ESAIM: PROCEEDINGS AND SURVEYS, October 2015, Vol. 51, p. 232-245
A. Garivier et al, Editors

ESTIMATOR SELECTION

M. Lerasle[1]

Abstract. The paper presents recent developments of the theory of estimator selection. We introduce, in the density estimation framework, the main methods used by the participants of the session "Variable and estimator selection" in the Journées MAS 2014. The purpose of the selection is always to prove oracle inequalities, that is, to compare the selected estimator with the best estimator in the original collection via some risk function. The first part of the paper deals with the selection by minimization of a penalized empirical loss and the second presents the methods based on robust tests.

Résumé. L'article présente quelques développements récents de la théorie de la sélection d'estimateurs. Nous introduisons, dans le cadre élémentaire de l'estimation de la densité, les principales méthodes apparues dans les exposés de la session "Sélections de variables, sélection d'estimateurs" des Journées MAS 2014. L'objectif de la sélection est toujours l'obtention d'inégalités oracle comparant le risque de l'estimateur choisi au plus petit des risques des estimateurs de la collection initiale. Nous discuterons dans une première partie les méthodes par minimisation d'un critère pénalisé et dans une seconde celles utilisant les tests robustes.

Introduction

The paper introduces some recent developments on the theory of estimator selection, from the oracle point of view. The subject is quite general and this presentation focuses on the methods related to the talks given in the Journées MAS 2014 in the session "estimator selection". By oracle point of view, we mean that we are interested in the following problem : given a collection $(\widehat{f}_\lambda)_{\lambda\in\Lambda}$ of estimators and a risk function $R : \Lambda \to \mathbb{R}_+^*$, the aim is to choose, using only the data, an estimator $\widehat{\lambda} \in \Lambda$ such that $R(\widehat{\lambda})$ is as close as possible to $\inf_{\lambda\in\Lambda} R(\lambda)$. We focus on the density estimation framework where one wants to estimate the unknown density $f^\star$ with respect to a known measure μ of a distribution P, based on the observation of an i.i.d. sample $X_1, \dots, X_n$ of real valued random variables with common distribution P. Among popular risk functions in density estimation, one can mention the risks L^1, L^2 or least-squares, L^p, Küllback, in total variation or in Hellinger distance. In the first part of this talk, R denotes the least-squares risk, that is $f^\star$ is supposed to belong to the space of square integrable functions L^2 and

$$\forall \lambda \in \Lambda, \qquad R(\lambda) = \mathbb{E}\left[\left\| f^\star - \widehat{f}_\lambda \right\|^2 \right] .$$

Estimator selection covers several classical problems, let us present here some well known examples that are discussed in the following. In the sequel, X denotes a copy of X_1, independent of $X_1, \dots, X_n$.

[1] Univ. Nice Sophia Antipolis LJAD CNRS UMR 7351
06100 Nice France

© EDP Sciences, SMAI 2015

Article published online by EDP Sciences and available at http://www.esaim-proc.org or http://dx.doi.org/10.1051/proc/201551013

Example 1 (Model selection). The unknown density $f^\star$ is equal to

$$f^\star = \arg\min_{t\in L^2} \left\{ \|t\|^2 - 2\mathbb{E}\left[t(X)\right] \right\} \ .$$

Given a collection of linear subspaces $(S_\lambda)_{\lambda\in\Lambda}$ (the models) of L^2, the empirical risk minimizers of $f^\star$ are defined by

$$\widehat{f}_\lambda = \arg\min_{t\in S_\lambda} \left\{ \|t\|^2 - \frac{2}{n}\sum_{i=1}^n t(X_i) \right\} \ .$$

The estimator $\widehat{f}_\lambda$ is also called the projection estimator of $f^\star$ on the linear space S_λ. The problem of model selection is the problem of selecting a space S_λ or, equivalently, an estimator $\widehat{f}_\lambda$ in order to minimize the risk R.

