 for large-complexity estimators— with the partial theoretical results available (Section 4), several settings can be identified where we conjecture that a minimal-penalty algorithm such as Algorithms 5–6 could be used fruitfully. Among them, we select below the most challenging ones, in terms of both practical applications and theoretical interest.

8.3.1. Supervised classification

A classical setting where minimal-penalty algorithms would be quite useful is supervised classification with the 0–1 loss and corresponding empirical risk minimizers. No theoretical result is available up to now, and it seems tough to prove any because the 0–1 contrast is far from being regular. Nevertheless, [Boucheron and Massart, 2011] provides a key ingredient of the proof, that is, a concentration inequality for $p_2(m)$ that applies easily to 0–1 classification, even when fast learning rates are possible. Given Proposition 1 and the general strategy detailed in Section 5, what remains is to prove a similar concentration inequality for $p_1(m)$, and to be able to estimate (up to the same unknown constant) $\mathbb{E}[p_1(m)]$ and $\mathbb{E}[p_2(m)]$.

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

A probably easier open problem is to provide theory for the case of classification with a convex loss, such as the logistic loss —at the basis of logistic regression— or the hinge loss —at the basis of support vector machines. At least, for the hinge loss, the numerical experiments of [Zwald, 2005, Section 6.4.3] suggest that minimal-penalty algorithms can work with $\mathrm{pen}_0(m) = \mathscr{C}_m = D_m$ and $\mathrm{pen}_1(m) = 2D_m$.

8.3.2. Model-based clustering, choice of the number of clusters

Estimating the number of clusters for (unsupervised) clustering is another problem where fine tuning of penalties is a major challenge. A classical approach —called model-based clustering— is to estimate the data density by maximum-likelihood on a mixture model, and to define clusters by a maximum a posteriori rule. Then, the number of clusters can be chosen by maximizing the penalized log-likelihood. Minimal-penalty algorithms with $\mathrm{pen}_0(m) = \mathscr{C}_m = D_m$ and $\mathrm{pen}_1(m) = 2D_m$ are shown successful by experiments on synthetic and real data, for various problems following this strategy (up to modifications that are specified below):

- clustering with Gaussian [Baudry, 2015], Poisson [Rau et al., 2015], multinomial [Derman and Le Pennec, 2017], or some functional [Bouveyron et al., 2015a] mixture models.
- clustering with Gaussian mixtures and the *conditional* log-likelihood instead of the log-likelihood, which leads to slightly different kinds of clusters [Baudry, 2015].
- joint clustering and variable selection —identifying which features are relevant for clustering— with mixtures of Gaussian [Maugis and Michel, 2011a] or multinomial [Bontemps and Toussile, 2013] variables.
- efficient joint clustering and high-dimensional variable selection with maximum-likelihood estimators trained on *data-driven models* obtained by a first step of L^1 penalization [Meynet and Maugis-Rabusseau, 2012]; here, the shape of the minimal penalty seems to be close to a linear combination of D_m and $D_m \log \frac{p}{D_m}$ [Meynet and Maugis-Rabusseau, 2012, Fig. 6], and the algorithm described in Section 7.4 can be used.
- when an additional feature vector is provided for each observation, clustering a mixture of Gaussian regression models, jointly done with feature selection [Devijver, 2017b, Devijver et al., 2015] or partitioning [Devijver et al., 2017], via two-steps procedures similar to the one of [Meynet and Maugis-Rabusseau, 2012].

We conjecture that minimal-penalty algorithms indeed work in these settings, that is, as one can observe on synthetic or real data: (i) $p_2(m)$ is (close to) a linear function of the number of parameters D_m, (ii) Algorithm 5 or 6 with $\mathrm{pen}_0(m) = D_m$ and $\mathrm{pen}_1(m) = \alpha D_m$ provides an estimator with a small Kullback-Leibler risk, for some $\alpha > 1$ to be determined, and (iii) the number of clusters selected by this algorithm is equal to the true one $K^\star$ with large probability when n is large and the data distribution is close to a mixture with $K^\star$ components. Note that (ii) is a density estimation guarantee —hence, slightly different from clustering, but classical for justifying theoretically a penalty shape [Derman and Le Pennec, 2017, Maugis and Michel, 2011b, Bontemps and Toussile, 2013, Meynet and Maugis-Rabusseau, 2012, Devijver, 2017a]— and that (ii) and (iii) may require different values of α since estimation and model identification are different goals for model selection, see Section 8.3.6. When variable selection is done jointly with clustering, (iii) can be completed by the fact that the true set of relevant variables is selected with large probability.

Up to now, only oracle inequalities with theoretical penalties are available in some of these settings. Of course, proving the above conjecture would be less difficult for pure model-based clustering [Baudry, 2015, Rau et al., 2015, Derman and Le Pennec, 2017, Bouveyron et al., 2015a] than for the two-steps algorithm using L^1-penalized maximum-likelihood for defining data-driven models [Meynet and Maugis-Rabusseau, 2012, Devijver, 2017b, Devijver et al., 2015].

Note also that variable selection, without any order among variables, means that the model collection considered is large —with the terminology of Section 4.7—, at least implicitly; this fact raises specific issues that are addressed in Section 8.3.4.

8.3.3. High-dimensional statistics

Hyperparameter tuning is a major issue for high-dimensional statistics [Giraud, 2014, Chapter 5], which can often be addressed by penalization. Despite several positive numerical results, providing a full theoretical proof of a minimal-penalty algorithm in this context remains an open problem.

Clustering Numerical results about minimal-penalty algorithms for joint (model-based) clustering and variable selection are reviewed in Section 8.3.2. For block-diagonal estimation of the covariance matrix of a high-dimensional Gaussian vector (graphical model), [Devijver and Gallopin, 2018] provides positive numerical results for a similar algorithm —maximum likelihood on data-driven models obtained by thresholding the empirical covariance matrix. Nevertheless, given the difficulty of proving an oracle inequality [Devijver and Gallopin, 2018, for instance], it seems hard to obtain a full theoretical validation of the minimal-penalty algorithms of [Meynet and Maugis-Rabusseau, 2012, Devijver, 2017b, Devijver and Gallopin, 2018].

Regression: the Lasso and related algorithms One of the most classical high-dimensional statistics problem is variable selection in linear regression, for which the (group) Lasso and related algorithms are popular. Penalized least-squares can be used for choosing their parameters, thanks to covariance penalties [Efron, 2004], which have a simple expression of the form $\sigma^2 \mathrm{df}/n$ when the noise is Gaussian, where σ^2 denotes the residual noise-level and df are the degrees of freedom. Easy-to-compute estimators of df exist for the Lasso [Tibshirani and Taylor, 2012, Dossal et al., 2013] and group Lasso [Vaiter et al., 2012], among others. The remaining issue is to estimate σ^2 without knowing any small correct model, for which minimal penalties are a natural approach (see Section 6.1).

Connault's Ph.D. thesis [Connault, 2011] provides an extensive numerical study of minimal penalties for calibrating the Lasso or least-squares estimators trained on models selected by the Lasso ('Lasso+LS'). In short, the major difficulty is that the natural candidate for $\mathrm{pen}_0(m)$, that is $\mathbb{E}[p_2(m)]/\sigma^2$, cannot be used because it depends on the (unknown) signal $\beta^\star$. Simplified penalty shapes —that is, $\mathbb{E}[p_2(m)]/\sigma^2$ for a zero-signal, or for a zero-signal and an identity design matrix— often work for Lasso+LS, and sometimes for the Lasso, depending on the signal-to-noise ratio and on the sparsity of the signal. This is not satisfactory because these minimal-penalty algorithms sometimes fail for the Lasso or Lasso+LS, and [Connault, 2011] proposes "antidotes" for detecting the failure but nothing for correcting it, except using cross-validation.

It nevertheless seems possible to solve the case of an orthogonal design matrix, when the Lasso is soft thresholding and Lasso+LS is hard thresholding. Taking the number of selected variables k as tuning parameter, Loubes and Massart [Loubes and Massart, 2004, section 2] conjecture that for soft thresholding, $\mathbb{E}[p_2(k)] \approx \frac{3}{2}\frac{\sigma^2 k}{n}$. If one could prove this conjecture, a minimal-penalty algorithm could be used for estimating σ^2 —hence for calibrating soft (or hard) thresholding. Section 8.3.4 discusses the case of hard thresholding.

For a general design matrix and for other estimators, we think that the key question is to find the good parametrization of the estimator to be calibrated. For instance, the Lasso can be parametrized by the regularization parameter or by the number of selected variables, and [Connault, 2011] shows that the theoretical minimal penalty $\mathbb{E}[p_2(m)]$ and the performance of minimal-penalty algorithms strongly depend on the chosen parametrization. We also conjecture that the solution might not come from a direct application of Algorithms 5–6, but by the more general approach of identifying an observable phase transition —with respect to some well-chosen parameter— that provides the key information for an optimal calibration of the algorithm considered —for instance, an estimation of σ^2 for the (group) Lasso. This idea has already been proposed in a few settings that are detailed in Section 8.5. For high-dimensional regression, theoretical results prove the existence of phase transitions in the risk of the Lasso [Bellec, 2018, Section 5] [Bellec, 2017, Section 4] and of the constrained formulation of the Lasso [Chatterjee, 2014, Section 2.1]. Nevertheless, it is not clear whether these phase transitions are observable, so they might not be useful for choosing hyperparameters. The approach of Section 6.6 seems to be another promising direction for tackling this problem.

Let us finally mention that concentration inequalities for $p_1(m)$ are available for the Lasso and some related algorithms —see Section 5.3 for details. They can be useful for validating minimal-penalty algorithms.

8.3.4. Large collection of models

As recalled in Section 4.7, the nature of the model-selection problem depends on the size of the model collection. All full proofs and almost all partial results available for minimal-penalty algorithms are for small collections, for which optimal model selection can be obtained with the unbiased risk estimation heuristics. For large collections, only a few partial theoretical results are available, as reported in Section 4.7. Therefore, the case of large collections remains a widely open problem of major interest.

The most classical situation is variable selection among $p \gtrsim n$ variables, which amounts to select among a collection of 2^p models. Let us start by focusing on the two settings where minimal-penalty algorithms are best understood: (i) variable selection with $p = n$ and an orthonormal design —so that penalizing the least-squares criterion by a function of the number of variables is equivalent to hard thresholding—, and (ii) change-point detection —finding the locations of abrupt changes in the distribution of a sequence of n observations— which can be casted as a variable-selection problem with $p = n - 1$ variables —the possible breakpoint locations— and solved by penalized least-squares.

Let us also recall that the notations (β), (β'), (γ), and $(\widetilde{\gamma})$ refer to partial results about minimal-penalty algorithms; they are defined in Sections 4.1–4.4.

Orthonormal variable selection by hard thresholding For variable selection with an orthonormal design and Gaussian noise, Birgé and Massart [Birgé and Massart, 2007, Proposition 2] prove (β') and $(\widetilde{\gamma})$ with a minimal penalty of order $2\sigma^2 \frac{D_m}{n} \log \frac{n}{D_m}$ —at least for $1 \ll D_m \ll n$. We conjecture that (β) holds true in the same setting: a proof of (β^-) derives from the proof of [Birgé and Massart, 2007, Proposition 2], and (β^+) seems easy to obtain given the results of [Birgé and Massart, 2007]. Then, (β) and $(\widetilde{\gamma})$ would prove that Algorithm 5, with $\mathrm{pen}_0(m) \approx D_m \log \frac{n}{D_m}$ and $\mathrm{pen}_1(m)$ given by [Birgé and Massart, 2007, Section 3.1.3], provides a good data-driven variable-selection procedure, satisfying an oracle inequality close to being optimal.

The main remaining challenge for having a full proof of a first-order optimal procedure is problem (γ): find a first-order optimal penalty of the form $\sigma^2 \mathrm{pen}_1(m)$ with $\mathrm{pen}_1(m)$ known. We think this is a hard problem, whose resolution would have a great impact on model-selection theory in general, since even the *value* of the optimal excess risk of such a variable-selection procedure is not exactly known at first order. We only know by minimax arguments that it should be of order $\log \frac{n}{D_{m^\star}}$ times the oracle excess risk, up to a constant factor. Given the numerical experiments of [Lebarbier, 2005, Arlot et al., 2012], we conjecture that the ratio between the optimal and minimal penalty belongs to $(1,2]$; it may be model-dependent.

Another open problem for orthonormal variable selection is to determine the minimal penalty for non-Gaussian noise. Contrary to small model collections, this is not a straightforward extension of Gaussian results since the experiments of [Lebarbier, 2005, Section 5] for another large collection problem —change-point detection— suggest that the minimal penalty then is different for Laplace and for Gaussian noise.

Change-point detection Penalized least-squares is a classical approach to change-point detection, for which an oracle inequality $(\widetilde{\gamma})$ holds true for Gaussian noise and a penalty of the form $\sigma^2 \frac{D_m}{n} \left[c_1 \log \frac{n}{D_m} + c_2 \right]$ where $c_1, c_2 > 0$ are two numerical constants and σ^2 is the residual noise-level [Lebarbier, 2005]; a similar result in a slightly different setting is proved by [Arlot et al., 2012]. The numerical experiments of [Lebarbier, 2005, Arlot et al., 2012, Sorba, 2017] and the partial theoretical results of [Sorba, 2017] —see Section 4.7— suggest that a minimal-penalty algorithm with $\mathscr{C}_m = D_m$ and $\mathrm{pen}_0(m)$ proportional to a linear combination of D_m and $D_m \log \frac{n}{D_m}$ —or close to it, see [Birgé and Massart, 2007, Sorba, 2017]— should work well in this setting. Proving it formally would require to solve two open problems: (a) prove the existence of a dimension jump for some known pen_0 —or, equivalently, prove that step 2 of Algorithm 6, or its generalization of Section 7.4, works well—, and (b) prove an optimal oracle inequality for a penalty $C^\star \mathrm{pen}_1$ that can be derived from the minimal penalty. Problem (b) is very hard, as for orthonormal variable selection. Problem (a) seems less difficult: [Sorba, 2017, Chapter 8] is close to proving it, but there is still a gap between "large enough" and "too small" penalties, which leaves open the possibility of having no clear dimension jump. It remains a challenge, as emphasized by the fact that the shape of the minimal penalty seems to depend on the noise distribution not only through its variance [Lebarbier, 2005, Section 5]. Note that proving (a) with $C^\star = \sigma^2$ would be sufficient to get a good data-driven penalty for change-point detection, since an oracle inequality —maybe suboptimal— is already available for a penalty depending on σ^2 and known quantities.

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

To conclude on change-point detection via penalized least-squares, let us recall that [Rozenholc, 2012] proposes a related but different approach to penalty tuning —see Section 7.5— that might be even more efficient than minimal-penalty algorithms for change-point detection with penalized least-squares. Justifying it theoretically would therefore be of great interest.

Beyond penalized least-squares, slope heuristics algorithms empirically work well for change-point detection with dependent data in two settings: causal processes with the maximizer of a penalized log-likelihood [Bardet et al., 2012], and long-memory processes with the minimizer of a local Whittle contrast [Bardet and Guenaizi, 2018]. [Bardet et al., 2012] even shows the remarkable fact that minimal penalties numerically adapt to variations of the optimal constant $C^\star$ —which can be of order $\log(n)$ or $\sqrt{n}$— when the dependence structure varies. A theoretical validation of these results seems quite a challenge, since handling small model collections in the same settings already is an open problem.

General setting Understanding minimal-penalty algorithms for more general variable-selection problems seems a too high theoretical challenge for the next few years. We nevertheless conjecture that minimal-penalty algorithms work well beyond orthonormal variable selection and change-point detection, given the successful experiments of [Maugis and Michel, 2011a, Bontemps and Toussile, 2013] for joint variable selection and model-based clustering (see Section 8.3.2).

Minimal-penalty algorithms experimentally work well in several other settings mentioned previously, where large model collections are implicitly considered through L^1 penalization or thresholding: model-based clustering [Meynet and Maugis-Rabusseau, 2012], multivariate regression with a mixture of linear models [Devijver, 2017b, Devijver et al., 2017], and Gaussian graphical model estimation [Devijver and Gallopin, 2018].

8.3.5. Infinite estimator collections

Throughout the paper, the estimator collection $(\widehat{s}_m)_{m\in\mathscr{M}}$ is assumed to be finite. Nevertheless, Algorithms 5–6 can still be used for some infinite collections (at least theoretically, due to computational issues) that behave as small (finite) collections, with the terminology of Sections 4.7 and 8.3.4.

Indeed, [Arlot and Bach, 2011] proves that Algorithm 5 can be used for selecting a tuning parameter within a *continuous set* $\mathscr{M}$. The proof of [Arlot and Bach, 2011] for kernel ridge regression mostly relies on two facts: (i) Algorithm 5 would work for a collection of Cn^α such estimators for any fixed $C, \alpha > 0$, (ii) the collection of kernel ridge regressors can be well approached by a finite collection of Cn^α estimators for some fixed $C, \alpha > 0$. We conjecture that a similar approach can be used for proving that Algorithm 5 works with some other continuous collections, starting by multiple kernel ridge regression with a fixed number of kernels.

8.3.6. Model selection for identification of the true model

Throughout this survey, we assume that the goal is to choose a data-driven $\widehat{m} \in \mathscr{M}$ such that the risk of $\widehat{s}_{\widehat{m}}$ is minimal, that is, satisfies a nonasymptotic oracle inequality (2). Model selection

can target a different goal, which is to identify the smallest true model $m^\star \in \mathscr{M}$ with probability one asymptotically, assuming that some true model exists; a procedure $\widehat{m}$ achieving this goal is said "model-consistent". Then, the exact same procedure cannot achieve the two goals in general [Yang, 2005]. Can minimal-penalty algorithms still be useful for identification? Some experiments and theoretical arguments suggest a positive answer.

Role of the size of the model collection For large collections —see Section 8.3.4—, it turns out that both estimation and identification require to overpenalize compared to the unbiased risk estimation principle. Therefore, the minimal-penalty algorithms suggested in Section 8.3.4 should also work for identification. This conjecture is supported by the experiments of [Bontemps and Toussile, 2013, Fig 2] about variable selection in multinomial mixture models, and by those of [Arlot et al., 2012, Fig 2] and [Garreau and Arlot, 2018, Fig 5–6] about change-point detection.

For small collections, the picture is different. A typical example is least-squares fixed-design regression with projection estimators, as in Section 2. Let us focus on this setting here for simplicity. The slope heuristics then leads to a model-selection procedure equivalent to C_p, which is first-order optimal for estimation but inconsistent for identification [Shao, 1997, Theorem 1]. A simple way to fix this failure is to replace the factor 2 between minimal and optimal penalties in the slope heuristics by, say, a $\log(n)$ factor, in order to get a BIC-type penalty, hence consistent for identification [Shao, 1997, Theorem 2]. Such a correction of the slope heuristics may seem unsatisfactory, so one may consider to combine it with a procedure choosing from data between AIC and BIC-type penalties [Yang, 2005, van Erven et al., 2012].

Change-point detection For change-point detection with well-chosen small collections of models, [Gey and Lebarbier, 2008, Akakpo, 2011, Durot et al., 2009] propose specific hybrid procedures for identification of the change-point locations. In short, they consist in two-steps procedures, with minimal-penalty algorithms in both steps. The first step selects within a small model collection, providing an oversegmentation of the data sequence. The second step removes the unnecessary change-points. Proving the consistency of these procedures —including the minimal-penalty algorithms— is an open problem. Another natural question is to generalize such two-steps procedures to other model-selection problems with an identification goal.

Minimal penalties for consistent identification Let us finally mention some theoretical results about the minimal level of penalization needed for model consistency, that is, for having $\widehat{m} = m^\star$ a.s. asymptotically. Even without a corresponding calibration algorithm —since the minimal penalty is not observable here—, this question remains of interest for theory.