Example 2 (Selection of linear estimators). Let $k : \mathbb{R}^2 \to \mathbb{R}$ denote a function such that $k(x,y) = k(y,x)$ and

$$\sup_{(x,y)} |k(x,y)| \vee \sup_x \int k(x,y)^2 d\mu(y) < \infty \ .$$

The function k is called a kernel. The associated *linear estimator* is defined, for any $x \in \mathbb{R}$, by

$$\widehat{f}_k(x) = \frac{1}{n}\sum_{i=1}^n k(X_i, x) \ .$$

In the linear estimator selection problem, or kernel selection problem, we want to select $\widehat{k}$ in a collection $\mathcal{K}$ in order to minimize the least-squares risk. It is an estimator selection problem.
As a first example of kernel estimator, consider a finite dimensional linear space $S \subset L^2$ and an orthonormal basis $(\varphi_i)_{i=1,\dots,p}$ of S. A projection kernel onto S is defined by $k_S(x,y) = \sum_{i=1}^p \varphi_i(x)\varphi_i(y)$. The associated kernel estimator $\widehat{f}_k = \sum_{j=1}^p (\frac{1}{n}\sum_{i=1}^n \varphi_j(X_i))\varphi_j$ is the projection estimator onto S. Hence, model selection can be seen as a kernel selection problem.
To present a second classical example, let $K : \mathbb{R} \to \mathbb{R}$ denote a bounded function such that $K(x) = K(-x)$ and $\|K\|_1 = 1$. Let $h > 0$ denote a real number called a bandwidth. The approximation kernel (or Parzen's kernel) associated to K and h is defined by $k_{K,h}(x,y) = \frac{1}{h}K\left(\frac{x-y}{h}\right)$. The associated estimator $\widehat{f}_{k_{K,h}}(x) = \frac{1}{nh}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)$ is sometimes called simply kernel estimator. The kernel selection problem is then to choose the function K and/or the bandwidth $h > 0$ to minimize the L^2-risk.

Example 3 (Aggregation). Let $(\varphi_1, \dots, \varphi_p)$ denote a fixed collection of functions, for example an orthonormal system in L^2 or some estimators built with an independent sample. For any $\lambda \in \Lambda = \mathbb{R}^p$, define

$$\widehat{f}_\lambda = \sum_{i=1}^p \lambda_i \varphi_i \ .$$

Any estimator $\widehat{f}_\lambda$ is called an aggregate and the problem of choosing the aggregation weights λ minimizing the L^2-risk is another estimator selection problem. Notice that $\widehat{f}_\lambda$ is a fixed function when λ is, it is still denoted $\widehat{f}_\lambda$ though to keep uniform notations. A first important example of aggregated estimator is given by thresholded estimators. Suppose that $(\varphi_i)_{i=1}^p$ is an orthonormal system in L^2 and let $\tau \geq 0$ be a threshold, define $\widehat{\lambda}_i = \widehat{\beta}_i \mathbf{1}_{\left|\widehat{\beta}_i\right| > \tau}$, with $\widehat{\beta}_i = \frac{1}{n}\sum_{j=1}^n \varphi_i(X_j)$. The thresholded estimator is then defined by

$$\widehat{f}_\tau^{\text{thresh}} = \sum_{i=1}^p (\widehat{\beta}_i \mathbf{1}_{\left|\widehat{\beta}_i\right| > \tau})\varphi_i \ .$$

These estimators are mostly used when $(\varphi_1, \dots, \varphi_p)$ is a wavelet basis, see for example [DJKP96]. Remark that $\tau = 0$ gives the projection estimator onto the linear subspace of L^2 spanned by the family $(\varphi_1, \dots, \varphi_p)$. Another popular example is given by LASSO estimators [Tib96]. Let $c > 0$ and

$$\widehat{\lambda}_c^{\text{lasso}} = \arg\min_{\lambda \in \Lambda} \left\{ \left\| \widehat{f}_\lambda \right\|^2 - \frac{2}{n} \sum_{i=1}^n \widehat{f}_\lambda(X_i) + c \sum_{j=1}^p |\lambda_j| \right\} \ .$$

The Lasso estimator with regularization parameter c is $\widehat{f}_{\widehat{\lambda}_c^{\text{lasso}}}$. These estimators are particularly interesting since they can be computed in practice and have nice theoretical properties in large dimension such as adaptation to sparsity as long as c is sufficiently large. Notice that the choice of the optimal threshold τ or constant c are other estimator selection problems.

The paper is organized as follows. Section 1 deals with minimization of an empirical contrast. Deterministic penalties, in particular the Lasso, are presented in Section 1.1. Resampling methods and cross-validation are discussed in Section 1.2, Section 1.3 presents the minimal penalty phenomenon and some practical applications to the slope heuristic. The second part of this paper, Section 2 briefly introduces estimator selection using tests.

1. Selection by minimization of a penalized empirical contrast

The purpose is to select among $(\widehat{f}_\lambda)_{\lambda \in \Lambda}$ an estimate $\widehat{f}_{\widehat{\lambda}}$ from the data such that $R(\widehat{\lambda}) = \mathbb{E}\left[\left\| \widehat{f}_{\widehat{\lambda}} - f^\star \right\|^2 \right]$ is as close as possible to $\inf_{\lambda \in \Lambda} R(\lambda)$. More precisely our aim is to choose $\widehat{\lambda}$ such that

$$R(\widehat{\lambda}) \leq C_n \inf_{\lambda \in \Lambda} R(\lambda) + r_n \ ,$$

where $C_n \geq 1$ and $r_n > 0$. In this case, $\widehat{f}_{\widehat{\lambda}}$ is said to satisfy an *oracle inequality*, see [DJ94], as long as r_n is asymptotically comparable to $\inf_{\lambda \in \Lambda} R(\lambda)$ and C_n does not explode, that is C_n is smaller than a power of $\log n$. This means that the selected estimate does as well as the best estimate in the family up to some multiplicative power of $\log n$. The case where C_n is close to 1 is the best case one can expect. If $C_n \to_{n \to \infty} 1$, the corresponding oracle inequality is called *asymptotically optimal*. When $C_n = 1$, the oracle inequality is called sharp, see for example [Cat04] or [DT08].

To do so, we study minimizers of *penalized least-squares criteria*. Let P_n denote the empirical measure defined for any real valued function t by

$$P_n(t) := \frac{1}{n} \sum_{i=1}^n t(X_i) \ .$$

For any $t \in L^1(P)$, let also

$$P(t) := \mathbb{E}\left[t(X) \right] = \int_{\mathbb{X}} t(x) f^\star(x) d\mu(x) \ .$$

The *least-squares contrast* is given for any $t \in L^2$ by

$$\gamma(t) := \|t\|^2 - 2t \ .$$

Then for any given function $\text{pen} : \Lambda \to \mathbb{R}$, the *least-squares penalized criterion* is defined by

$$\mathcal{C}_{\text{pen}}(\lambda) := P_n \gamma(\widehat{f}_\lambda) + \text{pen}(\lambda) \ . \tag{1}$$

Then the selected $\widehat{\lambda} \in \Lambda$ is given by any minimizer $\widehat{\lambda}$ of $\mathcal{C}_{\text{pen}}(\lambda)$, that is

$$\widehat{\lambda} \in \arg\min_{\lambda\in\Lambda} \{ \mathcal{C}_{\mathrm{pen}}(\lambda) \} \quad . \tag{2}$$

1.1. **Selection with deterministic penalties**

Consider the aggregation setting where $\Lambda = \mathbb{R}^p$ and one wants to select $\widehat{\lambda} \in \Lambda$ to guarantee a small risk of $\widehat{f}_{\widehat{\lambda}} = \sum_{i=1}^p \widehat{\lambda}_i \varphi_i$. Recall that the functions $(\varphi_i)_{i=1}^p$ are assumed to be fixed in this setting. The penalty $\mathrm{pen}(\lambda)$ is usually a deterministic quantity proportional to the norm $|\lambda|$. A classical choice is the ℓ_0-norm $|\lambda|_0$ that is the number of non-zero coordinates of λ. These penalties yield nice adaptive results and ensure sparsity of $\widehat{\lambda}$. However, it is usually hard to compute the estimator $\widehat{\lambda}$ when p is large. An interesting particular case though where this estimator can be computed is when the dictionary $(\varphi_i)_{i=1}^p$ is an orthonormal basis. Actually, the corresponding estimator is simply the hard thresholded estimator (see for example [BBM99])

$$\widehat{f}_{\widehat{\lambda}} = \sum_{i=1}^p (P_n\varphi_i \mathbf{1}_{|P_n(\varphi_i)|>\tau})\varphi_i \quad ,$$

where the threshold τ is the constant in front of the penalty. Laure Sansonnet presented her paper [San14] that uses thresholded estimators in a Poissonian interactions model in the Session Estimator selection in the Journées MAS 2014.