In least-squares regression, [Shao, 1997, Theorem 1 (iii)] shows that C_p is not model consistent, assuming only that some true model $m' \in \mathscr{M}$ exists with $D_{m^\star} < D_{m'} \leqslant D_{m^\star} + \kappa$ for some fixed $\kappa > 0$. Such a result actually holds for any penalty of the form CD_m/n with $C \geqslant 0$ fixed as n grows, which can be proved from arguments used in the proof of Theorem 1. Conversely, using a penalty of the form $\lambda_n D_m/n$ with $\lambda_n \to +\infty$ and $\lambda_n/n \to 0$ as n tends to infinity provides a model-consistent procedure [Shao, 1997, Theorem 2]. Therefore, the minimal level of penalization for identification is of the form $\lambda_n D_m/n$ with $\lambda_n \to +\infty$.

For maximum-likelihood estimators, at least two results are available. BIC-type penalties are minimal for estimating the order of a Markov chain without any prior upper bound on its order

[van Handel, 2011]. For density estimation with i.i.d. data, identifying the true model among a nested family by minimizing the log-likelihood penalized by $\mathrm{pen}(m) = f(m)g(n)$ requires $g(n) > C^\star(s^\star) \log\log n$ for some constant $C^\star(s^\star) > 0$ [Gassiat and Van Handel, 2013].

8.3.7. Miscellaneous

Model selection Numerical experiments suggest that minimal-penalty algorithms work well for several other model-selection problems. We list them below, in order to help identifying settings where new theoretical results could be proved:

1. *Heteroscedastic regression* when the residual variance is known up to a constant —which can occur for inverse problems—, with least-squares risk and estimators [Villers, 2007, Section 2.6.2], beyond regressograms and strongly localized bases for which theoretical results are already known for a random design [Arlot and Massart, 2009, Navarro and Saumard, 2017]. The fixed-design case can be handled similarly to [Arlot and Bach, 2011]. The random-design case with general models is clearly more challenging.
2. Estimation of two kinds of *geometrical objects*, with least-squares risk and estimators: simplicial complices [Caillerie and Michel, 2011] and principal curves [Biau and Fischer, 2012].
3. *Least-squares risk and estimators* for Hawkes-process intensity estimation [Reynaud-Bouret and Schbath, 2010], clustering of compositional data [Godichon-Baggioni et al., 2019], and in a functional linear model [Roche, 2014, Section 2.4].
4. *Maximum-likelihood estimators with the Kullback-Leibler loss* for semiparametric regression with censored data via the Cox model [Letué, 2000], point-process intensity estimation [Michel, 2008, Section 6.3 and Appendix D.1.2], regression for counting processes under a proportional-hazard assumption [Oueslati and Lopez, 2013], and segmentation of spectral images via spatialized Gaussian mixtures [Cohen and Le Pennec, 2014].

Estimator selection We emphasize in this survey that minimal-penalty algorithms can be useful for estimator selection in general. Beyond the few theoretical results pointed out in Section 4.2, we conjecture that Algorithms 5–6 work for several estimator-selection problems.

First, numerical experiments suggest that they can be used for selecting among maximum-likelihood or least-squares estimators trained on *data-driven models*, obtained by L^1 penalization in a variable-selection setting [Meynet and Maugis-Rabusseau, 2012, Devijver, 2017b], by thresholding the empirical covariance matrix [Devijver and Gallopin, 2018, Devijver et al., 2017], by k-means [Caillerie and Michel, 2011], or by (kernel) PCA in classification [Zwald, 2005, Section 6.4.3] or functional data analysis [Roche, 2014, Section 2.4]. Some of these results are detailed above in Sections 8.3.2 and 8.3.3.

Second, minimal-penalty algorithms can be used successfully for the pruning step of CART in regression [Gey and Lebarbier, 2008] and of a spatial variant of CART [Bar-Hen et al., 2018], according to numerical experiments.

Finally, Algorithms 5–6 certainly cannot succeed for any kind of estimator collection. Section 8.5 describes a natural and promising way to generalize the minimal-penalty approach beyond Algorithms 5–6.

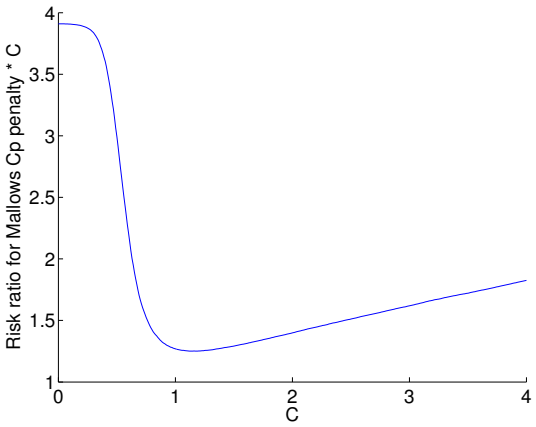


FIGURE 8. *Overpenalization with Mallows' C_p penalty, 'easy' setting (see Appendix D.2 for details). Error bars are so small that they would not be visible on the graph. The optimal overpenalization factor is $C = 1.12$, leading to an improvement by a factor* 1.015 *compared to taking* $C = 1$.

8.4. Overpenalization

It is known empirically that a better model-selection performance can be obtained by overpenalizing a bit: the penalty $C\,\mathrm{pen}^{\mathrm{gal}}_{\mathrm{opt}}(m) = C\mathbb{E}\left[\mathscr{R}(\widehat{s}_m) - \widehat{\mathscr{R}}_n(\widehat{s}_m)\right]$ has optimal performance when C is slightly above 1, see for instance [Arlot and Baudry, 2002], [Arlot, 2007, Chapter 11], and [Arlot, 2009, Section 6.3.2] in the regression setting, and [Arlot and Lerasle, 2016, Figure 3] in least-squares density estimation. A similar phenomenon holds in the experiments of Section 7.1, as shown by Figure 8. For histogram selection in density estimation, [Saumard and Navarro, 2018a] proposes a natural way to overpenalize automatically, which leads to a new corrected version of the AIC criterion. Nevertheless, choosing from data an appropriate overpenalization factor remains an open problem.

For reasons detailed below, we conjecture that minimal penalties can help solving this issue. More precisely, when Algorithms 5–6 are known to be first-order optimal, we conjecture that they automatically overpenalize —by a factor close to 1 when n is large—, and that this overpenalization decreases the risk of the final estimator compared to penalization by $\mathrm{pen}^{\mathrm{gal}}_{\mathrm{opt}}(m)$. Another way to formulate this conjecture, following [Lacour and Massart, 2016] and [Lacour et al., 2017, Section 2.4], is to state that the optimal constant $C^\star$ depends on n differently from what first-order asymptotics suggest, and that Algorithm 5 estimates well the finite-sample value of $C^\star$. Section 2 certainly provides the less difficult setting for proving this conjecture, even if the challenge is high: it requires to analyze penalization procedures at a precision level an order of magnitude higher than ever.

Several results support the above conjecture. First, several simulation experiments show that minimal-penalty algorithms overpenalize slightly in most settings: this is reported by [Villers, 2007, Section 2.6.2.4], [Arlot and Bach, 2011], and [Solnon et al., 2012, Section 6], and this holds for the experiments of Section 7.1 (see Figure 6 in Section 7.1 and Tables 2–3 in the Appendix).

Second, Theorem 1 is consistent with the fact that $\widehat{C}_{\mathrm{jump}}$ might overestimate σ^2. Taking $T_n = n/2$ for instance, Theorem 1 implies that on a large-probability event,

$$(1-\eta_n^-)\sigma^2 \leqslant \widehat{C}_{\mathrm{thr.}} \leqslant (1+\eta_n^+)\sigma^2$$

with $\eta_n^+ > \eta_n^-$; Proposition 3 in Section 6.1 provides a precise statement. If these bounds are tight, it means that $\widehat{C}_{\mathrm{thr.}}$ is slightly biased upwards as an estimator of σ^2, which corresponds to overpenalization.

Third, minimal-penalty algorithms take into account the full collection of estimators $(\widehat{s}_m)_{m\in\mathscr{M}}$ in their definition, and Section 8.3.4 details why they should automatically adapt to the richness of the collection $\mathscr{M}$. We claim that the need for overpenalization might be mostly related to the richness of $\mathscr{M}$, so that the conjectures of Section 8.3.4 could help solving the above overpenalization conjecture. Let us explain briefly why, by considering fixed-design regression with projection estimators, using some results and the vocabulary of Section 4.7. When $\mathscr{M}$ is small —say, one model per dimension—, $\mathrm{pen}_{C_p}(m) = 2\sigma^2 D_m/n$ is an (asymptotically) optimal penalty and the minimal penalty is $\sigma^2 D_m/n$. When $\mathscr{M}$ is large —say, $\binom{n}{D}$ models of dimension D—, the minimal amount of penalization required is multiplied by $1+2\log(n/D_m)$, which is of order $2\log(n) \gg 1$ (except for the largest models), and good performances can be obtained with some penalties of the same order of magnitude. For a given sample size n, between these two extreme settings, there is a continuum of collections $\mathscr{M}$ of increasing sizes, for which the optimal amount of penalization is $C_n^\star(\mathscr{M})\,\mathrm{pen}_{C_p}$ for some $C_n^\star(\mathscr{M})$ between 1 and $2\log(n)$, approximately: this is an instance of the overpenalization phenomenon. So, if a minimal-penalty algorithm adapts to the size of $\mathscr{M}$, it would capture the need for overpenalization by $C_n^\star(\mathscr{M})$ in the constant $\widehat{C}_{\mathrm{jump}}$. At least, $\widehat{C}_{\mathrm{jump}}$ would be asymptotically of the correct order for both small and large $\mathscr{M}$, which cannot be done with some estimator $\widehat{\sigma}^2$ that does not take into account the collection $\mathscr{M}$.

8.5. *Beyond Algorithms 5–6: phase transitions for estimator selection*

Most, if not all, non-parametric estimators depend on one or several parameters, whose optimal data-driven choice is often a challenge. In this survey, we focus on a single parameter C that is a multiplicative constant in front of a penalty, and we show that

(i) an observable phase transition occurs in several settings for the estimator $\widehat{s}_{\widehat{m}^{(0)}_{\min}(C)}$ around $C = C^\star$, and

(ii) $C^\star$ can be used for the optimal calibration of $\widehat{s}_{\widehat{m}^{(1)}_{\mathrm{opt}}(C)}$ through Algorithm 5.

If a similar phenomenon occurs for other types of tuning parameters, this would lead to highly interesting generalizations of Algorithm 5. This subsection collects partial theoretical results — using the notation (β), (β'), (γ) and $(\widetilde{\gamma})$ defined in Sections 4.1–4.4— and experiments going into this direction, as well as several conjectures and open problems.

8.5.1. *Goldenshluger-Lepski's and related procedures*

Goldenshluger-Lepski's method [Goldenshluger and Lepski, 2011, Bertin et al., 2016] is a classical estimator-selection procedure, which does not rely on penalization of an empirical risk but on pairwise comparisons between estimators.

Goldenshluger-Lepski's method For choosing the bandwidth h of kernel density estimators with a fixed kernel, in order to minimize the least-squares risk, [Lacour and Massart, 2016] studies a slightly simplified version of Goldenshluger-Lepski's method, that depends on a single parameter a that can be interpreted as a constant in front of a penalty. If we define a complexity $\mathscr{C}_h = \|K\|^2/(nh)$ as usual for kernel density estimation, [Lacour and Massart, 2016, Theorem 3] proves an equivalent of (β^-) and (β'^-) if $a < 1$, and ($\widetilde{\gamma}$) if $a > 1$. Simulation experiments suggest that there is indeed a phase transition for the selected bandwidth around some value $a^\star$ of a — hence, (β^+) should also hold true— and that the optimal value of a is slightly above $a^\star$. Despite the theoretical results showing that $a^\star \to 1$ as $n \to +\infty$, for a finite sample size $a^\star$ is not necessarily close to 1. This leads to a minimal-penalty algorithm for estimating $a^\star$, hence calibrating Goldenshluger-Lepski's method. The numerical experiments of two papers show the interest of this algorithm, with $\widehat{a}$ defined similarly to $\widehat{C}_{\text{maxj.}}$: for estimating the stationary distribution of a bifurcating Markov chain on $\mathbb{R}^d$ [Bitseki Penda and Roche, 2017], and for state-by-state inference of the emission densities of a hidden Markov model [Lehéricy, 2018]. Although [Lehéricy, 2018] takes a penalty multiplied by $2\widehat{a}$ for defining the final estimator —similarly to the slope heuristics—, [Bitseki Penda and Roche, 2017] takes a just above $\widehat{a}$, by selecting the estimator immediately after the jump of $\mathscr{C}_h$. This choice is supported by the fact that the minimal and optimal penalties are almost equal in the results of [Lacour and Massart, 2016].

Penalized comparison to overfitting (PCO) In the same framework, with a possibly multivariate bandwidth $h \in \mathbb{R}^d$, [Lacour et al., 2017] proposes a new procedure called penalized comparison to overfitting (PCO), which lies between penalization and Goldenshluger-Lepski's method. PCO depends on a parameter λ which is a multiplicative factor in front of one of the two terms of some kind of penalty. Considering again the complexity $\mathscr{C}_h = \|K\|^2/(nh)$, PCO satisfies an equivalent of (β^-) if $\lambda < 0$ [Lacour et al., 2017, Theorems 3–4], (β'^+) if $\lambda > 0$, and (γ) around $\lambda = \lambda^\star = 1$ [Lacour et al., 2017, Theorems 2 and 5]. The optimality of $\lambda^\star = 1$ is assessed by numerical experiments on synthetic data [Varet et al., 2018].

Although PCO does not seem to require a data-driven calibration of λ according to the above result, $\lambda = 1$ may not always be a good choice outside the least-squares density estimation setting. Therefore, the theoretical results of [Lacour et al., 2017] suggest the following minimal-penalty algorithm for calibrating PCO: first, detect $\widehat{\lambda}$ around which $\lambda \mapsto \mathscr{C}_{\widehat{h}(\lambda)}$ jumps, then, take

$$(a) \quad \lambda = \widehat{\lambda} + 1 \qquad \text{or} \qquad (b) \quad \lambda = 2(\widehat{\lambda} + 1) - 1 = 2\widehat{\lambda} + 1$$

for defining the final estimator. Option (a) is suggested by the fact that the difference between the minimal λ —zero— and the optimal λ —one— is equal to 1. Option (b) is suggested by the slope heuristics, since the penalty $\lambda\,\mathrm{pen}_0$ is equivalent to $C\,\mathrm{pen}_0 - \mathrm{pen}_0$ with $C = \lambda + 1$, for which the minimal penalty occurs at $C = 1$ and the (asymptotically) optimal penalty occurs for $C = 2$ —hence a factor 2 between the minimal and the optimal penalty.

The experiments of [Comte et al., 2017, Section 5] suggest that a similar way to calibrate PCO works well for selecting the bandwidth of a kernel estimator of the stationary density of the solution of a stochastic differential equation. For state-by-state inference of the emission densities of a hidden Markov model, [Lehéricy, 2018, Section 4.3.2] proposes a variant of PCO that can be well calibrated by a minimal-penalty algorithm according to numerical experiments.

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

A full theoretical validation of this calibration strategy remains an open problem. Providing theoretical guidelines for choosing between options (a) and (b) would also be interesting. Another natural question is to generalize PCO to other settings where Goldenshluger-Lepski's method applies, such as density estimation with the L^p risk or regression; to the best of our knowledge, this remains an open problem.

8.5.2. Choice of a threshold

For some thresholding estimators, with a threshold depending (non-linearly) on some parameter $\gamma \in (0, +\infty)$, an equivalent of (β') is proved —for a particular basis and assuming that $s^\star = \mathbb{1}_{[0,1]}$—, as well as an equivalent of $(\widetilde{\gamma})$ in the general case, in two settings: density estimation on $\mathbb{R}$ [Reynaud-Bouret et al., 2011] and estimation of a Poisson intensity on $\mathbb{R}$ [Reynaud-Bouret and Rivoirard, 2010].

For some Dantzig estimator (given some dictionary), with a parameter $\gamma > 0$ appearing in the Dantzig constraints, [Bertin et al., 2011] proves an equivalent of (β') —for a particular dictionary and assuming that $s^\star = \mathbb{1}_{[0,1]}$—, as well as an equivalent of $(\widetilde{\gamma})$ in the general case.

In all the above results [Bertin et al., 2011, Reynaud-Bouret and Rivoirard, 2010, Reynaud-Bouret et al., 2011], an equivalent of Algorithm 5 is proposed for a data-driven choice of γ, and numerical experiments suggest that the optimal γ is often very close to the minimal γ. Therefore, a generalization of the slope heuristics ($\gamma_{\mathrm{opt}} \approx 2\gamma_{\min}$) probably does not hold here.

8.5.3. Generalization

The above results, obtained for two different kinds of problems, suggest that phase transitions could be used much more generally for estimator selection, including the optimal calibration of learning algorithms. Section 8.3.3 proposes it for the Lasso and related procedures. We conjecture that the same idea can be used fruitfully in several other settings.

The key question is to find the good parametrization of the estimator collection. The successes of minimal-penalty algorithms rely on the parametrization by the constant C in front of a well-chosen penalty shape pen_0. Finding an appropriate parametrization for the Lasso, for instance, remains an open problem to the best of our knowledge.

8.6. Related challenges in probability theory

Addressing the statistical open problems listed above mostly relies on a few corresponding open problems in probability theory. As detailed in Section 5, for each estimator considered $\widehat{s}_m$, given some deterministic $s_m^\star$ —which can be the best estimator in the associated model, or the expectation of $\widehat{s}_m$, for instance—, the key theoretical quantities are the following:

— the excess risk $p_1(m) = \mathscr{R}(\widehat{s}_m) - \mathscr{R}(s_m^\star)$,
— the excess empirical risk $p_2(m) = \widehat{\mathscr{R}}_n(s_m^\star) - \widehat{\mathscr{R}}_n(\widehat{s}_m)$,
— the empirical process at $s_m^\star$, $\delta(m) = \mathscr{R}(s_m^\star) - \widehat{\mathscr{R}}_n(s_m^\star)$.

Since the empirical process is well-understood in general, the main challenges are about $p_1(m)$ and $p_2(m)$. One either has to show that $|p_1(m) - p_2(m)|/p_1(m)$ is small on a large-probability

event —using Proposition 2 in Section 5.2.2—, or to show non-asymptotic concentration inequalities for $p_1(m)$ and $p_2(m)$ around deterministic quantities that are known up to a multiplicative factor.

We strongly encourage further work on these questions, especially on the concentration of the excess risk $p_1(m)$ and the excess empirical risk $p_2(m)$, which are difficult theoretical problems of interest for statisticians beyond minimal penalties.
Indeed, concentrating the excess risk provides lower bounds on the risk of the estimator $\widehat{s}_m$ for a given statistical problem —and not in the minimax sense, as most statistical lower bounds—, which can be much informative for practicioners. This problem has attracted some attention in the last few years, and we review the recent work on this topic in Section 5.3.
When $\widehat{s}_m$ is the empirical risk minimizer over some model S_m, the excess empirical risk can be rewritten as the supremum of an empirical process

$$p_2(m) = \sup_{t \in S_m} \left\{ \widehat{\mathscr{R}}_n(s_m^\star) - \widehat{\mathscr{R}}_n(t) \right\},$$

which is an object of interest for empirical process theory in general. Its concentration can also be seen as a non-asymptotic version of the Wilks phenomenon, which is another reason for tackling the theoretical challenge of proving that $p_2(m)$ concentrates around some deterministic quantity. Section 5.2.3 reviews such theoretical results.

Handling large collections of estimators —see Sections 4.7 and 8.3.4— induces additional issues, since we cannot expect $p_1(m)$, $p_2(m)$, and $\delta(m)$ to concentrate tightly *uniformly* over $m \in \mathscr{M}$. This raises the challenge of understanding precisely their uniform deviations among such large collections, with high-probability upper *and lower* bounds on these deviations. In the case of model selection, by grouping models of the same dimension as explained in Section 4.7, this problem reduces to concentrating $p_1(m)$ and $p_2(m)$ for empirical risk minimizers over models that are *unions* of a large number of vector spaces of the same dimension. Note that the same probabilistic challenge arises in the problem of understanding the overpenalization phenomenon —see Section 8.4—, for both large and small collections of estimators.

Acknowledgments

The author is also member of the Select project-team of Inria Saclay - Île-de-France, and acknowledges the support of the French Agence Nationale de la Recherche (Blanc SIMI 1 2011 projet Calibration). Part of this work was done while the author was financed by CNRS and member of the Sierra team in the Département d'Informatique de l'École normale supérieure (CNRS / ENS / Inria UMR 8548), 45 rue d'Ulm, 75005 Paris, France.

I warmly thank the numerous colleagues who kindly answered my questions about their own works related with minimal penalties, during the (long) period of preparation of this article or before. Special thanks to Matthieu Lerasle and Pascal Massart for many inspiring discussions on the topic, and to Yves Rozenholc for a long discussion at a CIRM workshop about his "statistical base jumping" idea that is described in Section 7.5.

Finally, I deeply thank Patricia Reynaud-Bouret who introduced me to the topic quite early (at the beginning of 2002!) —I owe her my first numerical experiments on the slope heuristics [Arlot and Baudry, 2002]—, and my coauthors on the minimal-penalty related articles I wrote: Pascal Massart, Francis Bach, Matthieu Solnon, Alain Celisse, Zaïd Harchaoui, and Damien Garreau.

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

References

[Abramovich et al., 2006] Abramovich, F., Benjamini, Y., Donoho, D. L., and Johnstone, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.*, 34(2):584–653.

[Akaike, 1969] Akaike, H. (1969). Fitting autoregressive models for prediction. *Ann. Inst. Statist. Math.*, 21:243–247.

[Akaike, 1970] Akaike, H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math.*, 22:203–217.

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

[Akakpo, 2011] Akakpo, N. (2011). Estimating a discrete distribution via histogram selection. *ESAIM: Probability and Statistics*, 15:1–29.

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

[Andresen and Spokoiny, 2014] Andresen, A. and Spokoiny, V. (2014). Critical dimension in profile semiparametric estimation. *Electron. J. Statist.*, 8(2):3077–3125.

[Arlot, 2007] Arlot, S. (2007). *Resampling and Model Selection*. PhD thesis, University Paris-Sud 11. Available at https://tel.archives-ouvertes.fr/tel-00198803v1.

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

[Arlot, 2011] Arlot, S. (2011). Sélection de modèles et sélection d'estimateurs pour l'apprentissage statistique. Cours Peccot. Collège de France. Available at http://www.di.ens.fr/~arlot/peccot.htm.

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

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

[Arlot and Baudry, 2002] Arlot, S. and Baudry, J.-P. (2002). Sélection de modèles. In French. Master 1 report, ENS Paris. Available at https://www.math.u-psud.fr/~arlot/papers/02selection_modeles.pdf. Advisor: Yannick Baraud. Report about the paper "Gaussian model selection" by L. Birgé & P. Massart, JEMS 3(3):203–268, 2001.

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

[Arlot et al., 2012] Arlot, S., Celisse, A., and Harchaoui, Z. (2012). Kernel change-point detection. arXiv:1202.3878v2.

[Arlot and Lerasle, 2016] Arlot, S. and Lerasle, M. (2016). Choice of V for V-fold cross-validation in least-squares density estimation. *J. Mach. Learn. Res.*, 17(208):1–50.

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

[Audibert and Catoni, 2011] Audibert, J.-Y. and Catoni, O. (2011). Robust linear least squares regression. *Ann. Statist.*, 39(5):2766–2794.

[Bar-Hen et al., 2018] Bar-Hen, A., Gey, S., and Poggi, J.-M. (2018). Spatial CART Classification Trees. Available at https://hal.archives-ouvertes.fr/hal-01837065v1.

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

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

[Baraud et al., 2014] Baraud, Y., Giraud, C., and Huet, S. (2014). Estimator selection in the Gaussian setting. *Ann. Inst. Henri Poincaré Probab. Stat.*, 50(3):1092–1119.

[Bardet and Guenaizi, 2018] Bardet, J.-M. and Guenaizi, A. (2018). Semi-parametric detection of multiple changes in long-range dependent processes. arXiv:1801.02515v1.

[Bardet et al., 2012] Bardet, J.-M., Kengne, W. C., and Wintenberger, O. (2012). Multiple breaks detection in general causal time series using penalized quasi-likelihood. *Electron. J. Stat.*, 6:435–477 (electronic).

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

[Barron et al., 1999] Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413.

[Bartlett et al., 2005] Bartlett, P. L., Bousquet, O., and Mendelson, S. (2005). Local Rademacher complexities. *Ann. Statist.*, 33(4):1497–1537.

[Bartlett and Mendelson, 2006] Bartlett, P. L. and Mendelson, S. (2006). Empirical minimization. *Probability Theory and Related Fields*, 135(3):311–334.

[Baudry, 2009] Baudry, J.-P. (2009). *Model selection for clustering. Choosing the number of classes*. PhD thesis, University Paris-Sud. Available at https://tel.archives-ouvertes.fr/tel-00461550v1.

[Baudry, 2015] Baudry, J.-P. (2015). Estimation and model selection for model-based clustering with the conditional classification likelihood. *Electron. J. Statist.*, 9(1):1041–1077.

[Baudry et al., 2012] Baudry, J.-P., Maugis, C., and Michel, B. (2012). Slope heuristics: overview and implementation. *Statistics and Computing*, 22(2):455–470.

[Bellec, 2017] Bellec, P. (2017). Optimistic lower bounds for convex regularized least-squares. arXiv:1703.01332v3.

[Bellec, 2018] Bellec, P. (2018). The noise barrier and the large signal bias of the lasso and other convex estimators. arXiv:1804.01230v1.

[Bellec and Tsybakov, 2017] Bellec, P. and Tsybakov, A. (2017). Bounds on the prediction error of penalized least squares estimators with convex penalty. In Panov, V., editor, *Modern Problems of Stochastic Analysis and Statistics*, pages 315–333, Cham. Springer International Publishing.

[Bellec, 2014] Bellec, P. C. (2014). Optimal bounds for aggregation of affine estimators. Technical report, arXiv. arXiv:1410.0346v3.

[Bertin et al., 2016] Bertin, K., Lacour, C., and Rivoirard, V. (2016). Adaptive pointwise estimation of conditional density function. *Ann. Inst. Henri Poincaré Probab. Stat.*, 52(2):939–980.

[Bertin et al., 2011] Bertin, K., Le Pennec, E., and Rivoirard, V. (2011). Adaptive Dantzig density estimation. *Ann. Inst. H. Poincaré Probab. Statist.*, 47(1):43–74.

[Biau and Fischer, 2012] Biau, G. and Fischer, A. (2012). Parameter selection for principal curves. *IEEE Transactions on Information Theory*, 58(3):1924–1939.

[Birgé and Massart, 1997] Birgé, L. and Massart, P. (1997). From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York.

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

[Birgé and Massart, 2001] Birgé, L. and Massart, P. (2001). A generalized Cp criterion for Gaussian model selection. Technical report, Universités de Paris 6 et Paris 7. Prépublication 647, 39 pages. Available at http://massart.pascal.free.fr/Site/publications_files/Cp.pdf.

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

[Bitseki Penda and Roche, 2017] Bitseki Penda, S. V. and Roche, A. (2017). Local bandwidth selection for kernel density estimation in bifurcating markov chain model. arXiv:1706.07034v1.

[Blanchard and Massart, 2006] Blanchard, G. and Massart, P. (2006). Discussion: "Local Rademacher complexities and oracle inequalities in risk minimization" [Ann. Statist. **34** (2006), no. 6, 2593–2656] by V. Koltchinskii. *Ann. Statist.*, 34(6):2664–2671.

[Bontemps and Toussile, 2013] Bontemps, D. and Toussile, W. (2013). Clustering and variable selection for categorical multivariate data. *Electron. J. Stat.*, 7:2344–2371.

[Boucheron et al., 2013] Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford.

[Boucheron and Massart, 2011] Boucheron, S. and Massart, P. (2011). A high dimensional Wilks phenomenon. *Probab. Theory Related Fields*, 150(3-4):405–433.

[Bouveyron et al., 2015a] Bouveyron, C., Côme, E., and Jacques, J. (2015a). The discriminative functional mixture model for a comparative analysis of bike sharing systems. *Ann. Appl. Stat.*, 9(4):1726–1760.

[Bouveyron et al., 2015b] Bouveyron, C., Fauvel, M., and Girard, S. (2015b). Kernel discriminant analysis and clustering with parsimonious Gaussian process models. *Statistics and Computing*, 25(6):1143–1162.

[Breiman and Spector, 1992] Breiman, L. and Spector, P. (1992). Submodel Selection and Evaluation in Regression. The X-Random Case. *International Statistical Review*, 60(3):291–319.

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

[Brown and Levine, 2007] Brown, L. D. and Levine, M. (2007). Variance estimation in nonparametric regression via the difference sequence method. *Ann. Statist.*, 35(5):2219–2232.

[Buckley and Eagleson, 1989] Buckley, M. J. and Eagleson, G. K. (1989). A graphical method for estimating the residual variance in nonparametric regression. *Biometrika*, 76(2):203–210.

[Burnham and Anderson, 2002] Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference*. Springer-Verlag, New York, second edition. A practical information-theoretic approach.

[Caillerie and Michel, 2011] Caillerie, C. and Michel, B. (2011). Model selection for simplicial approximation. *Foundations of Computational Mathematics*, 11(6):707–731.

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

[Carter and Eagleson, 1992] Carter, C. K. and Eagleson, G. K. (1992). A comparison of variance estimators in nonparametric regression. *J. Roy. Statist. Soc. Ser. B*, 54(3):773–780.

[Castellan, 1999] Castellan, G. (1999). Modified Akaike's criterion for histogram density estimation. Technical Report 1999-61, University Paris-Sud. Available at https://www.math.u-psud.fr/~biblio/pub/1999/abs/ppo1999_61.html.

[Castellanos et al., 2002] Castellanos, J. L., Gómez, S., and Guerra, V. (2002). The triangle method for finding the corner of the L-curve. *Appl. Numer. Math.*, 43(4):359–373.

[Catherine and Vincent, 2017] Catherine, M. and Vincent, M. (2017). Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1119–1141.

[Cattell, 1966] Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behav. Res.*, 1(2):245–276.

[Cattell and Vogelmann, 1977] Cattell, R. B. and Vogelmann, S. (1977). A comprehensive trial of the scree and k.g. criteria for determining the number of factors. *Multivariate Behav. Res.*, 12(3):289–325.

[Chagny, 2013] Chagny, G. (2013). Penalization versus Goldenshluger-Lepski strategies in warped bases regression. *ESAIM Probab. Stat.*, 17:328–358.

[Chatterjee, 2014] Chatterjee, S. (2014). A new perspective on least squares under convex constraint. *Ann. Statist.*, 42(6):2340–2381.

[Chatterjee, 2015] Chatterjee, S. (2015). High dimensional regression and matrix estimation without tuning parameters. arXiv:1510.07294v3.

[Chen et al., 2017] Chen, X., Guntuboyina, A., and Zhang, Y. (2017). A note on the approximate admissibility of regularized estimators in the gaussian sequence model. arXiv:1703.00542v1.

[Cohen and Le Pennec, 2014] Cohen, S. X. and Le Pennec, E. (2014). Unsupervised segmentation of spectral images with a spatialized gaussian mixture model and model selection. *Oil Gas Sci. Technol. – Rev. IFP Energies nouvelles*, 69(2):245–259.

[Comte et al., 2017] Comte, F., Prieur, C., and Samson, A. (2017). Adaptive estimation for stochastic damping Hamiltonian systems under partial observation. *Stochastic Process. Appl.*, 127(11):3689–3718.

[Comte and Rozenholc, 2004] Comte, F. and Rozenholc, Y. (2004). A new algorithm for fixed design regression and denoising. *Ann. Inst. Statist. Math.*, 56(3):449–473.

[Connault, 2011] Connault, P. (2011). *Calibration d'algorithmes de type Lasso et analyse statistique de données métallurgiques en aéronautique*. PhD thesis, Université Paris-Sud.

[Craven and Wahba, 1978] Craven, P. and Wahba, G. (1978). 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.

[Derman and Le Pennec, 2017] Derman, E. and Le Pennec, E. (2017). Clustering and model selection via penalized likelihood for different-sized categorical data vectors. arXiv:1709.02294v1.

[Dette et al., 1998] Dette, H., Munk, A., and Wagner, T. (1998). Estimating the variance in nonparametric regression—what is a reasonable choice? *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 60(4):751–764.

[Devijver, 2017a] Devijver, É. (2017a). Joint rank and variable selection for parsimonious estimation in a high-dimensional finite mixture regression model. *Journal of Multivariate Analysis*, 157:1–13.

[Devijver, 2017b] Devijver, É. (2017b). Model-based regression clustering for high-dimensional data: application to functional data. *Adv. Data Analysis and Classification*, 11(2):243–279.

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

[Devijver and Gallopin, 2018] Devijver, É. and Gallopin, M. (2018). Block-diagonal covariance selection for high-dimensional Gaussian graphical models. *Journal of the American Statistical Association*, pages 306–314.

[Devijver et al., 2017] Devijver, É., Gallopin, M., and Perthame, E. (2017). Nonlinear network-based quantitative trait prediction from transcriptomic data. arXiv:1701.07899v5.

[Devijver et al., 2015] Devijver, É., Goude, Y., and Poggi, J.-M. (2015). Clustering electricity consumers using high-dimensional regression mixture models. arXiv:1507.00167v1.

[Devroye et al., 1996] Devroye, L., Györfi, L., and Lugosi, G. (1996). *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York.

[Donoho et al., 1995] Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1995). Wavelet shrinkage: asymptopia? *J. Roy. Statist. Soc. Ser. B*, 57(2):301–369. With discussion and a reply by the authors.

[Dossal et al., 2013] Dossal, C., Kachour, M., Fadili, J. M., Peyré, G., and Chesneau, C. (2013). The degrees of freedom of the lasso for general design matrix. *Statistica Sinica*, 23(2):809–828.

[Du and Schick, 2009] Du, J. and Schick, A. (2009). A covariate-matched estimator of the error variance in non-parametric regression. *J. Nonparametr. Stat.*, 21(3):263–285.

[Durot et al., 2009] Durot, C., Lebarbier, É., and Tocquet, A.-S. (2009). Estimating the joint distribution of independent categorical variables via model selection. *Bernoulli*, 15(2):475–507.

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

[Efron, 2004] Efron, B. (2004). The estimation of prediction error: covariance penalties and cross-validation. *J. Amer. Statist. Assoc.*, 99(467):619–642. With comments and a rejoinder by the author.

[Engl and Grever, 1994] Engl, H. W. and Grever, W. (1994). Using the *L*-curve for determining optimal regularization parameters. *Numer. Math.*, 69(1):25–31.

[Frontier, 1976] Frontier, S. (1976). Étude de la décroissance des valeurs propres dans une analyse en composantes principales: Comparaison avec le modèle du bâton brisé. *Journal of Experimental Marine Biology and Ecology*, 25(1):67–75.

[Garivier and Lerasle, 2011] Garivier, A. and Lerasle, M. (2011). Oracle approach and slope heuristic in context tree estimation. arXiv:1111.2191v1.

[Garreau and Arlot, 2018] Garreau, D. and Arlot, S. (2018). Consistent change-point detection with kernels. *Electron. J. Statist.*, 12(2):4440–4486.

[Gassiat and Van Handel, 2013] Gassiat, E. and Van Handel, R. (2013). Consistent order estimation and minimal penalties. *IEEE Trans. Inform. Theory*, 59(2):1115–1128.

[Gavish and Donoho, 2014] Gavish, M. and Donoho, D. L. (2014). The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Trans. Inform. Theory*, 60(8):5040–5053.

[Gendre, 2008] Gendre, X. (2008). Simultaneous estimation of the mean and the variance in heteroscedastic Gaussian regression. *Electron. J. Stat.*, 2:1345–1372.

[Gey and Lebarbier, 2008] Gey, S. and Lebarbier, É. (2008). Using CART to Detect Multiple Change Points in the Mean for large samples. Technical Report 12, Statistics for Systems Biology. Available at https://hal.archives-ouvertes.fr/hal-00327146v1.

[Giacobino et al., 2017] Giacobino, C., Sardy, S., Diaz-Rodriguez, J., and Hengartner, N. (2017). Quantile universal threshold. *Electron. J. Statist.*, 11(2):4701–4722.

[Giraud, 2008] Giraud, C. (2008). Estimation of gaussian graphs by model selection. *Electron. J. Stat.*, 2:542–563 (electronic).

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

[Giraud, 2014] Giraud, C. (2014). *Introduction to High-Dimensional Statistics*, volume 139 of *Monographs on Statistics and Applied Probability*. Chapman and Hall/CRC, Boca Raton, FL.

[Giraud et al., 2012] Giraud, C., Huet, S., and Verzelen, N. (2012). High-dimensional regression with unknown variance. *Statist. Sci.*, 27(4):500–518.

[Godichon-Baggioni et al., 2019] Godichon-Baggioni, A., Maugis-Rabusseau, C., and Rau, A. (2019). Clustering transformed compositional data using K-means, with applications in gene expression and bicycle sharing system data. *Journal of Applied Statistics*, 46(1):47–65.

[Goldenshluger and Lepski, 2011] Goldenshluger, A. and Lepski, O. (2011). Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *Ann. Statist.*, 39(3):1608–1632.

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

[Grodzevich and Wolkowicz, 2009] Grodzevich, O. and Wolkowicz, H. (2009). Regularization using a parameterized trust region subproblem. *Math. Program.*, 116(1-2):193–220.

[Hall et al., 1990] Hall, P., Kay, J. W., and Titterington, D. M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika*, 77(3):521–528.

[Hall and Marron, 1990] Hall, P. and Marron, J. S. (1990). On variance estimation in nonparametric regression. *Biometrika*, 77(2):415–419.

[Hanke, 1996] Hanke, M. (1996). Limitations of the *L*-curve method in ill-posed problems. *BIT*, 36(2):287–301.

[Hansen, 1992] Hansen, P. C. (1992). Analysis of discrete ill-posed problems by means of the *L*-curve. *SIAM Rev.*, 34(4):561–580.

[Hansen et al., 2007] Hansen, P. C., Jensen, T. K., and Rodriguez, G. (2007). An adaptive pruning algorithm for the discrete L-curve criterion. *J. Comput. Appl. Math.*, 198(2):483–492.

[Hansen and O'Leary, 1993] Hansen, P. C. and O'Leary, D. P. (1993). The use of the *L*-curve in the regularization of discrete ill-posed problems. *SIAM J. Sci. Comput.*, 14(6):1487–1503.

[Heng et al., 2010] Heng, Y., Lu, S., Mhamdi, A., and Pereverzev, S. V. (2010). Model functions in the modified *L*-curve method—case study: the heat flux reconstruction in pool boiling. *Inverse Problems*, 26(5):055006, 13.

[Horn, 1965] Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185.

[Horn and Engstrom, 1979] Horn, J. L. and Engstrom, R. (1979). Cattell's scree test in relation to Bartlett's chi-square test and other observations on the number of factors problem. *Multivariate Behavioral Research*, 14(3):283–300.

[Jackson, 1993] Jackson, D. A. (1993). Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology*, 74(8).

[Koltchinskii, 2001] Koltchinskii, V. (2001). Rademacher penalties and structural risk minimization. *IEEE Trans. Inform. Theory*, 47(5):1902–1914.

[Koltchinskii, 2006] Koltchinskii, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6):2593–2656.

[Lacour and Massart, 2016] Lacour, C. and Massart, P. (2016). Minimal penalty for Goldenshluger-Lepski method. *Stochastic Processes and their Applications*, 126(12):3774–3789. In Memoriam: Evarist Giné.

[Lacour et al., 2017] Lacour, C., Massart, P., and Rivoirard, V. (2017). Estimator selection: a new method with applications to kernel density estimation. *Sankhya A*, 79(2):298–335.

[Lavielle, 2005] Lavielle, M. (2005). Using penalized contrasts for the change-point problem. *Signal Proces.*, 85(8):1501–1510.