In general, the ℓ_0 criterion has to be relaxed and ℓ_q-norms, with $q = 1, 2$, have also been studied yielding respectively to the LASSO [Tib96] and ridge estimators. Ridge estimators, as named by Hoerl, are very easy to compute since a closed form is available, but the estimator $\widehat{\lambda}$ is usually not sparse which yields to bad performances when the dimension p is large. LASSO estimators offer a nice alternative, the criterion can be minimized in practice and efficient algorithm have been developed. Moreover, the selected estimator behaves well in large dimension. These reasons made LASSO estimators very popular over the last few years, even if theoretical results on LASSO estimators usually rely on strong assumptions on the dictionary $(\varphi_i)_{i=1}^p$ such as the RIP condition that may not be easy to check in practice. We refer to the book [BvdG11] and the references therein for more details on LASSO estimators and related topics. Mixed strategies, such as the Elastic net estimators [ZH05] that takes a linear combination of ℓ_1 and ℓ_2-norms, have also been studied. In this section, we'll shortly present a study of the LASSO estimator in density estimation. We follow the presentation of [BTWB10], who study these estimators for general $f^\star$ and apply their results to the case of mixtures of Gaussian densities. The empirical risk

$$P_n\gamma(\widehat{f}_\lambda) = \left\|\widehat{f}_\lambda\right\|^2 - 2\sum_{i=1}^p \lambda_i P_n\varphi_i = \left\|\widehat{f}_\lambda - f^\star\right\|^2 - \|f^\star\|^2 - 2\sum_{i=1}^p \lambda_i \left[(P_n - P)\varphi_i\right] \quad .$$

Hence, any minimizer $\widehat{\lambda}$ defined by (2) satisfies, for any $\lambda \in \mathbb{R}^p$,

$$\left\|\widehat{f}_{\widehat{\lambda}} - f^\star\right\|^2 \le \left\|\widehat{f}_\lambda - f^\star\right\|^2 + \mathrm{pen}(\lambda) - \mathrm{pen}(\widehat{\lambda}) + 2\sum_{i=1}^p (\widehat{\lambda}_i - \lambda_i)\left[(P_n - P)\varphi_i\right] \quad .$$

Suppose that the functions φ_i are uniformly bounded by M. Hoeffding's inequality ensures that

$$\forall x > 0, \qquad \mathbb{P}\left(\forall i = 1, \dots, p, \quad |(P_n - P)\varphi_i| \le M\sqrt{\frac{2(\log(p) + x)}{n}}\right) \ge 1 - 2e^{-x} \quad .$$

On the event, $\Omega_{good} = \left\{ \forall i = 1, \dots, p, \quad |(P_n - P)\varphi_i| \leq M\sqrt{\frac{2(\log(p)+x)}{n}} \right\}$, for any $\lambda \in \mathbb{R}^p$,

$$\left\| \widehat{f}_{\widehat{\lambda}} - f^\star \right\|^2 \leq \left\| \widehat{f}_\lambda - f^\star \right\|^2 + \mathrm{pen}(\lambda) - \mathrm{pen}(\widehat{\lambda}) + 2\sum_{i=1}^{p} \left| \widehat{\lambda}_i - \lambda_i \right| M\sqrt{\frac{2(\log(p)+x)}{n}} \ .$$

Then various assumptions allow to derive from this inequality some oracle properties of $\widehat{f}_{\widehat{\lambda}}$, they usually involve some connection between the ℓ_2-norm of λ and the L^2-norm of $\widehat{f}_\lambda$. Assume for example that, for any λ with support $\mathrm{supp}(\lambda)$ of cardinal smaller than s,

$$\|\lambda\|^2 \leq A_s \left\| \widehat{f}_\lambda \right\|^2 \ .$$

Suppose furthermore that, for any $s \leq p$ and any λ such that $|\lambda|_0 \leq s$, there exists a constant $C_s \geq 1$ such that $\left\| \widehat{f}_\lambda \right\|^2 \leq C \left\| \widehat{f}_{\lambda'} \right\|^2$ for any λ' such that $\lambda'_i = \lambda_i$ for any i such that $\lambda_i \neq 0$. Lasso estimators are defined by (2) with $\mathrm{pen}(\lambda) = c\sum_{i=1}^p |\lambda_i|$, for $c = 2M\sqrt{\frac{2(\log(p)+x)}{n}}$. Using successively Cauchy-Schwarz inequality, our assumptions on the dictionary and the classical inequality $2ab \leq \theta a^2 + b^2/\theta$, we get that, for any $\theta > 0$, and any λ with $|\lambda|_0 \leq s$, they satisfy

$$\begin{aligned}
\left\| \widehat{f}_{\widehat{\lambda}} - f^\star \right\|^2 &\leq \left\| \widehat{f}_\lambda - f^\star \right\|^2 + 4M\sqrt{\frac{2(\log(p)+x)}{n}} \sum_{i \in \mathrm{supp}(\lambda)} \left| \widehat{\lambda}_i - \lambda_i \right| \\
&\leq \left\| \widehat{f}_\lambda - f^\star \right\|^2 + 4M\sqrt{s\frac{2(\log(p)+x)}{n}} \sqrt{\sum_{i \in \mathrm{supp}(\lambda)} \left| \widehat{\lambda}_i - \lambda_i \right|^2} \\
&\leq \left\| \widehat{f}_\lambda - f^\star \right\|^2 + 4M\sqrt{C_s A_s s\frac{2(\log(p)+x)}{n}} \left\| f^\star_\lambda - f^\star_{\widehat{\lambda}} \right\| \\
&\leq (1+\theta) \left\| \widehat{f}_\lambda - f^\star \right\|^2 + \theta \left\| f^\star - \widehat{f}_{\widehat{\lambda}} \right\|^2 + \frac{16 C_s A_s M^2}{\theta} \frac{s(\log(p)+x)}{n} \ .
\end{aligned}$$

i.e., for any $x > 0$, denoting by $\Lambda_s = \{ \lambda \in \Lambda, \ \text{s.t.} \ \ |\lambda|_0 \leq s \}$ for any $s \leq p$,

$$\mathbb{P}\left(\forall \theta > 0, \ \frac{1-\theta}{1+\theta} \left\| \widehat{f}_{\widehat{\lambda}} - f^\star \right\|^2 \leq \inf_{s \leq p} \left\{ \inf_{\lambda \in \Lambda_s} \left\| \widehat{f}_\lambda - f^\star \right\|^2 + \frac{16 C_s A_s M^2}{\theta} \frac{s(\log(p)+x)}{n} \right\} \right) \geq 1 - 2e^{-x} \ .$$

This equation shows several nice properties of Lasso's estimators; in particular, even if the dimension p is larger than n, the remainder term is controlled when $s \log p < n$. Therefore, the risk of $\widehat{f}_{\widehat{\lambda}}$ is well controlled if $f^\star$ is close to some subset of $(\widehat{f}_\lambda)_{\lambda \in \Lambda_s}$ with a small s.

1.2. **Selection with resampling methods**

We would like to minimize an ideal criterion $\mathcal{C}_{\mathrm{id}}(\lambda) := P\gamma(\widehat{f}_\lambda) = \left\| f^\star - \widehat{f}_\lambda \right\|^2 - \|f^\star\|^2$ and we replaced it by a penalized one $\mathcal{C}_{\mathrm{pen}}(\lambda) = P_n\gamma(\widehat{f}_\lambda) + \mathrm{pen}(\lambda)$, therefore, an ideal penalty (as called by [Arl09]) is given by

$$\mathrm{pen}_{\mathrm{id}}(\lambda) := \mathcal{C}_{\mathrm{id}}(\lambda) - P_n\gamma(\widehat{f}_\lambda) = (P - P_n)\gamma(\widehat{f}_\lambda) = 2(P_n - P)\widehat{f}_\lambda \ .$$

In this section, we shall focus on the model selection framework where one wants to select among a finite collection of projection estimators $\widehat{f}_\lambda = \sum_{i \in \mathcal{I}_\lambda} (P_n\varphi_i)\varphi_i$ one with a small L^2-risk. The ideal penalty can be

written

$$\mathrm{pen}_{\mathrm{id}}(\lambda) = 2\sum_{i\in\mathcal{I}_\lambda}(P_n\varphi_i)(P_n - P)\varphi_i \ .$$