[Lawson and Hanson, 1974] Lawson, C. L. and Hanson, R. J. (1974). *Solving least squares problems*. Prentice-Hall Inc., Englewood Cliffs, N.J. Prentice-Hall Series in Automatic Computation.

[Lebarbier, 2002] Lebarbier, É. (2002). *Quelques approches pour la détection de ruptures à horizon fini*. PhD thesis, Université Paris-Sud.

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

[Lehéricy, 2018] Lehéricy, L. (2018). State-by-state minimax adaptive estimation for nonparametric hidden markov models. *Journal of Machine Learning Research*, 19(39):1–46.

[Lerasle, 2009] Lerasle, M. (2009). *Rééchantillonnage et sélection de modèles optimale pour l'estimation de la densité de variables indépendantes ou mélangeantes*. PhD thesis, INSA de Toulouse. Available at http://lerasle.perso.math.cnrs.fr/docs/these.pdf.

[Lerasle, 2010] Lerasle, M. (2010). Optimal model selection in density estimation. arXiv:0910.1654v2.

[Lerasle, 2011] Lerasle, M. (2011). Optimal model selection for stationary data under various mixing conditions. *Ann. Statist.*, 39(4):1852–1877.

[Lerasle, 2012] Lerasle, M. (2012). Optimal model selection in density estimation. *Ann. Inst. Henri Poincaré Probab. Stat.*, 48(3):884–908.

[Lerasle et al., 2016] Lerasle, M., Magalhães, N., and Reynaud-Bouret, P. (2016). Optimal kernel selection for density estimation. In *High Dimensional Probability VII: The Cargese Volume*, volume 71 of *Progress in Probability*, pages 425–460. Springer. Preliminary version available at arXiv:1511.02112.

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

[Lerasle and Takahashi, 2011] Lerasle, M. and Takahashi, D. Y. (2011). An oracle approach for interaction neighborhood estimation in random fields. *Electron. J. Stat.*, 5:534–571 (electronic).

[Lerasle and Takahashi, 2016] Lerasle, M. and Takahashi, D. Y. (2016). Sharp oracle inequalities and slope heuristic for specification probabilities estimation in discrete random fields. *Bernoulli*, 22(1):325–344.

[Letué, 2000] Letué, F. (2000). *Modèle de Cox: estimation par sélection de modèle et modèle de chocs bivarié*. PhD thesis, Université Paris-Sud. Available at http://www-ljk.imag.fr/membres/Frederique.Letue/These3.pdf.

[Li, 1985] Li, K.-C. (1985). From Stein's unbiased risk estimates to the method of generalized cross validation. *Ann. Statist.*, 13(4):1352–1377.

[Li, 1986] Li, K.-C. (1986). Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. *Ann. Statist.*, 14(3):1101–1112.

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

[Liitiäinen et al., 2010] Liitiäinen, E., Corona, F., and Lendasse, A. (2010). Residual variance estimation using a nearest neighbor statistic. *J. Multivariate Anal.*, 101(4):811–823.

[Liitiäinen et al., 2009] Liitiäinen, E., Verleysen, M., Corona, F., and Lendasse, A. (2009). Residual variance estimation in machine learning. *Neurocomputing*, 72(16):3692–3703. Financial Engineering Computational and Ambient Intelligence (IWANN 2007).

[Loubes and Massart, 2004] Loubes, J.-M. and Massart, P. (2004). Discussion: "Least angle regression" [Ann. Statist. **32** (2004), no. 2, 407–451] by B. Efron, T. Hastie, I. Johnstone and R. Tibshirani. *Ann. Statist.*, 32(2):460–465.

[Lozano, 2000] Lozano, F. (2000). Model selection using Rademacher penalization. In *Proceedings of the 2nd ICSC Symp. on Neural Computation (NC2000). Berlin, Germany*. ICSC Academic Press.

[Magalhães, 2015] Magalhães, N. (2015). *Cross-Validation and Penalization for Density Estimation*. PhD thesis, Université Paris Sud - Paris XI. Available at https://tel.archives-ouvertes.fr/tel-01164581v1.

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

[Mammen and Tsybakov, 1999] Mammen, E. and Tsybakov, A. B. (1999). Smooth discrimination analysis. *Ann. Statist.*, 27(6):1808–1829.

[Massart, 2005] Massart, P. (2005). A non-asymptotic theory for model selection. In *European Congress of Mathematics*, pages 309–323. Eur. Math. Soc., Zürich.

[Massart, 2007] Massart, P. (2007). *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[Massart, 2008] Massart, P. (2008). Sélection de modèles: de la théorie à la pratique. *Journal de la SFdS*, 149(4):5–28.

[Massart and Nédélec, 2006] Massart, P. and Nédélec, É. (2006). Risk bounds for statistical learning. *Ann. Statist.*, 34(5):2326–2366.

[Maugis, 2008] Maugis, C. (2008). *Variable selection for model-based clustering. Application for transcriptome data analysis.* PhD thesis, Université Paris-Sud. Available at https://tel.archives-ouvertes.fr/tel-00344120v1.

[Maugis and Michel, 2011a] Maugis, C. and Michel, B. (2011a). Data-driven penalty calibration: A case study for gaussian model selection. *ESAIM Probab. Stat.*, 15:320–339.

[Maugis and Michel, 2011b] Maugis, C. and Michel, B. (2011b). A non asymptotic penalized criterion for gaussian mixture model selection. *ESAIM Probab. Stat.*, 15:41–68.

[Mendelson, 2018] Mendelson, S. (2018). Learning without concentration for general loss functions. *Probab. Theory Related Fields*, 171(1-2):459–502.

[Meynet and Maugis-Rabusseau, 2012] Meynet, C. and Maugis-Rabusseau, C. (2012). A sparse variable selection procedure in model-based clustering. Available at https://hal.inria.fr/hal-00734316v1.

[Michel, 2008] Michel, B. (2008). *Modélisation de la production d'hydrocarbures dans un bassin pétrolier*. PhD thesis, Université Paris-Sud. Available at http://tel.archives-ouvertes.fr/tel-00345753v1.

[Miller, 1970] Miller, K. (1970). Least squares methods for ill-posed problems with a prescribed bound. *SIAM J. Math. Anal.*, 1:52–74.

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

[Müller et al., 2003] Müller, U. U., Schick, A., and Wefelmeyer, W. (2003). Estimating the error variance in nonparametric regression by a covariate-matched U-statistic. *Statistics*, 37(3):179–188.

[Muro and Geer, 2018] Muro, A. and Geer, S. (2018). Concentration behavior of the penalized least squares estimator. *Statistica Neerlandica*, 72(2):109–125.

[Navarro and Saumard, 2017] Navarro, F. and Saumard, A. (2017). Slope heuristics and V-fold model selection in heteroscedastic regression using strongly localized bases. *ESAIM Probab. Stat.*, 21:412–451.

[Oueslati and Lopez, 2013] Oueslati, A. and Lopez, O. (2013). A proportional hazards regression model with change-points in the baseline function. *Lifetime Data Analysis*, 19(1):59–78.

[Rau et al., 2015] Rau, A., Maugis-Rabusseau, C., Martin-Magniette, M.-L., and Celeux, G. (2015). Co-expression analysis of high-throughput transcriptome sequencing data with poisson mixture models. *Bioinformatics*, 31(9):1420–1427.

[Regińska, 1996] Regińska, T. (1996). A regularization parameter in discrete ill-posed problems. *SIAM J. Sci. Comput.*, 17(3):740–749.

[Reichel and Rodriguez, 2013] Reichel, L. and Rodriguez, G. (2013). Old and new parameter choice rules for discrete ill-posed problems. *Numerical Algorithms*, 63(1):65–87.

[Reid et al., 2016] Reid, S., Tibshirani, R., and Friedman, J. (2016). A study of error variance estimation in Lasso regression. *Statist. Sinica*, 26(1):35–67.

[Reynaud-Bouret and Rivoirard, 2010] Reynaud-Bouret, P. and Rivoirard, V. (2010). Near optimal thresholding estimation of a poisson intensity on the real line. *Electron. J. Stat.*, 4:172–238 (electronic).

[Reynaud-Bouret et al., 2011] Reynaud-Bouret, P., Rivoirard, V., and Tuleau-Malot, C. (2011). Adaptive density estimation: a curse of support? *J. Statist. Plann. Inference*, 141(1):115–139.

[Reynaud-Bouret and Schbath, 2010] Reynaud-Bouret, P. and Schbath, S. (2010). Adaptive estimation for hawkes processes; application to genome analysis. *Ann. Statist.*, 38(5):2781–2822.

[Rice, 1984] Rice, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.*, 12(4):1215–1230.

[Roche, 2014] Roche, A. (2014). *Statistical modeling for functional data: non-asymptotic approaches and adaptive methods*. PhD thesis, Université Montpellier II - Sciences et Techniques du Languedoc. Available at https://tel.archives-ouvertes.fr/tel-01023919v1.

[Rozenholc, 2012] Rozenholc, Y. (2012). Statistical base jumping: A simple and fully data-driven answer to penalized model selection. Séminaire de Statistique du MAP5, February 3rd.

[Saumard, 2010a] Saumard, A. (2010a). Convergence in sup-norm of least-squares estimators in regression with random design and nonparametric heteroscedastic noise. Available at http://hal.archives-ouvertes.fr/hal-00528539v2.

[Saumard, 2010b] Saumard, A. (2010b). *Estimation par Minimum de Contraste Régulier et Heuristique de Pente en Sélection de Modèles*. PhD thesis, Université de Rennes 1. Available at http://tel.archives-ouvertes.fr/tel-00569372v1.

[Saumard, 2010c] Saumard, A. (2010c). Nonasymptotic quasi-optimality of AIC and the slope heuristics in maximum likelihood estimation of density using histogram models. Available at https://hal.archives-ouvertes.fr/hal-00512310v1.

[Saumard, 2012] Saumard, A. (2012). Optimal upper and lower bounds for the true and empirical excess risks in heteroscedastic least-squares regression. *Electron. J. Stat.*, 6:579–655.

[Saumard, 2013] Saumard, A. (2013). Optimal model selection in heteroscedastic regression using piecewise polynomial functions. *Electron. J. Stat.*, 7:1184–1223.

[Saumard, 2017] Saumard, A. (2017). A concentration inequality for the excess risk in least-squares regression with random design and heteroscedastic noise. arXiv:1702.05063v2.

[Saumard and Navarro, 2018a] Saumard, A. and Navarro, F. (2018a). Finite sample improvement of akaike's information criterion. arXiv:1803.02078v4.

[Saumard and Navarro, 2018b] Saumard, A. and Navarro, F. (2018b). Model selection as a multiple testing procedure: Improving Akaike's information criterion. arXiv:1803.02078v2.

[Schwarz, 1978] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464.

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

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

[Solnon, 2013] Solnon, M. (2013). *Apprentissage statistique multi-tâches*. PhD thesis, Université Pierre et Marie Curie - Paris VI. Available at https://hal.inria.fr/tel-00911498v1.

[Solnon et al., 2012] Solnon, M., Arlot, S., and Bach, F. (2012). Multi-task regression using minimal penalties. *J. Mach. Learn. Res.*, 13:2773–2812 (electronic).

[Sorba, 2017] Sorba, O. (2017). *Minimal penalties for model selection*. PhD thesis, Université Paris-Saclay. Available at https://tel.archives-ouvertes.fr/tel-01515957v1.

[Spokoiny, 2002] Spokoiny, V. (2002). Variance estimation for high-dimensional regression models. *J. Multivariate Anal.*, 82(1):111–133.

[Spokoiny, 2012] Spokoiny, V. (2012). Parametric estimation. finite sample theory. *Ann. Statist.*, 40(6):2877–2909.

[Spokoiny, 2017] Spokoiny, V. (2017). Penalized maximum likelihood estimation and effective dimension. *Ann. Inst. Henri Poincaré Probab. Stat.*, 53(1):389–429.

[Stein, 1981] Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 9(6):1135–1151.

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

[Sugar and James, 2003] Sugar, C. A. and James, G. M. (2003). Finding the number of clusters in a dataset: an information-theoretic approach. *J. Amer. Statist. Assoc.*, 98(463):750–763.

[Tibshirani et al., 2001] Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 63(2):411–423.

[Tibshirani and Taylor, 2012] Tibshirani, R. J. and Taylor, J. (2012). Degrees of freedom in lasso problems. *Ann. Statist.*, 40(2):1198–1232.

[Tong et al., 2013] Tong, T., Ma, Y., and Wang, Y. (2013). Optimal variance estimation without estimating the mean function. *Bernoulli*, 19(5A):1839–1854.

[Ullah and Zinde-Walsh, 1992] Ullah, A. and Zinde-Walsh, V. (1992). On the estimation of residual variance in nonparametric regression. *J. Nonparametr. Statist.*, 1(3):263–265.

[Vaiter et al., 2012] Vaiter, S., Deledalle, C., Peyré, G., Fadili, J. M., and Dossal, C. (2012). The Degrees of Freedom of the Group Lasso. In *International Conference on Machine Learning Workshop (ICML)*, Edinburgh, United Kingdom. Available at https://hal.archives-ouvertes.fr/hal-00695292.

[van de Geer and Wainwright, 2017] van de Geer, S. and Wainwright, M. J. (2017). On concentration for (regularized) empirical risk minimization. *Sankhya A*, 79(2):159–200.

[van Erven et al., 2012] van Erven, T., Grünwald, P. D., and de Rooij, S. (2012). Catching up faster by switching sooner: a predictive approach to adaptive estimation with an application to the AIC-BIC dilemma. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):361–417.

[van Handel, 2011] van Handel, R. (2011). On the minimal penalty for markov order estimation. *Probability Theory and Related Fields*, 150(3):709–738.

[Varet et al., 2018] Varet, S., Lacour, C., Massart, P., and Rivoirard, V. (2018). Performances numériques du choix de fenêtre par PCO lors de l'estimation de densités multivariées par méthode à noyau. In *50èmes Journées de Statistique, SFdS*, EDF Lab Paris-Saclay, France.

[Vert, 2006] Vert, R. (2006). *Theoretical Insights on Density Level Set Estimation, Application to Anomaly Detection*. PhD thesis, Université Paris Sud. Available at http://sites.google.com/site/regisvert/Home/publications/files-1/thesis.pdf.

[Verzelen, 2010] Verzelen, N. (2010). Data-driven neighborhood selection of a gaussian field. *Comput. Statist. Data Anal.*, 54(5):1355–1371.

[Villers, 2007] Villers, F. (2007). *Tests et Sélection de Modèles pour l'Analyse de Données Protéomiques et Transcriptomiques*. PhD thesis, University Paris XI. Available at http://www.proba.jussieu.fr/~villers/manuscript.pdf.

[Vogel, 1996] Vogel, C. R. (1996). Non-convergence of the *L*-curve regularization parameter selection method. *Inverse Problems*, 12(4):535–547.

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

[Wahba, 1977] Wahba, G. (1977). A survey of some smoothing problems and the method of generalized cross-validation for solving them. In *Applications of statistics (Proc. Sympos., Wright State Univ., Dayton, Ohio, 1976)*, pages 507–523. North-Holland, Amsterdam.

[Wilks, 1938] Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statistics*, 9:60–62.

[Yang, 2005] Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950.

[Zwald, 2005] Zwald, L. (2005). *Statistical performances of learning algorithm : Kernel Projection Machine and Kernel Principal Component Analysis.* PhD thesis, Université Paris Sud. Available at http://tel.archives-ouvertes.fr/tel-00012011v1.

Appendix A: Some proofs

A.1. Proof of Proposition 1

Proof of Eq. (61) and (62) By definition of $\widehat{m}^{(0)}_{\min}(C)$, for every $m \in \mathscr{M}$,

$$\mathscr{R}\Big(s^\star_{\widehat{m}^{(0)}_{\min}(C)}\Big) - \delta\Big(\widehat{m}^{(0)}_{\min}(C)\Big) + (C-1)p_2\Big(\widehat{m}^{(0)}_{\min}(C)\Big) \leqslant \mathscr{R}(s^\star_m) - \delta(m) + (C-1)p_2(m)$$

hence

$$\begin{aligned}(1-C)p_2\Big(\widehat{m}^{(0)}_{\min}(C)\Big) \geqslant \mathscr{R}\Big(s^\star_{\widehat{m}^{(0)}_{\min}(C)}\Big) - \mathscr{R}(s^\star) - \big(\mathscr{R}(s^\star_m) - \mathscr{R}(s^\star)\big)\\ + \delta(m) - \delta\Big(\widehat{m}^{(0)}_{\min}(C)\Big) + (1-C)p_2(m)\end{aligned}$$

which implies, using Eq. (60), that

$$(1-C)p_2\Big(\widehat{m}^{(0)}_{\min}(C)\Big) \geqslant -2\big(\mathscr{R}(s^\star_m) - \mathscr{R}(s^\star)\big) + (1-C)p_2(m) - \varepsilon'_\delta$$

hence Eq. (61) by dividing by $(1-C)$. Eq. (62) is a straightforward consequence of Eq. (61) since $p_2(m_1) > 0$.

General proof of Eq. (60) If $\xi_1, \dots, \xi_n \in \mathscr{X}$ are i.i.d. and some contrast function $\gamma : \Xi \times \mathbb{S} \to \mathbb{R}$ and constants $A, L > 0$ exist such that (63)–(65) hold true. Then, for every $x \geqslant 0$, with probability at least $1 - 2\operatorname{card}(\mathscr{M})\mathrm{e}^{-x}$, for every $\theta > 0$, Eq. (60) holds with

$$\varepsilon_\delta = \theta \quad \text{and} \quad \varepsilon'_\delta = 2\left(\frac{L}{2\theta} + \frac{2A}{3}\right)\frac{x}{n}.$$

Indeed, for every fixed $m \in \mathscr{M}$,

$$\delta(m) - \big[\mathscr{R}(s^\star) - \widehat{\mathscr{R}}_n(s^\star)\big] = \frac{1}{n}\sum_{i=1}^n \big(X_{i,m} - \mathbb{E}[X_{i,m}]\big) \quad \text{where} \quad X_{i,m} = \gamma(\xi_i, s^\star) - \gamma(\xi_i, s^\star_m)$$

are i.i.d. random variables satisfying $|X_{i,m}| \leqslant 2A$ a.s. Therefore, by Bernstein's inequality [Boucheron et al., 2013, Theorem 2.10], for every $x \geqslant 0$, with probability at least $1 - 2\mathrm{e}^{-x}$,

$$\Big|\delta(m) - \big[\mathscr{R}(s^\star) - \widehat{\mathscr{R}}_n(s^\star)\big]\Big| \leqslant \sqrt{\frac{2x\operatorname{var}(X_{i,m})}{n}} + \frac{2Ax}{3n} \leqslant \theta\big(\mathscr{R}(s^\star_m) - \mathscr{R}(s^\star)\big) + \left(\frac{L}{2\theta} + \frac{2A}{3}\right)\frac{x}{n}$$

and the result follows by the union bound. □

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

Extension Note that assuming only $(1-\varepsilon_0)p_2(m) \leqslant \mathrm{pen}_0(m) \leqslant (1+\varepsilon_0)p_2(m)$ for some ε_0 with $C(1-\varepsilon_0) < 1$ —which implies $p_2(m) \geqslant 0$—, instead of Eq. (61) we get that

$$p_2\Big(\widehat{m}^{(0)}_{\min}(C)\Big) \geqslant \sup_{m\in\mathscr{M}} \left\{ \frac{1-(1+\varepsilon_0)C}{1-(1-\varepsilon_0)C} p_2(m) - \frac{2\big[\mathscr{R}(s^\star_m) - \mathscr{R}(s^\star)\big]}{1-(1-\varepsilon_0)C} \right\} - \frac{\varepsilon'_\delta}{1-(1-\varepsilon_0)C}\,.$$

If in addition some $m_1 \in \mathscr{M}$ exists such that $\mathscr{R}(s^\star_{m_1}) = \mathscr{R}(s^\star)$ and $p_2(m_1) > 0$, we get that for any $\alpha \in (0,1)$,

$$p_2\Big(\widehat{m}^{(0)}_{\min}(C)\Big) \geqslant (1-\alpha)p_2(m_1)$$

if

$$C \leqslant 1-\eta_\alpha \qquad \text{where} \qquad \eta_\alpha := 1 - \left(1 - \frac{\varepsilon'_\delta}{\alpha p_2(m_1)}\right)\left[1+\varepsilon_0\left(\frac{2}{\alpha}-1\right)\right]^{-1}.$$

A.2. Proof of Proposition 2

By definition of $\widehat{m}^{(0)}_{\min}(C)$, for every $m \in \mathscr{M}$ and $C > 1$,

$$\begin{aligned}
&\mathscr{R}\Big(\widehat{s}_{\widehat{m}^{(0)}_{\min}(C)}\Big) - p_1\Big(\widehat{m}^{(0)}_{\min}(C)\Big) - \delta\Big(\widehat{m}^{(0)}_{\min}(C)\Big) + (C-1)p_2\Big(\widehat{m}^{(0)}_{\min}(C)\Big) \\
&\leqslant \mathscr{R}(\widehat{s}_m) - p_1(m) - \delta(m) + (C-1)p_2(m)
\end{aligned}$$

hence, using Eq. (66),

$$\begin{aligned}
&\mathscr{R}\Big(\widehat{s}_{\widehat{m}^{(0)}_{\min}(C)}\Big) - \mathscr{R}(s^\star) + \big[(C-1)(1-\varepsilon_p) - 1\big] p_1\Big(\widehat{m}^{(0)}_{\min}(C)\Big) \\
&\leqslant \mathscr{R}(\widehat{s}_m) - \mathscr{R}(s^\star) + \Big[(C-1)(1+\varepsilon_p) - 1\Big] p_1(m) + \delta\Big(\widehat{m}^{(0)}_{\min}(C)\Big) - \delta(m)\,.
\end{aligned}$$

By Eq. (60), we get that for every $m \in \mathscr{M}$ and $C > 1$,

$$\begin{aligned}
&\mathscr{R}\Big(\widehat{s}_{\widehat{m}^{(0)}_{\min}(C)}\Big) - \mathscr{R}(s^\star) + \big[(C-1)(1-\varepsilon_p) - 1\big] p_1\Big(\widehat{m}^{(0)}_{\min}(C)\Big) - \varepsilon_\delta\left[\mathscr{R}\Big(s^\star_{\widehat{m}^{(0)}_{\min}(C)}\Big) - \mathscr{R}(s^\star)\right] \\
&\leqslant \mathscr{R}(\widehat{s}_m) - \mathscr{R}(s^\star) + \big[(C-1)(1+\varepsilon_p) - 1\big] p_1(m) + \varepsilon_\delta\big[\mathscr{R}(s^\star_m) - \mathscr{R}(s^\star)\big] + \varepsilon'_\delta
\end{aligned}$$

that is,

$$\begin{aligned}
&\Big(1 - \max\big\{1-(C-1)(1-\varepsilon_p)\,,\,\varepsilon_\delta\big\}\Big)\left[\mathscr{R}\Big(\widehat{s}_{\widehat{m}^{(0)}_{\min}(C)}\Big) - \mathscr{R}(s^\star)\right] \\
&\leqslant \Big(1 + \max\big\{(C-1)(1+\varepsilon_p) - 1\,,\,\varepsilon_\delta\big\}\Big) \inf_{m\in\mathscr{M}}\big\{\mathscr{R}(\widehat{s}_m) - \mathscr{R}(s^\star)\big\} + \varepsilon'_\delta
\end{aligned}$$

which proves Eq. (67). Using again Eq. (66), we get

$$\begin{aligned}
p_2\Big(\widehat{m}^{(0)}_{\min}(C)\Big) &\leqslant (1+\varepsilon_p) p_1\Big(\widehat{m}^{(0)}_{\min}(C)\Big) \leqslant (1+\varepsilon_p)\left[\mathscr{R}\Big(\widehat{s}_{\widehat{m}^{(0)}_{\min}(C)}\Big) - \mathscr{R}(s^\star)\right] \\
&\leqslant K(C)(1+\varepsilon_p) \inf_{m\in\mathscr{M}}\big\{\mathscr{R}(\widehat{s}_m) - \mathscr{R}(s^\star)\big\} + K'(C)(1+\varepsilon_p)\,. \quad \square
\end{aligned}$$

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

Extension Note that if $\mathrm{pen}_0(m) = p_2(m)$ is replaced by $(1-\varepsilon_0)p_2(m) \leqslant \mathrm{pen}_0(m) \leqslant (1+\varepsilon_0)p_2(m)$ for some $\varepsilon_0 \geqslant 0$ with $C(1-\varepsilon_0) > 1$, and if Eq. (66) is replaced by

$$\forall m \in \mathscr{M}\,, \quad -\varepsilon'_p + \varepsilon_p^- p_1(m) \leqslant p_2(m) \leqslant \varepsilon_p^+ p_1(m) + \varepsilon'_p \tag{83}$$

for some $\varepsilon_p^-, \varepsilon_p^+ > 0$ and $\varepsilon'_p \geqslant 0$, the same proof shows that Eq. (67) holds true with

$$K(C) := \frac{\max\left\{[C(1+\varepsilon_0)-1]\varepsilon_p^+\,,\, 1+\varepsilon_\delta\right\}}{\min\left\{[C(1-\varepsilon_0)-1]\varepsilon_p^-\,,\, 1-\varepsilon_\delta\right\}}$$

$$\text{and} \qquad K'(C) := \frac{\varepsilon'_\delta + 2(C-1)\varepsilon'_p}{\min\left\{[C(1-\varepsilon_0)-1]\varepsilon_p^-\,,\, 1-\varepsilon_\delta\right\}}\,,$$

and Eq. (68) replaced by

$$p_2\left(\widehat{m}^{(0)}_{\min}(C)\right) \leqslant K(C)\varepsilon_p^+ \inf_{m\in\mathscr{M}}\left\{\mathscr{R}(\widehat{s}_m) - \mathscr{R}(s^\star)\right\} + K'(C)\varepsilon_p^+ + \varepsilon'_p\,. \tag{84}$$

A.3. Proof of Proposition 3

We first state two general lemmas for $\widehat{C}_{\text{thr.}}$ and $\widehat{C}_{\text{window}}$, as defined by Eq. (20) and (19) in Section 2.5, respectively. These lemmas do not assume a specific definition for $\widehat{m}(C)$ and D_m, so they apply to Algorithms 1, 3, 4, and 5 (possibly up to a rescaling of $\mathscr{C}_m$ for Lemma 2 and Algorithm 5).

Lemma 1. *Let $T_n \in \mathbb{R}$, $\widehat{m} : [0,+\infty) \to \mathscr{M}$ some function, and $\widehat{C}_{\text{thr.}}(T_n) := \inf\left\{C \geqslant 0 \,/\, D_{\widehat{m}(C)} \leqslant T_n\right\}$.*
If $\Gamma_1 \geqslant 0$ and $D_{\widehat{m}(C)} > T_n$ for all $C < \Gamma_1$, then, $\widehat{C}_{\text{thr.}}(T_n) \geqslant \Gamma_1$.
If $\Gamma_2 \geqslant 0$ and $D_{\widehat{m}(\Gamma_2)} \leqslant T_n$, then, $\widehat{C}_{\text{thr.}}(T_n) \leqslant \Gamma_2$.

Proof. The proof is straightforward from the definition of $\widehat{C}_{\text{thr.}}$. □

Lemma 2. *Let $\eta > 0$, $\widehat{m} : [0,+\infty) \to \mathscr{M}$ some function, and*

$$\widehat{C}_{\text{window}}(\eta) \in \operatorname*{argmax}_{C\geqslant 0}\left\{D_{\widehat{m}(C/(1+\eta))} - D_{\widehat{m}(C(1+\eta))}\right\}.$$

Assume that $0 \leqslant D_m \leqslant n$ for every $m \in \mathscr{M}$. If $a_n, b_n \in \mathbb{R}$ and $\Gamma_2 > \Gamma_1 > 0$ exist such that $a_n - b_n > \max\{n-a_n, b_n\}$,

$$\forall C \leqslant \Gamma_1, \quad D_{\widehat{m}(C)} \geqslant a_n\,, \qquad \forall C \geqslant \Gamma_2, \quad D_{\widehat{m}(C)} \leqslant b_n\,, \qquad \text{and} \qquad (1+\eta)^2 \geqslant \frac{\Gamma_2}{\Gamma_1}\,,$$

then,

$$\frac{\Gamma_1}{1+\eta} < \widehat{C}_{\text{window}}(\eta) < \Gamma_2(1+\eta)\,.$$

Proof. First, for $C = \sqrt{\Gamma_1\Gamma_2}$, we have $C/(1+\eta) \leqslant \Gamma_1$ and $C(1+\eta) \geqslant \Gamma_2$, hence $D_{\widehat{m}(C/(1+\eta))} - D_{\widehat{m}(C(1+\eta))} \geqslant a_n - b_n$. Second, for any $C \leqslant \Gamma_1/(1+\eta)$, $D_{\widehat{m}(C/(1+\eta))} - D_{\widehat{m}(C(1+\eta))} \leqslant n - a_n$. Third, for any $C \geqslant \Gamma_2(1+\eta)$, $D_{\widehat{m}(C/(1+\eta))} - D_{\widehat{m}(C(1+\eta))} \leqslant b_n$. The result follows since $a_n - b_n > \max\{n-a_n, b_n\}$. □

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

Let us now prove Proposition 3.

By Eq. (30) and Eq. (35) in the proof of Theorem 1, for all $x \geqslant 0$, on the event Ω_x which has a probability larger than $1-4\,\mathrm{card}(\mathscr{M})\mathrm{e}^{-x}$, we have

$$\forall C \leqslant C_1(x;a_n), \quad D_{\widehat{m}(C)} \geqslant a_n \qquad \text{and} \qquad \forall C \geqslant C_2(x;b_n;c_n), \quad D_{\widehat{m}(C)} \leqslant b_n\,, \tag{85}$$

whatever $0 \leqslant c_n < b_n \leqslant n$ and $0 \leqslant a_n < n$, provided that $\mathscr{M}$ contains at least one model of dimension at most c_n.

Proof of Eq. (74) First, note that

$$C_1\left(x;\frac{2n}{3}\right) = \sigma^2\left[1-\left(12\sqrt{\frac{x}{n}}+18\frac{x}{n}\right)\right] > 0$$

since $\sqrt{x/n} \leqslant (\sqrt{6}-2)/6$. Therefore, by continuity of C_1 and using that $c_n < n/3$, some $\varepsilon_1 \in (0, \min\{1/2, 1-3c_n/n\})$ exists such that $C_1(x;\frac{2n}{3}(1+\varepsilon)) > 0$ for any $\varepsilon \in [0,\varepsilon_1]$. Let $\varepsilon \in (0,\varepsilon_1]$, $a_n = \frac{2n}{3}(1+\varepsilon) \in (2n/3, n)$, and $b_n = \frac{n}{3}(1-\varepsilon) \in (c_n, n/3)$. By Lemma 2 with a_n, b_n as above, $\Gamma_1 = C_1(x;a_n)$, and $\Gamma_2 = C_2(x;b_n;c_n)$, on Ω_x, since Eq. (85) holds true, we get that

$$\forall \eta \geqslant \sqrt{\frac{C_2\left(x;\frac{n}{3}(1-\varepsilon);c_n\right)}{C_1\left(x;\frac{2n}{3}(1+\varepsilon)\right)}}-1\,, \qquad \frac{C_1\left(x;\frac{2n}{3}(1+\varepsilon)\right)}{1+\eta} < \widehat{C}_{\text{window}}(\eta) < C_2\left(x;\frac{n}{3}(1-\varepsilon);c_n\right)(1+\eta)\,.$$

Now, since $z \mapsto C_1(x;z)$ is continuous and different from zero at $z = 2n/3$, and since $z \mapsto C_2(x;z;c_n)$ is continuous at $z = n/3$ (using that $c_n < n/3$), for any

$$\eta > \sqrt{\frac{C_2\left(x;\frac{n}{3};c_n\right)}{C_1\left(x;\frac{2n}{3}\right)}}-1\,,$$

some $\varepsilon_2 \in (0,\varepsilon_1]$ exists such that

$$\forall \varepsilon \in (0,\varepsilon_2]\,, \qquad \eta \geqslant \sqrt{\frac{C_2\left(x;\frac{n}{3}(1-\varepsilon);c_n\right)}{C_1\left(x;\frac{2n}{3}(1+\varepsilon)\right)}}-1\,.$$

So, for such η, on Ω_x, for every $\varepsilon \in (0,\varepsilon_2]$,

$$\frac{C_1\left(x;\frac{2n}{3}(1+\varepsilon)\right)}{1+\eta} < \widehat{C}_{\text{window}}(\eta) < C_2\left(x;\frac{n}{3}(1-\varepsilon);c_n\right)(1+\eta)\,.$$

Making ε tend to zero in the above inequality yields the result.

Proof of Eq. (75) For any $a_n \in (T_n, n)$, by Eq. (85), on Ω_x, we have $D_{\widehat{m}(C)} \geqslant a_n > T_n$ for every $C \leqslant C_1(x;a_n)$, hence $\widehat{C}_{\text{thr.}}(T_n) \geqslant C_1(x;a_n)$ by Lemma 1 with $\Gamma_1 = C_1(x;a_n)$. So, on Ω_x,

$$\widehat{C}_{\text{thr.}}(T_n) \geqslant \sup_{a_n \in (T_n,n)} \left\{C_1(x;a_n)\right\} = C_1(x;T_n)\,.$$

By Eq. (85) with $b_n = T_n > c_n$, on Ω_x we have $D_{\widehat{m}(C)} \leqslant T_n$ for every $C \geqslant C_2(x;T_n;c_n)$, hence $\widehat{C}_{\text{thr.}}(T_n) \leqslant C_2(x;T_n;c_n)$ by Lemma 1 with $\Gamma_2 = C_2(x;T_n;c_n)$.

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

Proof of Eq. (76) Let $\alpha = \log(4\operatorname{card}(\mathscr{M})) \geqslant \log(8)$, since $\operatorname{card}(\mathscr{M}) \geqslant 2$ under the assumptions of Proposition 3. For every $z \geqslant 0$, by Eq. (75) with $x = z + \alpha$ and c_n replaced by $T_n/2$,

$$\mathbb{P}\left(\left(\widehat{C}_{\text{thr.}} - \sigma^2\right)^2 \geqslant 4\max\left\{ \left(1 - \frac{T_n}{n}\right)^{-2}, \left(\frac{T_n}{2n}\right)^{-2} \right\} \times \left[\mathscr{B}\left(\frac{T_n}{2}\right) + 2\sigma^2 \left(\sqrt{\frac{z+\alpha}{n}} + \frac{2(z+\alpha)}{n} \right) \right]^2 \right) \leqslant \mathrm{e}^{-z}. \tag{86}$$

Then, integrating Eq. (86) with respect to z —that is, using Lemma 3 below— we get that Eq. (76) holds true. Note that much smaller constants can be obtained by assuming that $\operatorname{card}(\mathscr{M})$ is large enough, or that $\log(\operatorname{card}(\mathscr{M}))/n$ is small enough. For instance, assuming that $100 \leqslant \operatorname{card}(\mathscr{M}) \leqslant \exp(n/100)$, we get

$$\mathbb{E}\left[\left(\widehat{C}_{\text{thr.}} - \sigma^2\right)^2 \right] \leqslant \max\left\{ \left(1 - \frac{T_n}{n}\right)^{-2}, \left(\frac{T_n}{2n}\right)^{-2} \right\} \left[12\mathscr{B}\left(\frac{T_n}{2}\right)^2 + \frac{102\sigma^4 \log\left(\operatorname{card}(\mathscr{M})\right)}{n} \right].$$

□

Lemma 3. *For a real-valued random variable Z, if some $a, b, c \geqslant 0$ exist such that for every $z \geqslant 0$,*

$$\mathbb{P}\left(Z \geqslant a + bz + cz^2\right) \leqslant \mathrm{e}^{-z},$$
$$\textit{then,} \qquad \mathbb{E}[Z] \leqslant a + 2b + 4c.$$

Lemma 3 is a classical integration exercise.

A.4. Computations about $\widehat{\sigma}^2_{m_0}$

The following proposition gives a general formula for the variance and MSE of the residual-variance estimator $\widehat{\sigma}^2_{m_0}$ defined by Eq. (70) in Section 6.1. Note that Proposition 4 and Lemma 4 below are classical results, see for instance [Ullah and Zinde-Walsh, 1992, Eq. (5)] or [Dette et al., 1998, Eq. (6)]. We state and prove them here for completeness.

In this subsection, for any matrix $M \in \mathscr{M}_n(\mathbb{R})$, $\operatorname{diag}(M)$ denotes the diagonal matrix of the diagonal elements of M and $\mathbf{1} = (1, \ldots, 1) \in \mathbb{R}^n$.

Proposition 4. *Let $F \in \mathbb{R}^n$, $\Pi \in \mathscr{M}_n(\mathbb{R})$ some orthogonal projection matrix such that $D = \operatorname{tr}(\Pi) < n$, and $\varepsilon \in \mathbb{R}^n$ some random vector with independent components. Assume that for all $i \in \{1, \ldots, n\}$*

$$\mathbb{E}[\varepsilon_i] = 0, \qquad \mathbb{E}\left[\varepsilon_i^2\right] = \sigma^2, \qquad \mathbb{E}\left[\varepsilon_i^3\right] = m_3, \qquad \textit{and} \qquad \mathbb{E}\left[\varepsilon_i^4\right] = m_4.$$

Let

$$\widehat{\sigma}^2 := \frac{1}{n - D}\left\| (I_n - \Pi)(F + \varepsilon) \right\|^2.$$

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

Then,

$$\operatorname{var}\left(\widehat{\sigma}^2\right) = V + \frac{4\left\|(I_n-\Pi)F\right\|^2\sigma^2}{(n-D)^2} + \frac{4\left\langle F,\,(I_n-\Pi)\operatorname{diag}(I_n-\Pi)\mathbf{1}\right\rangle}{(n-D)^2}m_3 \tag{87}$$

$$\textit{where}\quad V := \frac{1}{(n-D)^2}\left(\sum_{i=1}^{n}(1-\Pi_{i,i})^2\right)\left(m_4-3\sigma^4\right) + \frac{2}{n-D}\sigma^4\,,$$

$$\textit{and}\quad \mathbb{E}\left[\left(\widehat{\sigma}^2-\sigma^2\right)^2\right] = V + \frac{4\left\|(I_n-\Pi)F\right\|^2\sigma^2}{(n-D)^2} + \frac{\left\|(I_n-\Pi)F\right\|^4}{(n-D)^2} + \frac{4\left\langle F,\,(I_n-\Pi)\operatorname{diag}(I_n-\Pi)\mathbf{1}\right\rangle}{(n-D)^2}m_3\,. \tag{88}$$

In particular, if the ε_i are Gaussian,

$$\operatorname{var}\left(\widehat{\sigma}^2\right) = \frac{2\sigma^4}{n-D} + \frac{4\left\|(I_n-\Pi)F\right\|^2}{(n-D)^2}\sigma^2 \tag{89}$$

$$\mathbb{E}\left[\left(\widehat{\sigma}^2-\sigma^2\right)^2\right] = \frac{2\sigma^4}{n-D} + \frac{4\left\|(I_n-\Pi)F\right\|^2}{(n-D)^2}\sigma^2 + \frac{\left\|(I_n-\Pi)F\right\|^4}{(n-D)^2}\,. \tag{90}$$

Proof of Proposition 4. Applying Lemma 4 below with $M = I_n - \Pi$ yields Eq. (87), since $\widehat{\sigma}^2 = Z/(n-D)$ and M is an orthogonal projection matrix with $\operatorname{tr}(M) = n-D$. Eq. (88) follows, in combination with Eq. (71), since

$$\mathbb{E}\left[\left(\widehat{\sigma}^2-\sigma^2\right)^2\right] = \left(\mathbb{E}\left[\widehat{\sigma}^2\right]-\sigma^2\right)^2 + \operatorname{var}\left(\widehat{\sigma}^2\right).$$