1.2.1. *Resampling penalties in the model selection framework*

Resampling penalties, originally introduced by Efron [Efr83], have recently been studied from a non-asymptotic point of view in a series of works including [Fro04, Arl09, Ler12].
The idea of Efron [Efr83] is to estimate the ideal penalty using the resampling heuristics. A resampling scheme is a vector $W = (W_1, \dots, W_n)$ independent of $X_1, \dots, X_n$ of random variables such that $\sum_{i=1}^n W_i = n$. The associated resampling estimator of $\mathrm{pen}_{\mathrm{id}}(\lambda)$ is defined by

$$\mathrm{pen}_W(\lambda) = \mathbb{E}_W\left[(P_n - P_n^W)\gamma(\widehat{f}_\lambda^W)\right] = 2\mathbb{E}_W\left[(P_n^W - P_n)\widehat{f}_\lambda^W\right] = 2\mathbb{E}_W\left[\sum_{i\in\mathcal{I}_\lambda}(P_n^W\varphi_i)(P_n^W - P_n)\varphi_i\right] \ ,$$

where the resampling empirical process P_n^W is defined for any function g by

$$P_n^W g = \frac{1}{n}\sum_{i=1}^n W_i g(X_i) \ ,$$

the resampling estimator $\widehat{f}_\lambda^W = \sum_{i\in\mathcal{I}_\lambda}(P_n^W\varphi_i)\varphi_i$ and the expectation $\mathbb{E}_W[.]$ is the expectation conditionally on the sample $X_1, \dots, X_n$.
The most classical weights W are exchangeable, which means that, for any permutation τ of $\{1, \dots, n\}$,

$$W \overset{\mathrm{dist}}{=} (W_{\tau(1)}, \dots, W_{\tau(n)}) \ .$$

For example, Efron's weights, where W has multinomial distribution $\mathcal{M}(1/n, \dots, 1/n)$ are exchangeable. It is proved in [Ler12] that all resampling penalties built with exchangeable weights are proportional in this problem. In particular, they are all proportional to the *leave-one-out* penalties that will be presented in the next section. Therefore, the results for these penalties can be derived from those on cross-validation methods.

1.2.2. *Cross-validation*

Another important method for estimator selection is cross-validation. In order to evaluate the risk of a procedure of estimation, the validation principle proposes to divide the sample into a training sample $X_T = (X_i)_{i\in T}$, with $T \subset \{1, \dots, n\}$ and a validation one $X_{T^c} = (X_i)_{i\in T^c}$, with $T^c = \{1, \dots, n\}\setminus T$, both non-empty. The training sample is used to build the estimators $\widehat{f}_\lambda^T = \sum_{i\in\mathcal{I}_\lambda}(P_T\varphi_i)\varphi_i$, where $P_T g = \frac{1}{|T|}\sum_{i\in T} g(X_i)$ while the validation sample is used to estimate the risk of $\widehat{f}_\lambda^T$, $P\gamma(\widehat{f}_\lambda^T)$ by $P_{T^c}\gamma(\widehat{f}_\lambda^T)$. This yields the validation criterion, or hold-out criterion

$$\mathcal{C}_{\mathrm{HO},T}(\lambda) = P_{T^c}\gamma(\widehat{f}_\lambda^T) \ .$$

and the estimator $\widehat{f}_{\widehat{\lambda}}$, where $\widehat{\lambda}$ is a minimizer of $\mathcal{C}_{\mathrm{HO},T}(\lambda)$ is called hold-out estimator of $f^\star$. In practice, hold-out estimators are highly dependent on the choice of T and are therefore unstable. In order to reduce this variability, cross-validation criteria propose to take several training sets $\mathcal{E}$ and to minimize the criterion obtained by taking the empirical mean of the hold-out criteria :

$$\mathcal{C}_{\mathcal{E}}(\lambda) = \frac{1}{|\mathcal{E}|}\sum_{T\in\mathcal{E}} P_{T^c}\gamma(\widehat{f}_\lambda^T) \ .$$