In the Gaussian case, $m_3 = 0$ and $m_4 = 3\sigma^4$, hence

$$V = \frac{2\sigma^4}{n-D}\,,$$

which leads to Eq. (89) and (90). □

Lemma 4. *Let $F \in \mathbb{R}^n$, $M \in \mathscr{M}_n(\mathbb{R})$ a symmetric matrix, $\varepsilon \in \mathbb{R}^n$ some random vector with independent components such that for all $i \in \{1,\dots,n\}$,*

$$\mathbb{E}[\varepsilon_i] = 0\,, \qquad \mathbb{E}\left[\varepsilon_i^2\right] = \sigma^2\,, \qquad \mathbb{E}\left[\varepsilon_i^3\right] = m_3\,, \qquad \textit{and} \qquad \mathbb{E}\left[\varepsilon_i^4\right] = m_4\,.$$

Then, if $Z = \left\langle F+\varepsilon, M(F+\varepsilon)\right\rangle$,

$$\operatorname{var}(Z) = W + 4\left\|MF\right\|^2\sigma^2 + 4\left\langle F, M\operatorname{diag}(M)\mathbf{1}\right\rangle m_3 \tag{91}$$

$$\textit{where}\qquad W := \operatorname{tr}\left(\operatorname{diag}(M)^2\right)\left(m_4-3\sigma^4\right) + 2\operatorname{tr}(M^2)\sigma^4$$

$$\textit{satisfies}\qquad 0 \leqslant W \leqslant \left(2\sigma^4+m_4\right)\operatorname{tr}(M^2)\,. \tag{92}$$

Proof of Lemma 4. First note that

$$Z = \langle F, MF\rangle + \langle \varepsilon, M\varepsilon\rangle + 2\langle MF, \varepsilon\rangle\,.$$

Then,

$$\mathbb{E}[Z] = \langle F, MF\rangle + \mathbb{E}\big[\langle \varepsilon, M\varepsilon\rangle\big] = \langle F, MF\rangle + \sigma^2 \operatorname{tr}(M)\,.$$

Furthermore,

$$\begin{aligned}\mathbb{E}\big[Z^2\big] &= \langle F, MF\rangle^2 + \mathbb{E}\big[\langle \varepsilon, M\varepsilon\rangle^2\big] + 4\mathbb{E}\big[\langle MF, \varepsilon\rangle^2\big] \\ &\quad + 2\langle F, MF\rangle\sigma^2 \operatorname{tr}(M) + 4\mathbb{E}\big[\langle \varepsilon, M\varepsilon\rangle\langle MF, \varepsilon\rangle\big] \\ &= \langle F, MF\rangle^2 + \left(\sum_{i=1}^n M_{i,i}^2\right)\left(m_4 - 3\sigma^4\right) + \Big[\operatorname{tr}(M)^2 + 2\operatorname{tr}(M^\top M)\Big]\sigma^4 \\ &\quad + 4\|MF\|^2\sigma^2 + 2\langle F, MF\rangle\sigma^2\operatorname{tr}(M) + 4\left(\sum_{i,j=1}^n M_{i,i}M_{i,j}F_j\right)m_3\end{aligned}$$

where we used that

$$\begin{aligned}\mathbb{E}\big[\langle \varepsilon, M\varepsilon\rangle^2\big] &= \mathbb{E}\left[\left(\sum_{i,j}\varepsilon_i M_{i,j}\varepsilon_j\right)^2\right] = \sum_{i,j,k,\ell} M_{i,j}M_{k,\ell}\mathbb{E}[\varepsilon_i\varepsilon_j\varepsilon_k\varepsilon_\ell] \\ &= \left(\sum_{i=1}^n M_{i,i}^2\right)\left(m_4 - 3\sigma^4\right) + \Big(\operatorname{tr}(M)^2 + \operatorname{tr}(M^2) + \operatorname{tr}(M^\top M)\Big)\sigma^4\end{aligned}$$

and for any $G \in \mathbb{R}^n$ (here, $G = MF$),

$$\mathbb{E}\big[\langle G, \varepsilon\rangle^2\big] = \|G\|^2\sigma^2$$

$$\text{and} \qquad \mathbb{E}\big[\langle \varepsilon, M\varepsilon\rangle\langle G, \varepsilon\rangle\big] = \mathbb{E}\left[\sum_{i,j,k}\varepsilon_i M_{i,j}\varepsilon_j G_k\varepsilon_k\right] = \left(\sum_{i=1}^n M_{i,i}G_i\right)m_3\,.$$

Eq. (91) follows since M is symmetric, $\operatorname{var}(Z) = \mathbb{E}[Z^2] - \mathbb{E}[Z]^2$,

$$\sum_{i,j=1}^n M_{i,i}M_{i,j}F_j = \Big\langle F, M^\top \operatorname{diag}(M)\mathbf{1}\Big\rangle \qquad \text{and} \qquad \sum_{i=1}^n M_{i,i}^2 = \operatorname{tr}\left(\operatorname{diag}(M)^2\right)\,.$$

For proving Eq. (92), we remark that

$$\begin{gathered}W = \operatorname{tr}\big(\operatorname{diag}(M)^2\big)\left(m_4 - \sigma^4\right) + 2\Big[\operatorname{tr}(M^2) - \operatorname{tr}\big(\operatorname{diag}(M)^2\big)\Big]\sigma^4\,, \\ m_4 \geqslant \sigma^4\,, \quad \text{and} \quad 0 \leqslant \operatorname{tr}\left(\operatorname{diag}(M)^2\right) \leqslant \operatorname{tr}(M^2)\end{gathered}$$

since M is symmetric. □

Appendix B: Algorithms

B.1. Computation of the full path $(\widehat{m}(C))_{C\geqslant 0}$ in Algorithms 1, 3, 4, and 5

One can formulate the first step in Algorithms 1, 3, 4, and 5 as computing, for every $C \geqslant 0$,

$$\widehat{m}(C) \in \operatorname*{argmin}_{m\in\mathscr{M}}\big\{f(m) + Cg(m)\big\} \tag{93}$$

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

for some functions $f, g : \mathscr{M} \to \mathbb{R}$, where $\mathscr{M}$ is assumed to be finite. In the most general case (Algorithm 5), $f(m) = \widehat{\mathscr{R}}_n(\widehat{s}_m)$ and $g(m) = \mathrm{pen}_0(m)$. Particular cases (Algorithms 1, 3, and 4) follow.

This subsection explains how to compute the full path $(\widehat{m}(C))_{C \geqslant 0}$ defined by Eq. (93), given $(f(m))_{m \in \mathscr{M}}$ and $(g(m))_{m \in \mathscr{M}}$, with at most $\mathscr{O}(\mathrm{card}(\mathscr{M})^2)$ operations, and much less in practice. The material presented here is adapted from [Arlot and Massart, 2009, Section 3.2]. Similar results —with a bit less details and formulated in specific frameworks where $\mathscr{M} \subset \mathbb{N}$— have been proved earlier by [Lebarbier, 2002, Lemma 4.4.1], [Lavielle, 2005, Proposition 2.1], and [Zwald, 2005, section 6.4.3].

First, remark that the definition (93) of $\widehat{m}(C)$ can be ambiguous. Let us choose a strict total order $\prec$ on $\mathscr{M}$ such that g is non-decreasing, which is always possible since $\mathscr{M}$ is finite. Then, by convention, for every $C \geqslant 0$, $\widehat{m}(C)$ is defined as

$$\widehat{m}(C) = \min_{\prec} \mathscr{E}(C) \quad \text{where} \quad \mathscr{E}(C) := \operatorname*{argmin}_{m \in \mathscr{M}} \left\{ f(m) + C g(m) \right\}. \tag{94}$$

The main reason why the whole trajectory $(\widehat{m}(C))_{C \geqslant 0}$ can be computed efficiently is its particular shape. Indeed, the proof of Proposition 5 below shows that $C \mapsto \widehat{m}(C)$ is piecewise constant and non-increasing for $\prec$. Then, the whole trajectory $(\widehat{m}(C))_{C \geqslant 0}$ can be written as

$$\forall i \in \{0, \dots, i_{\max}\}, \quad \forall C \in [C_i, C_{i+1}), \qquad \widehat{m}(C) = m_i \tag{95}$$

where $i_{\max} \in \{0, \dots, \mathrm{card}(\mathscr{M}) - 1\}$ is the number of jumps, $(C_i)_{0 \leqslant i \leqslant i_{\max}+1}$ is an increasing sequence of nonnegative reals (the location of the jumps) with $C_0 = 0$ and $C_{i_{\max}+1} = +\infty$, and $(m_i)_{0 \leqslant i \leqslant i_{\max}}$ is a non-increasing sequence of elements of $\mathscr{M}$.

Algorithm 7. `Input`: $(f(m))_{m \in \mathscr{M}}$, $(g(m))_{m \in \mathscr{M}}$, *and* $\prec$ *some strict total order on* $\mathscr{M}$ *such that* g *is non-decreasing.*
Initialization: $C_0 := 0$ *and* $m_0 := \min_{\prec} \operatorname{argmin}_{m \in \mathscr{M}} \{ f(m) \}$.
Step i, $i \geqslant 1$: *Let*

$$\mathscr{G}(m_{i-1}) := \left\{ m \in \mathscr{M} \ \text{ s.t. } \ f(m) > f(m_{i-1}) \quad \text{and} \quad g(m) < g(m_{i-1}) \right\}.$$

If $\mathscr{G}(m_{i-1}) = \emptyset$, *then put* $C_i = +\infty$, $i_{\max} = i - 1$ *and stop. Otherwise, define*

$$C_i := \min \left\{ \frac{f(m) - f(m_{i-1})}{g(m_{i-1}) - g(m)} \ \text{ s.t. } \ m \in \mathscr{G}(m_{i-1}) \right\} \tag{96}$$

$$\text{and} \qquad m_i := \min_{\prec} \mathscr{F}_i \qquad \text{with} \quad \mathscr{F}_i := \operatorname*{argmin}_{m \in \mathscr{G}(m_{i-1})} \left\{ \frac{f(m) - f(m_{i-1})}{g(m_{i-1}) - g(m)} \right\}.$$

`Output`: $(C_i)_{0 \leqslant i \leqslant i_{\max}+1}$ *and* $(m_i)_{0 \leqslant i \leqslant i_{\max}}$, *which describe according to Eq.* (95) *the full trajectory* $(\widehat{m}(C))_{C \geqslant 0}$ *defined by Eq.* (94).

Proposition 5 (Correctness of Algorithm 7)**.** *For every* $C \geqslant 0$, *let* $\widehat{m}(C)$ *be defined by Eq.* (94). *Assume* $\mathscr{M}$ *is finite. Then, Algorithm 7 terminates and* $i_{\max} \leqslant \mathrm{card}(\mathscr{M}) - 1$. *Furthermore, Algorithm 7 is correct, that is,* $(C_i)_{0 \leqslant i \leqslant i_{\max}+1}$ *is increasing and* $\forall i \in \{0, \dots, i_{\max} - 1\}$, $\forall C \in [C_i, C_{i+1})$, $\widehat{m}(C) = m_i$.

Proposition 5 also gives an upper bound on the computational complexity of Algorithm 7: since the complexity of each step is $\mathcal{O}(\operatorname{card}\mathscr{M})$, Algorithm 7 requires less than $\mathcal{O}(i_{\max}\operatorname{card}\mathscr{M}) \leqslant \mathcal{O}((\operatorname{card}\mathscr{M})^2)$ operations. In general, this upper bound is pessimistic since $i_{\max} \ll \operatorname{card}\mathscr{M}$.

Proof of Proposition 5. First, since $\mathscr{M}$ is finite, $\mathscr{G}(m_{i-1})$ is also finite and m_i is well-defined as soon as $\mathscr{G}(m_{i-1}) \neq \emptyset$, which holds for every $i \leqslant i_{\max}$. Moreover, by construction, $g(m_i)$ decreases with i, so that all the $m_i \in \mathscr{M}$ are different; hence, Algorithm 7 terminates and $i_{\max}+1 \leqslant \operatorname{card}(\mathscr{M})$. Notice also that C_i can always be defined by Eq. (96) with the convention $\min\emptyset = +\infty$. We now prove by induction that the following property holds true for every $i \in \{0,\dots,i_{\max}\}$, which implies that Proposition 5 holds true:

$$\mathscr{P}_i : \qquad C_i < C_{i+1} \quad \text{and} \quad \forall C \in [C_i, C_{i+1}), \quad \widehat{m}(C) = m_i \,.$$

$\mathscr{P}_0$ holds true By definition of C_1, since $\mathscr{M}$ is finite, $C_1 > 0$ (it may be equal to $+\infty$ if $\mathscr{G}(m_0) = \emptyset$). For $C = C_0 = 0$, the definition of m_0 is the one of $\widehat{m}(0)$, so that $\widehat{m}(C) = m_0$. For $C \in (0, C_1)$, Lemma 5 below shows that either $\widehat{m}(C) = \widehat{m}(0) = m_0$ or $\widehat{m}(C) \in \mathscr{G}(m_0)$. In the latter case, by definition of C_1,

$$\frac{f\big(\widehat{m}(C)\big) - f(m_0)}{g(m_0) - g\big(\widehat{m}(C)\big)} \geqslant C_1 > C$$

hence

$$f\big(\widehat{m}(C)\big) + Cg\big(\widehat{m}(C)\big) > f(m_0) + Cg(m_0)$$

which contradicts the definition of $\widehat{m}(C)$. Therefore, $\mathscr{P}_0$ holds true.

$\mathscr{P}_i \Rightarrow \mathscr{P}_{i+1}$ for every $i \in \{0,\dots,i_{\max}-1\}$ Assume that $\mathscr{P}_i$ holds true. First, we have to prove that $C_{i+2} > C_{i+1}$. If $i = i_{\max}-1$, this is clear since $C_{i_{\max}+1} = +\infty$. Otherwise, $C_{i+2} < +\infty$ and m_{i+2} exists. Then, by definition of m_{i+2} and C_{i+2} (resp. m_{i+1} and C_{i+1}), we have

$$f(m_{i+2}) - f(m_{i+1}) = C_{i+2}\big[g(m_{i+1}) - g(m_{i+2})\big] \tag{97}$$
$$f(m_{i+1}) - f(m_i) = C_{i+1}\big[g(m_i) - g(m_{i+1})\big] \,. \tag{98}$$

Moreover, $m_{i+2} \in \mathscr{G}(m_{i+1}) \subset \mathscr{G}(m_i)$ and $m_{i+2} \prec m_{i+1}$ (because g is non-decreasing). Using again the definition of C_{i+1}, we have

$$f(m_{i+2}) - f(m_i) > C_{i+1}\big[g(m_i) - g(m_{i+2})\big] \tag{99}$$

(the inequality is strict since otherwise, we would have $m_{i+2} \in \mathscr{F}_{i+1}$ and $m_{i+2} \prec m_{i+1} = \min_{\prec}\mathscr{F}_{i+1}$, which is not possible). The difference of Eq. (99) and (98) yields

$$f(m_{i+2}) - f(m_{i+1}) > C_{i+1}\big[g(m_{i+1}) - g(m_{i+2})\big] \,.$$

By Eq. (97), we deduce that

$$C_{i+2}\big[g(m_{i+1}) - g(m_{i+2})\big] > C_{i+1}\big[g(m_{i+1}) - g(m_{i+2})\big] \,,$$

hence $C_{i+2} > C_{i+1}$ since $g(m_{i+1}) > g(m_{i+2})$.

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

Second, we prove that $\widehat{m}(C_{i+1}) = m_{i+1}$. From $\mathscr{P}_i$, we know that for every $m \in \mathscr{M}$, for every $C \in [C_i, C_{i+1})$, $f(m_i) + Cg(m_i) \leqslant f(m) + Cg(m)$. Taking the limit when C tends to C_{i+1}, it follows that $m_i \in \mathscr{E}(C_{i+1})$. By Eq. (98), we then have $m_{i+1} \in \mathscr{E}(C_{i+1})$. Now, let m' be any element of $\mathscr{E}(C_{i+1})$. By Lemma 5 with $C = C_i$, $m = m_i = \widehat{m}(C_i) \in \mathscr{E}(C_i)$ and $C' = C_{i+1} > C_i$, we have either (a) $f(m') = f(m_i)$ and $g(m') = g(m_i)$ or (b) $m' \in \mathscr{G}(m_i)$; case (c) is excluded since $m_i = \widehat{m}(C_i)$. In case (a), $g(m') = g(m_i) > g(m_{i+1})$, hence $m_{i+1} \prec m'$ because g is non-decreasing. In case (b), notice that $m_i, m' \in \mathscr{E}(C_{i+1})$ implies $f(m') + C_{i+1}g(m') = f(m_i) + C_{i+1}g(m_i)$. Since $m' \in \mathscr{G}(m_i)$, we get that $m' \in \mathscr{F}_{i+1}$. Then, by definition of m_{i+1}, we have $m_{i+1} \preceq m'$. Overall, we have proved that m_{i+1} belongs to $\mathscr{E}(C_{i+1})$ and is smaller than any element m' of $\mathscr{E}(C_{i+1})$, which proves that $m_{i+1} = \min_{\prec} \mathscr{E}(C_{i+1}) = \widehat{m}(C_{i+1})$.

Last, we have to prove that $\widehat{m}(C') = m_{i+1}$ for every $C' \in (C_{i+1}, C_{i+2})$. From the last statement of Lemma 5 with $C = C_{i+1}$, we have either $\widehat{m}(C') = \widehat{m}(C_{i+1}) = m_{i+1}$ or $\widehat{m}(C') \in \mathscr{G}(\widehat{m}(C_{i+1})) = \mathscr{G}(m_{i+1})$. In the latter case (in which $\mathscr{G}(m_{i+1}) \neq \emptyset$ hence $C_{i+2} < \infty$), by definition of C_{i+2},

$$\frac{f\big(\widehat{m}(C')\big) - f(m_{i+1})}{g(m_{i+1}) - g\big(\widehat{m}(C')\big)} \geqslant C_{i+2} > C'$$

so that

$$f\big(\widehat{m}(C')\big) + C'g\big(\widehat{m}(C')\big) > f(m_{i+1}) + C'g(m_{i+1})$$

which contradicts the definition of $\widehat{m}(C')$. Therefore, $\widehat{m}(C') = m_{i+1}$, which ends proving $\mathscr{P}_{i+1}$. □

The following lemma is used in the proof of Proposition 5 above.

Lemma 5. *With the notations of Proposition 5 and its proof, if $0 \leqslant C < C'$, $m \in \mathscr{E}(C)$ and $m' \in \mathscr{E}(C')$, then one of the following statements holds true*:

(*a*) $f(m) = f(m')$ *and* $g(m) = g(m')$.

(*b*) $f(m) < f(m')$ *and* $g(m) > g(m')$.

(*c*) $C = 0$, $f(m) = f(m')$ *and* $g(m) > g(m')$, *hence* $m \neq \widehat{m}(0)$.

In particular, for any $0 \leqslant C < C'$, we have either $\widehat{m}(C) = \widehat{m}(C')$ or $\widehat{m}(C') \in \mathscr{G}(\widehat{m}(C))$.

Proof of Lemma 5. By definition of $\mathscr{E}(C)$ and $\mathscr{E}(C')$,

$$f(m) + Cg(m) \leqslant f(m') + Cg(m') \tag{100}$$

$$f(m') + C'g(m') \leqslant f(m) + C'g(m)\,. \tag{101}$$

Summing Eq. (100) and (101) gives $(C' - C)g(m') \leqslant (C' - C)g(m)$ so that

$$g(m') \leqslant g(m)\,. \tag{102}$$

Since $C \geqslant 0$, Eq. (100) and (102) give $f(m) + Cg(m) \leqslant f(m') + Cg(m)$, that is

$$f(m) \leqslant f(m')\,. \tag{103}$$

If $g(m) = g(m')$, Eq. (101) and (103) imply $f(m') = f(m)$ hence (a) is satisfied. Otherwise, $g(m) > g(m')$ by Eq. (102), and Eq. (100) implies $f(m) < f(m')$ or $C = 0$. If $f(m) < f(m')$, (b) holds true. Otherwise, $f(m) = f(m')$ and $C = 0$. Since $g(m') < g(m)$, we get $m' \prec m$ hence $m \neq \widehat{m}(0)$.

The last statement follows by taking $m = \widehat{m}(C)$ and $m' = \widehat{m}(C')$, which excludes case (c). In case (a), $\mathscr{E}(C) = \mathscr{E}(C')$ hence $\widehat{m}(C) = \widehat{m}(C')$. In case (b), $\widehat{m}(C') \in \mathscr{G}(\widehat{m}(C))$. □

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

B.2. Computation of $\widehat{C}_{\mathrm{window}}$ in step 2 of Algorithms 1, 3, 4, and 5

Step 2 of Algorithms 1, 3, 4, and 5 require to localize a jump in the trajectory $(\mathscr{C}_{\widehat{m}(C)})_{C\geqslant 0}$, given the path $(\widehat{m}(C))_{C\geqslant 0}$ and some complexity measure $(\mathscr{C}_m)_{m\in\mathscr{M}}$. Although the maximal jump is straightforward to localize, Theorem 1 suggests to look for the largest jump over a geometrical window of values of C, that is, $\widehat{C}_{\mathrm{window}}$ as defined by Eq. (19) in Section 2.5. This section explains how $\widehat{C}_{\mathrm{window}}$ can be computed efficiently —with complexity $\mathscr{O}(i_{\max}\log(i_{\max})) = \mathscr{O}(\mathrm{card}(\mathscr{M})\log[\mathrm{card}(\mathscr{M})]$— given $(\mathscr{C}_{\widehat{m}(C)})_{C\geqslant 0}$.

Let us consider a slightly more general problem: given some $\alpha > \beta > 0$, compute

$$\widehat{\mathscr{E}}^{\mathrm{gal}}_{\mathrm{window}} := \operatorname*{argmax}_{C\geqslant 0} \left\{ \mathscr{C}_{\widehat{m}(\beta C)} - \mathscr{C}_{\widehat{m}(\alpha C)} \right\} . \tag{104}$$

Note that $\widehat{\mathscr{E}}^{\mathrm{gal}}_{\mathrm{window}}$ is usually not reduced to a singleton, but can be an interval or a finite union of intervals.

From Eq. (95) in Appendix B.1, the path $(\mathscr{C}_{\widehat{m}(C)})_{C\geqslant 0}$ is piecewise constant and can be fully described with a small number of parameters: writing $\mathscr{C}_i = \mathscr{C}_{m_i}$,

$$\forall i \in \{0,\dots,i_{\max}\}, \quad \forall C \in [C_i, C_{i+1}), \qquad \mathscr{C}_{\widehat{m}(C)} = \mathscr{C}_i . \tag{105}$$

Given this description of $(\mathscr{C}_{\widehat{m}(C)})_{C\geqslant 0}$, Algorithm 8 below determines the set $\widehat{\mathscr{E}}^{\mathrm{gal}}_{\mathrm{window}}$, as proved by Proposition 6 .

Algorithm 8. `Input`: *$(C_i)_{0\leqslant i\leqslant i_{\max}+1}$ an increasing sequence of nonnegative reals with $C_0 = 0$ and $C_{i_{\max}+1} = +\infty$, and $(\mathscr{C}_i)_{0\leqslant i\leqslant i_{\max}}$ a sequence of real numbers.*

1. *If $i_{\max} = 0$, define $\widehat{\mathscr{E}}^{\mathrm{gal}}_{\mathrm{window}} := [0,+\infty)$ and stop.*
2. *Otherwise, compute $\overline{C} = (C_1/\beta,\dots,C_{i_{\max}}/\beta, C_1/\alpha,\dots,C_{i_{\max}}/\alpha) \in \mathbb{R}^{2i_{\max}}$ and $\overline{\Delta} = (\mathscr{C}_1 - \mathscr{C}_0,\dots,\mathscr{C}_{i_{\max}} - \mathscr{C}_{i_{\max}-1}, \mathscr{C}_0 - \mathscr{C}_1,\dots,\mathscr{C}_{i_{\max}-1} - \mathscr{C}_{i_{\max}}) \in \mathbb{R}^{2i_{\max}}$.*
3. *Sort $\overline{C}$ and $\overline{\Delta}$ according to $\overline{C}$, that is, compute $\overline{C^\sigma} = (\overline{C}_{\sigma(i)})_{1\leqslant i\leqslant 2i_{\max}}$ and $\overline{\Delta^\sigma} = (\overline{\Delta}_{\sigma(i)})_{1\leqslant i\leqslant 2i_{\max}}$ where σ is some permutation of $\{1,\dots,2i_{\max}\}$ such that $\overline{C}_{\sigma(1)} \leqslant \dots \leqslant \overline{C}_{\sigma(2i_{\max})}$.*
4. *Compute $W := \mathrm{cumsum}(\overline{\Delta^\sigma}) \in \mathbb{R}^{2i_{\max}}$, that is, for every $i \in \{1,\dots,2i_{\max}\}$,*

$$W_i = \sum_{j=1}^{i} \overline{\Delta^\sigma_j} .$$

5. *Compute $V \in \mathbb{R}^{2i_{\max}}$ such that, for every $i \in \{1,\dots,2i_{\max}\}$, $V_i = W_i$ if $\overline{C^\sigma_i} < \overline{C^\sigma_{i+1}}$ and $V_i = -\infty$ otherwise.*
6. *Determine $\mathscr{K} := \mathrm{argmax}_{i\in\{1,\dots,2i_{\max}\}} V_i$.*
7. *Define $\widehat{\mathscr{E}}^{\mathrm{gal}}_{\mathrm{window}} := \bigcup_{k\in\mathscr{K}} [\overline{C^\sigma_k}, \overline{C^\sigma_{k+1}})$ with $\overline{C^\sigma_{2i_{\max}+1}} = +\infty$.*

`Output`: $\widehat{\mathscr{E}}^{\mathrm{gal}}_{\mathrm{window}}$.

Proposition 6 (Correctness of Algorithm 8)**.** *Algorithm 8 is correct, that is, it terminates and its output $\widehat{\mathscr{E}}^{\mathrm{gal}}_{\mathrm{window}}$ actually satisfies Eq. (104) provided Eq. (105) holds true.*

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

Proof of Proposition 6. If $i_{\max} = 0$, $C \mapsto \mathscr{C}_{\widehat{m}(C)}$ is constant over $[0, +\infty)$ so Algorithm 8 is correct. Otherwise, Eq. (105) can be rewritten as

$$\forall C \geqslant 0\,, \quad \mathscr{C}_{\widehat{m}(C)} = \mathscr{C}_0 + \sum_{i=1}^{i_{\max}} (\mathscr{C}_i - \mathscr{C}_{i-1}) \mathbb{1}_{C \geqslant C_i}$$

hence, for every $C \geqslant 0$, using the notations of Algorithm 8,

$$\begin{aligned} \mathscr{C}_{\widehat{m}(\beta C)} - \mathscr{C}_{\widehat{m}(\alpha C)} &= \sum_{i=1}^{i_{\max}} (\mathscr{C}_i - \mathscr{C}_{i-1}) \mathbb{1}_{\beta C \geqslant C_i} - \sum_{i=1}^{i_{\max}} (\mathscr{C}_i - \mathscr{C}_{i-1}) \mathbb{1}_{\alpha C \geqslant C_i} \\ &= \sum_{i=1}^{2i_{\max}} \overline{\Delta}_i \mathbb{1}_{C \geqslant \overline{C}_i} = \sum_{i=1}^{2i_{\max}} \overline{\Delta_i^\sigma} \mathbb{1}_{C \geqslant \overline{C_i^\sigma}} = \sum_{i=1}^{2i_{\max}} W_i \mathbb{1}_{C \in [\overline{C_i^\sigma}, \overline{C_{i+1}^\sigma})} \\ &= \sum_{i=1}^{2i_{\max}} V_i \mathbb{1}_{C \in [\overline{C_i^\sigma}, \overline{C_{i+1}^\sigma})} \end{aligned} \tag{106}$$

with the conventions $\overline{C^\sigma_{2i_{\max}+1}} = +\infty$ and $\infty \mathbb{1}_{C \in \emptyset} = 0$. For the last equality, we use the fact that when $[\overline{C_i^\sigma}, \overline{C_{i+1}^\sigma})$ is empty —which corresponds to values of $\overline{C_i^\sigma}$ that are equal to $C_j/\beta = C_k/\alpha$ for some $j, k \in \{1, \ldots, i_{\max}\}$—, the value of $W_i \mathbb{1}_{C \in [\overline{C_i^\sigma}, \overline{C_{i+1}^\sigma})}$ is zero whatever W_i, hence W_i can be changed into V_i.

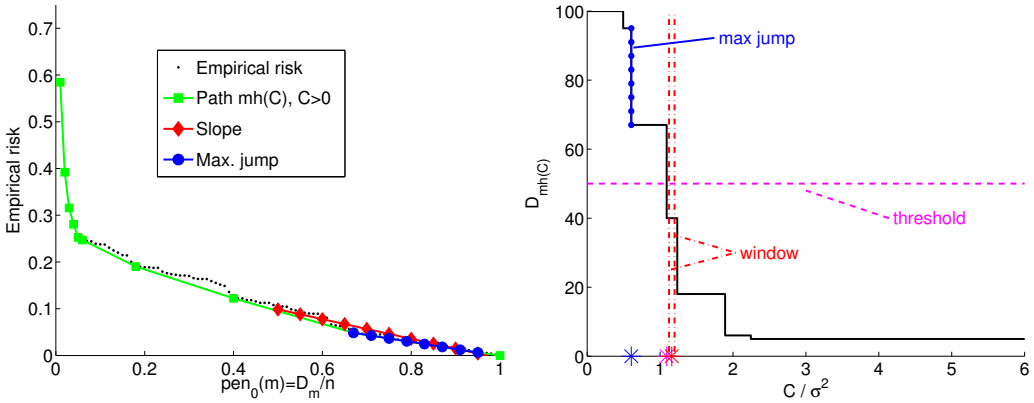
By Eq. (106),

$$\sup_{C \geqslant 0} \left\{ \mathscr{C}_{\widehat{m}(\beta C)} - \mathscr{C}_{\widehat{m}(\alpha C)} \right\} = \max_{1 \leqslant i \leqslant 2i_{\max}} V_i$$

and the supremum is attained exactly at the values of C belonging to some interval $[\overline{C_k^\sigma}, \overline{C_{k+1}^\sigma})$ with $k \in \mathscr{K} = \operatorname{argmax}_i V_i$. In other words, Algorithm 8 is correct. □

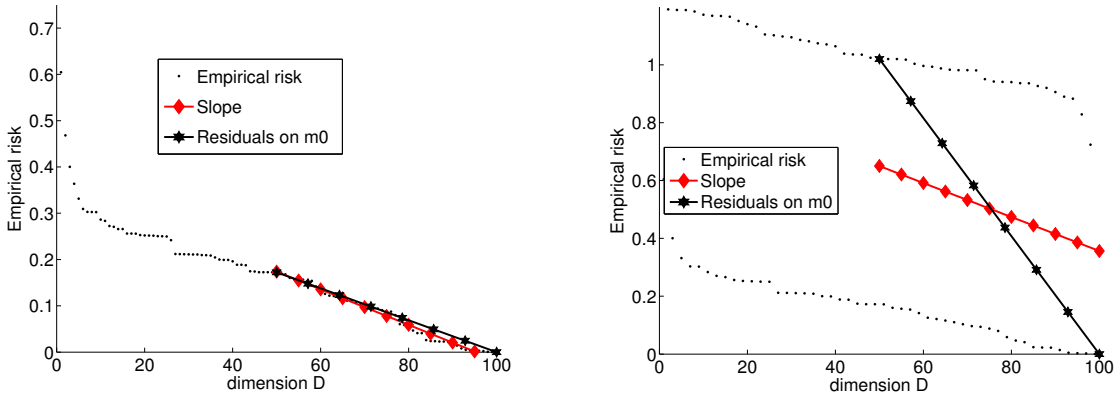
Appendix C: More figures and experimental results

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019



(a) L-curve (as in Figure 4). (b) Plot of $C \mapsto D_{\widehat{m}(C)}$.

FIGURE 9. *On the same sample, visualization of the three versions of Algorithm 1 and of* $\widehat{C}_{\text{slope}}$. *'Easy' setting, see Appendix D for details.*



(a) 'Easy' setting. (b) 'Hard' setting.

FIGURE 10. *Slope estimation* $\widehat{C}_{\text{slope}}$ *vs. residual-based variance estimator* $\widehat{\sigma}^2_{m_0}$. *See Appendix D for details.*

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

$\widehat{C}$	$\mathbb{E}[\widehat{C}/\sigma^2]$	$\sqrt{\mathrm{var}(\widehat{C})}/\sigma^2$	$\mathbb{E}[(\widehat{C}-\sigma^2)^2]/\sigma^4$	risk ratio
$\widehat{C}_{\mathrm{max\,j.}}$	1.09	0.257	0.0749	1.309 ± 0.003
$\widehat{C}_{\mathrm{thr.}}$, $T_n = n/10$	3.12	1.281	6.140	1.647 ± 0.004
$\widehat{C}_{\mathrm{thr.}}$, $T_n = n/\log(n)$	1.60	0.469	0.584	1.310 ± 0.002
$\widehat{C}_{\mathrm{thr.}}$, $\mathbf{T_n = n/2}$	1.13	0.229	0.0683	1.278 ± 0.003
$\widehat{C}_{\mathrm{thr.}}$, $T_n = 9n/10$	0.84	0.239	0.0826	1.621 ± 0.083
$\widehat{C}_{\mathrm{window}}$, $\eta = 1/n$	1.09	0.257	0.0745	1.309 ± 0.003
$\widehat{C}_{\mathrm{window}}$, $\eta = \mathbf{1}/\sqrt{\mathbf{n}}$	1.10	0.256	0.0752	1.308 ± 0.003
$\widehat{C}_{\mathrm{window}}$, $\eta = 1.5/\sqrt{n}$	1.10	0.258	0.0776	1.307 ± 0.003
$\widehat{C}_{\mathrm{window}}$, $\eta = \sqrt{\log(n)/n}$	1.12	0.263	0.0829	1.304 ± 0.003
$\widehat{C}_{\mathrm{window}}$, $\eta = 2\sqrt{\log(n)/n}$	1.17	0.286	0.110	1.294 ± 0.003
$\widehat{C}_{\mathrm{slope}}$, $D_0 = n/10$	1.15	0.181	0.0544	1.243 ± 0.002
$\widehat{C}_{\mathrm{slope}}$, $D_0 = n/\log(n)$	1.09	0.188	0.0437	1.260 ± 0.002
$\widehat{C}_{\mathrm{slope}}$, $\mathbf{D_0 = n/2}$	1.05	0.228	0.0543	1.313 ± 0.003
$\widehat{C}_{\mathrm{slope}}$, $D_0 = 9n/10$	1.02	0.478	0.229	1.672 ± 0.009
CAPUSHE	1.05	0.291	0.0873	1.410 ± 0.005
median	1.08	0.229	0.0588	1.301 ± 0.003
consensus	–	–	–	1.306 ± 0.003
consensus when no reject	–	–	–	1.298 ± 0.003
$\widehat{\sigma}^2_{m_0}$, $D_{m_0} = n/10$	1.23	0.180	0.0862	1.237 ± 0.002
$\widehat{\sigma}^2_{m_0}$, $D_{m_0} = n/\log(n)$	1.12	0.176	0.0443	1.241 ± 0.002
$\widehat{\sigma}^2_{m_0}$, $D_{m_0} = n/2$	1.05	0.211	0.0469	1.304 ± 0.003
$\widehat{\sigma}^2_{m_0}$, $D_{m_0} = n/2+1$	1.05	0.213	0.0478	1.305 ± 0.003
$\widehat{\sigma}^2_{m_0}$, $D_{m_0} = 9n/10$	1.02	0.455	0.2080	1.641 ± 0.008
C_p (known σ^2)	–	–	–	1.269 ± 0.003
$C_p \times 1.12$ (known σ^2)	–	–	–	1.251 ± 0.002

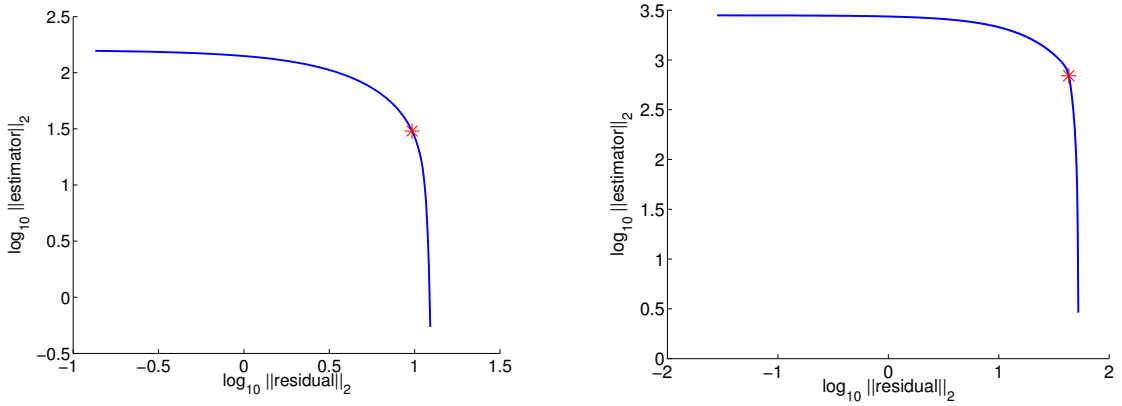
TABLE 2. *Algorithms 1–2, 'easy' setting: distribution of $\widehat{C}$ and model-selection performance, with various definitions for $\widehat{C}$ and various parameters for each definition. The risk ratio is $\mathbb{E}[\|\widehat{F}_{\widehat{m}} - F\|^2 / \inf_{m\in\mathscr{M}} \|\widehat{F}_m - F\|^2]$. Reported values are empirical estimates obtained from $N = 10\,000$ independent samples. For the risk ratio, error bars are equal to the standard deviation of the ratio $\|\widehat{F}_{\widehat{m}} - F\|^2 / \inf_{m\in\mathscr{M}} \|\widehat{F}_m - F\|^2$ divided by $\sqrt{N}$.*

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

$\widehat{C}$	$\mathbb{E}\left[\widehat{C}/\sigma^2\right]$	$\sqrt{\mathrm{var}(\widehat{C})}/\sigma^2$	$\mathbb{E}\left[(\widehat{C}-\sigma^2)^2\right]/\sigma^4$	risk ratio
$\widehat{C}_{\text{maxj.}}$	1.10	0.259	0.076	1.291 ± 0.003
$\widehat{C}_{\text{thr.}}$, $T_n = n/10$	3.38	1.390	7.57	1.661 ± 0.004
$\widehat{C}_{\text{thr.}}$, $T_n = n/\log(n)$	1.59	0.462	0.563	1.285 ± 0.002
$\widehat{C}_{\text{thr.}}$, $\mathbf{T_n = n/2}$	1.13	0.231	0.0703	1.258 ± 0.002
$\widehat{C}_{\text{thr.}}$, $T_n = 9n/10$	0.86	0.236	0.077	1.566 ± 0.008
$\widehat{C}_{\text{window}}$, $\eta = 1/n$	1.09	0.257	0.0746	1.292 ± 0.003
$\widehat{C}_{\text{window}}$, $\eta = \mathbf{1}/\sqrt{\mathbf{n}}$	1.10	0.257	0.0762	1.288 ± 0.003
$\widehat{C}_{\text{window}}$, $\eta = 1.5/\sqrt{n}$	1.11	0.258	0.078	1.288 ± 0.003
$\widehat{C}_{\text{window}}$, $\eta = \sqrt{\log(n)/n}$	1.12	0.263	0.0827	1.287 ± 0.003
$\widehat{C}_{\text{window}}$, $\eta = 2\sqrt{\log(n)/n}$	1.17	0.285	0.109	1.275 ± 0.003
$\widehat{C}_{\text{slope}}$, $D_0 = n/10$	1.54	0.188	0.328	1.268 ± 0.002
$\widehat{C}_{\text{slope}}$, $D_0 = n/\log(n)$	1.65	0.193	0.46	1.291 ± 0.002
$\widehat{C}_{\text{slope}}$, $\mathbf{D_0 = n/2}$	2.36	0.231	1.89	1.437 ± 0.003
$\widehat{C}_{\text{slope}}$, $D_0 = 9n/10$	20.2	2.07	374	3.68 ± 0.016
CAPUSHE	2.77	1.66	5.87	1.562 ± 0.005
median	1.16	0.253	0.0911	1.260 ± 0.002
consensus	–	–	–	1.285 ± 0.003
consensus when no reject	–	–	–	1.266 ± 0.003
$\widehat{\sigma}^2_{m_0}$, $D_{m_0} = n/10$	5.44	0.473	19.9	2.055 ± 0.006
$\widehat{\sigma}^2_{m_0}$, $D_{m_0} = n/\log(n)$	1.12	0.176	0.0443	1.223 ± 0.002
$\widehat{\sigma}^2_{m_0}$, $D_{m_0} = n/2$	8.94	0.828	63.7	2.577 ± 0.006
$\widehat{\sigma}^2_{m_0}$, $D_{m_0} = n/2+1$	1.05	0.213	0.0478	1.285 ± 0.003
$\widehat{\sigma}^2_{m_0}$, $D_{m_0} = 9n/10$	38.9	3.95	1450	6.11 ± 0.011
C_p (known σ^2)	–	–	–	1.252 ± 0.003
$C_p \times 1.12$ (known σ^2)	–	–	–	1.232 ± 0.002

TABLE 3. *Same as Table 2 for the 'hard' setting.*

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019



(a) $n = 100$ (b) $n = 2000$

FIGURE 11. *L-curve* $(\log_{10}\|Y - \widehat{F}_\lambda\|, \log_{10}\|\widehat{F}_\lambda\|)_{\lambda>0}$ *in the 'kernel ridge' framework. See Appendix D for details. The red star shows the position of the oracle (minimum of the risk* $n^{-1}\|F - \widehat{F}_\lambda\|^2$*). The "elbow" is clearly localized for* $n = 2000$ *(and close to the oracle), but not for* $n = 100$*.*

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

Appendix D: Detailed information about figures and simulation experiments

This section provides all details necessary to reproduce the figures and simulation experiments reported throughout the article.

D.1. Data and estimators

All experiments are made within the fixed-design regression framework described in Section 2.1, with two main kinds of estimator collections and data.

Least-squares framework ('easy'/'hard') All figures and tables, except Figures 3 and 11, consider data and estimators as follows. Data satisfy

$$Y = F + \varepsilon \in \mathbb{R}^n$$

with independent Gaussian noise $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, $\sigma^2 = 1/4$, $n = 100$,

$$F_i = \frac{C_n}{i} \quad \text{and} \quad C_n = \left(\sum_{i=1}^{n} \frac{1}{i^2} \right)^{-1/2} .$$

The choice of C_n ensures that $n^{-1} \|F\|^2 = 1$.

The estimators considered are least-squares (projection) estimators with one among the following two collections of models $(S_m)_{1 \leqslant m \leqslant n}$:

— 'easy' setting: for every $m \in \{1, \dots, n\}$, $S_m = S_m^{\text{easy}}$ is the linear span of the first m vectors of the canonical basis of $\mathbb{R}^n$.

— 'hard' setting: for every $m \in \{1, \dots, n\}$, $S_m = S_m^{\text{hard}}$ is the linear span of the first m vectors of the canonical basis of $\mathbb{R}^n$ if m is odd, and $S_m = S_m^{\text{hard}}$ is the linear span of the last m vectors of the canonical basis of $\mathbb{R}^n$ if m is even.

Both settings correspond to (ordered) variable selection with an orthogonal design, after having transformed the data conveniently according to the design matrix. In the easy case, the variables are ordered by decreasing order of magnitude. In the hard case, some incertainty remains about the correct order (ascending or descending), and the two options are considered alternatively (depending on the parity of n). Of course, models S_m^{hard} with m odd are very poor, but this can be unknown before seeing the data.

Kernel ridge framework Figures 3 and 11 consider data and estimators as follows. Data satisfy

$$Y = F + \varepsilon \in \mathbb{R}^n$$

with independent Gaussian noise $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, $\sigma^2 = 1$,

$$F_1 = \frac{1}{2} \qquad \text{and} \qquad \forall i \in \{2, \dots, n\}\,, \qquad F_i = \sin\left(25\pi x_i^3\right) \quad \text{with} \quad x_i = \frac{i-1}{n-1}\,.$$

The family of estimators considered is the family of kernel ridge estimators $(\widehat{F}_\lambda)_{\lambda > 0}$ where for every $\lambda > 0$,

$$\widehat{F}_\lambda = K(K + n\lambda I_n)^{-1} Y\,, \quad K = \left(k(x_i, x_j)\right)_{1 \leqslant i,j \leqslant n}\,, \quad \text{and} \quad \forall x, x' \in \mathbb{R},\, k(x, x') = \exp\left(-\alpha |x - x'|\right)$$

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

the Laplace kernel, with $\alpha = 8$. In the experiments, only a finite set $\{\lambda_0, \dots, \lambda_n\}$ of values of λ is considered, chosen such that the degrees of freedom $\mathrm{tr}(K(K + n\lambda_i I_n)^{-1})$ are equal to i for every $i = 0, \dots, n$.

D.2. Procedures

The exact definitions of all procedures considered in the experiments for computing some $\widehat{C}$ or choosing some model $\widehat{m}$ are the following. For the procedures depending on some parameter, its default value is used everywhere except in Tables 2–3. Note that the choice of the default values was made *prior to the simulations*: we can check afterwards on Tables 2–3 that these choices provide reasonably good results (which fortunately happened), so that results using only the default values of the parameters (for instance, Table 1) are meaningful.

Maximal jump ($\widehat{C}_{\mathrm{max\,j.}}$, 'Max. jump', 'max j.' or 'max') In Section 7.1, we define

$$\widehat{C}_{\mathrm{max\,j.}} \in \operatorname*{argmax}_{C \geqslant 0} \left\{ D_{\widehat{m}^{(0)}_{\min}(C^-)} - D_{\widehat{m}^{(0)}_{\min}(C^+)} \right\},$$

that is, the location of the maximal jump of $C \mapsto D_{\widehat{m}^{(0)}_{\min}(C)}$, assuming it is unique. In our experiments, when the argmax contains several values of C, we choose the largest one, that is, the last largest jump; this choice is natural, since it means taking the less complex model among those corresponding to a maximal jump, and it matches the choice made by [Lerasle and Takahashi, 2011].

Note that for change-point detection, [Lebarbier, 2005, Section 4.2] suggests an opposite convention —taking the smallest value of C in the argmax—, arguing from simulation experiments that otherwise too small models are selected. Nevertheless, [Lebarbier, 2005, Section 4.2] also reports that its convention can lead to taking $\widehat{C}_{\mathrm{max\,j.}}$ too small, so a rather complicated method is suggested for choosing some threshold α_{thr} and imposing $\widehat{C}_{\mathrm{max\,j.}} \geqslant \alpha_{\mathrm{thr}}$.

Threshold ($\widehat{C}_{\mathrm{thr.}}$ or 'thr') Eq. (20) in Section 2.5 defines

$$\widehat{C}_{\mathrm{thr.}} := \min \left\{ C \geqslant 0 \,/\, D_{\widehat{m}(C)} \leqslant T_n \right\},$$

which depends on some parameter T_n. The default value of T_n is $n/2$.

Note that Theorem 1 suggests that $T_n = \rho n$ works for any $\rho \in (0,1)$, and previous theoretical results [Arlot and Massart, 2009, Section 3.3] suggest to take $T_n \propto n/\log(n)$ or $n/(\log(n))^2$. Nevertheless, all these theoretical results involve pessimistic constants (as shown by the simulation experiments), so they cannot be used for a fine tuning of T_n. It turns out that $n/2$ does very good in the experiments of Tables 2–3, while other choices lead to much worse performance.

Window ($\widehat{C}_{\mathrm{window}}$ or 'win') Eq. (19) in Section 2.5 defines

$$\widehat{C}_{\mathrm{window}} \in \operatorname*{argmax}_{C \geqslant 0} \left\{ D_{\widehat{m}(C/(1+\eta))} - D_{\widehat{m}(C(1+\eta))} \right\},$$

which depends on some parameter $\eta > 0$. Similarly to $\widehat{C}_{\mathrm{max\,j.}}$, the argmax is usually not reduced to a single point, so a more precise definition must be given for $\widehat{C}_{\mathrm{window}}$. Actually, when $\eta > 0$,

Appendix B.2 shows that $\operatorname{argmax}_{C\geqslant 0}\{D_{\widehat{m}(C/(1+\eta))} - D_{\widehat{m}(C(1+\eta))}\}$ is a finite union of intervals. Denoting by $[\widehat{C}^{(1)}_{\text{window}}, \widehat{C}^{(2)}_{\text{window}})$ the last of these intervals —that is, the one corresponding to the largest values of C— we define

$$\widehat{C}_{\text{window}} = \sqrt{\widehat{C}^{(1)}_{\text{window}} \widehat{C}^{(2)}_{\text{window}}}\,.$$

Of course, other choices could be possible and we do not claim that our (arbitrary) choice is the best one.

In Figures 5 and 9b, the interval represented by the two red vertical lines is $[\widehat{C}^{(1)}_{\text{window}}, \widehat{C}^{(2)}_{\text{window}})$. Note that this interval often looks like $[\widehat{C}_{\text{window}}/(1+\eta), \widehat{C}_{\text{window}}(1+\eta))$ but they can also be quite different.

Taking the limit $\eta \to 0^+$ in the definition of $\widehat{C}_{\text{window}}$, we recover $\widehat{C}_{\max \text{j}.}$. Theorem 1 suggests to take $\eta \propto \eta_n^+ \geqslant \sqrt{\log(n)/n}$, that we consider in our experiments (see Tables 2–3). In our experiments, the default value for η is $n^{-1/2}$, a choice made to get a slightly smaller value than $\sqrt{\log(n)/n} \approx 0.22$ (we recall that $n = 100$ in the least-squares framework).

Slope ($\widehat{C}_{\text{slope}}$ or 'slope') In Algorithm 2, the definition of $\widehat{C}_{\text{slope}}$ is rather vague; it is a bit more precise in Section 7 where the range of models considered in the regression is defined by $\text{pen}_0(m) \in [p_{\min}, p_{\max}]$ for some $p_{\min}, p_{\max}$ to be chosen. In the experiments, since $\text{pen}_0(m) = D_m/n$ in the least-squares framework, we choose $p_{\min} = D_0/n$ and $p_{\max} = 1$ for some parameter $D_0 \in [1, n)$.

In other words, given $D_0 \in [1, n)$, we consider only models of dimension $D_m \geqslant D_0$ and we perform a (standard) linear regression of the empirical risk $n^{-1}\|\widehat{F}_m - F\|^2$ against $-D_m/n$, that is, we solve

$$(\alpha, \beta) \in \operatorname*{argmin}_{(a,b)\in\mathbb{R}^2} \sum_{m\in\mathscr{M}\,/\,D_m\geqslant D_0} \left(a - b\frac{D_m}{n} - \frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 \right)^2 ,$$

and we define $\widehat{C}_{\text{slope}}$ as the resulting slope β. The default value of D_0 is $n/2$.

Capushe ($\widehat{C}_{\text{CAPUSHE}}$, $\widehat{m}_{\text{CAPUSHE}}$ or 'CAP') The procedure called 'CAPUSHE' throughout this paper is the one proposed by [Baudry et al., 2012, Section 4.2] and implemented in the CAPUSHE package for Matlab and R. For completeness, let us recall its definition —which depends on some parameter $pct \in (0,1)$— in the least-squares framework.

- Step 1: If several models have the same dimension D, keep only the one with the smallest empirical risk. This step does not change anything in our experimental setting since there is exactly one model per dimension.
- Step 2: for all $D \in [1, n-2]$, compute by robust linear regression the slope $\widehat{C}_s(D)$ of $n^{-1}\|\widehat{F}_m - F\|^2$ against $-D_m/n$, among models of dimension $D_m \geqslant D$.
- Step 3: for all $D \in [1, n-2]$, compute the corresponding selected model

$$\widehat{m}_D = \widehat{m}\big(2\widehat{C}_s(D)\big) = \operatorname*{argmin}_{m\in\mathscr{M}} \left\{ \frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2 + \frac{2\widehat{C}_s(D) D_m}{n} \right\} .$$

Soumis au Journal de la Société Française de Statistique
File: survey_penmin.tex, compiled with jsfds, version : 2018/06/13
date: January 22, 2019

Then, $\widehat{m}_1, \ldots, \widehat{m}_{n-2}$ is piecewise constant, and some $1 = D_1 < \ldots < D_{I+1} = n-1$ and $m_1, \ldots, m_{I+1} \in \mathscr{M}$ exist such that

$$\forall i \in \{1, \ldots, I\}\,, \forall D \in [D_i, D_{i+1} - 1]\,, \quad \widehat{m}_D = m_i$$

with $m_1 \neq m_2$, …, $m_I \neq m_{I+1}$. The intervals $[D_i, D_{i+1} - 1]$ are called "plateau" (platforms) by [Baudry et al., 2012, Section 4.2] and their size is denoted by $N_i = D_{i+1} - D_i$.

— Step 4: Keep only the platforms of size N_i larger than pct times the total size $\sum_\ell N_\ell = n-2$, and among these, define $\widehat{\imath}$ the last platform, that is,

$$\widehat{\imath} = \max\left\{ i \in \{1, \ldots, I\} \,/\, N_i > pct \times (n-2) \right\}$$

and select

$$\widehat{m}_{\text{CAPUSHE}} = m_{\widehat{\imath}}.$$

Note that at step 4, it can happen that no platform is large enough. In such cases, we consider the last platform among the ones of largest size, that is,

$$\widehat{\imath} = \max\left\{ i \in \{1, \ldots, I\} \,/\, N_i = \max_j N_j \right\}.$$

We always take $pct = 0.15$ in our experiments, that is, the default value proposed by [Baudry et al., 2012].

Note that [Baudry et al., 2012] only provides a model-selection procedure $\widehat{m}_{\text{CAPUSHE}}$, and not a value $\widehat{C}$ of the constant in front of the penalty. In order to help understanding better $\widehat{m}_{\text{CAPUSHE}}$, we also report in our experiments the distribution of $\widehat{C}_{\text{CAPUSHE}}$ that we define as some median of

$$\left\{ \widehat{C}_s(D_{\widehat{\imath}}), \ldots, \widehat{C}_s(D_{\widehat{\imath}+1} - 1) \right\}.$$

This choice is arbitrary among many others that all lead to having $\widehat{m}_{\text{CAPUSHE}} = \widehat{m}(2\widehat{C}_{\text{CAPUSHE}})$.

Median ('med') As defined in the caption of Figure 6, 'median' refers to taking $\widehat{C}$ as the median of

$$\left\{ \widehat{C}_{\text{max j.}}, \widehat{C}_{\text{thr.}}, \widehat{C}_{\text{window}}, \widehat{C}_{\text{slope}}, \widehat{C}_{\text{CAPUSHE}} \right\}$$

(with their default parameter values for $\widehat{C}_{\text{thr.}}$, $\widehat{C}_{\text{window}}$, and $\widehat{C}_{\text{slope}}$), and $\widehat{m} = \widehat{m}(2\widehat{C})$.

Remark that the set of procedures considered is arbitrary, and other choices could be made. The idea of considering some median of several values of $\widehat{C}$ could also be used when there is some uncertainty about the parameter of some procedure (say, T_n for $\widehat{C}_{\text{thr.}}$), by considering the median of the set of values obtained on a grid of values of the parameter.

Residuals ($\widehat{\sigma}^2_{m_0}$, 'Residuals on m_0' or 'resid') The residual-based variance estimator $\widehat{\sigma}^2_{m_0}$ is defined by Eq. (70) in Section 6.1:

$$\widehat{\sigma}^2_{m_0} := \frac{1}{n - D_{m_0}} \left\| Y - \widehat{F}_{m_0} \right\|^2,$$

for some model m_0. In the experiments, there is one model per dimension so m_0 is given by the value of its dimension D_{m_0}, and the default choice is $D_{m_0} = n/2$. Since $n = 100$, the default

choice is $D_{m_0} = 50$ which is even, so the definition of S_m^{hard} —in which models of odd dimension are good and models of even dimension are very poor— is made on purpose.

Note that in Tables 2–3, the line "$D_{m_0} = n/\log(n)$" means "$D_{m_0} = 21$" (hence, for the 'hard' setting, it is a reasonably good model).

Consensus ('cons') As defined in the caption of Figure 7, a majority vote is performed among

$$\left\{\widehat{m}(2\widehat{C}_{\text{max j.}}), \widehat{m}(2\widehat{C}_{\text{thr.}}), \widehat{m}(2\widehat{C}_{\text{window}}), \widehat{m}(2\widehat{C}_{\text{slope}}), \widehat{m}_{\text{CAPUSHE}}\right\}$$

with their default parameters values. If no majority emerges (that is, if we do not have at least three of these procedures that agree), the default choice is $\widehat{m}(2\widehat{C}_{\text{window}})$. Remark that Table 1 shows that an agreement occurs for more than 96% of the samples in the 'easy' setting, and for more than 89% of the samples in the 'hard' setting.

Consensus when no reject ('no rej') This actually refers to the same procedure as 'consensus', but showing results (a boxplot or an estimation of the expectation of the loss ratio) only for the samples for which a majority emerged. Again, Table 1 shows that this only removes a small fraction of the $N = 10^4$ independent samples generated in our experiments.

Mallows' C_p When the variance σ^2 is known, a natural model-selection procedure for the 'least-squares' framework is Mallows' C_p [Mallows, 1973], that is, selecting

$$\widehat{m} = \widehat{m}(2\sigma^2) = \operatorname*{argmin}_{m\in\mathscr{M}} \left\{ \frac{1}{n} \left\| \widehat{F}_m - Y \right\|^2 + \frac{2\sigma^2 D_m}{n} \right\}.$$

Its performances are shown in Tables 2–3 for comparison.

Mallows' C_p is also considered for illustrating the overpenalization phenomenon in Figure 8 in Section 8.4. On the graph of Figure 8, what is plotted is the estimated value (from $N = 10^4$ independent samples) of the expected risk ratio

$$\mathbb{E}\left[\frac{\left\| \widehat{F}_{\widehat{m}(2C\sigma^2)} - F \right\|^2}{\inf_{m\in\mathscr{M}} \left\| \widehat{F}_m - F \right\|^2} \right]$$

when using Mallows' C_p penalty multiplied by some factor $C \in [0,4]$, as a function of C. For plotting the graph of Figure 8, a linear grid of values of C with stepsize $1/100$ is considered. The optimal performance is obtained for $C = 1.12$ in the 'easy' and 'hard' settings, and it is also included in Tables 2–3.

D.3. Additional remarks

Repeated experiments show results obtained from $N = 10^4$ independent samples.

Illustrations made on a single sample in the least-squares framework are showed in Figures 2, 4, 5, 9, 10. The samples considered have been chosen manually in order to illustrate either typical or rare (but still possible) configurations. The graphs of Figure 2, Figure 4, and Figure 10a are

made on the same sample (they correspond to a 'typical' situation). The graph of Figure 5 is made on a second sample (corresponding to a 'rare' situation). The two graphs of Figure 9 are made on a third sample (also corresponding to a 'rare' situation, similar to the one shown in Figure 5).

Figure 3 is taken from [Arlot and Bach, 2011, top right graph of Figure 2]. It is made from a single sample generated as in the 'kernel ridge' framework (see Appendix D.1), with a sample size $n = 200$.