For example, the leave-p-out criterion, $\mathcal{C}_{\mathrm{lpo}}(\lambda)$ is obtained when $\mathcal{E}$ is the collection of all subsets of $\{1, \dots, n\}$ with cardinal $n - p$. The V-fold criteria $\mathcal{C}_V(\lambda)$ are obtained by taking a partition $B_1, \dots, B_V$ of $\{1, \dots, n\}$

such that all B_i have cardinal n/V and to choose for $\mathcal{E}$ the collection of subsets $(B_k^c)_{k=1,\dots,V}$. Closed forms exist for leave-p-out criteria in this framework, see [Cel12] and the references therein, but, in general, V-fold cross-validation estimates are much faster to compute, especially when $V = 2, 5$ or 10 which are the usual choices in practice.
It is easy to apply the cross-validation principle to estimate the ideal penalty by

$$\mathrm{pen}_{\mathcal{E}}(\lambda) = 2\frac{1}{|\mathcal{E}|}\sum_{T\in\mathcal{E}}\sum_{i\in\mathcal{I}_\lambda}(P_T\varphi_i)(P_T - P_{T^c})\varphi_i \ .$$

We define $\mathrm{pen}_{\mathrm{lpo}}$ and pen_V using this principle and the collections of training sets respectively used in the definitions of $\mathcal{C}_{\mathrm{lpo}}$ and $\mathcal{C}_V$. Now, if all $T \in \mathcal{E}$ have the same cardinality $n - n/V$, as this is the case for the collections defining $\mathcal{C}_{\mathrm{lpo}}$, when $p = n/V$ and the one defining $\mathcal{C}_V$, then

$$\forall i \in \mathcal{I}_\lambda, \qquad (P_T - P_{T^c})\varphi_i = V(P_T - P_n)\varphi_i \ .$$

and, denoting by $W_i = \frac{V}{V-1}\mathbf{1}_{i\in T}$, $P_T g = P_n^W g$ for any g. Therefore,

$$\mathrm{pen}_{\mathcal{E}}(\lambda) = V\,\mathrm{pen}_W(\lambda) \ ,$$

where $W = (W_1, \dots, W_n)$, with $W_i = \frac{V}{V-1}\mathbf{1}_{i\in T}$ and T uniformly chosen among the subsets in $\mathcal{E}$. Moreover, the weights W associated to the leave-p-out criteria are exchangeable, hence $\mathrm{pen}_{\mathrm{lpo}}$ belongs to the family of penalties pen_W built with an exchangeable resampling scheme. The V-fold weights are exchangeable iff $V = n$, hence, pen_V is proportional to $\mathrm{pen}_{\mathrm{lpo}}$ iff $V = n$. Hence, any resampling penalty pen_W built with exchangeable weights W and any leave-p-out penalties belong to the set $(C\,\mathrm{pen}_V)_{C>0,V=n}$. In addition, the following results holds, see Lemma 1 in [AL14].

Lemma 1. *For any $\mathcal{E}$ such that $\forall T \in \mathcal{E}$, $|T| = n - n/V$ and $\frac{1}{|\mathcal{E}|}\sum_{T\in\mathcal{E}} P_T = P_n$,*

$$\mathcal{C}_{\mathcal{E}}(\lambda) = P_n\gamma(\widehat{f}_\lambda) + \frac{V-\frac{1}{2}}{V}\,\mathrm{pen}_{\mathcal{E}}(\lambda) = P_n\gamma(\widehat{f}_\lambda) + \left(V - \frac{1}{2}\right)\mathrm{pen}_V(\lambda) \ .$$

Since the collections $\mathcal{E}$ defining $\mathrm{pen}_{\mathrm{lpo}}$ and pen_V satisfy the assumptions of the lemma, this result implies, together with our previous discussion, that we can study all cross-validation criteria, all V-fold penalized criteria and all criteria penalized by a resampling penalty with exchangeable weights by studying the following penalized criteria.

$$\mathcal{C}_{V,C}(\lambda) = P_n\gamma(\widehat{f}_\lambda) + C(V-1)\,\mathrm{pen}_V(\lambda) \ , \tag{3}$$

for any constant C, the factor $V - 1$ being here for normalization, see [AL14].

1.2.3. *An oracle inequality for V-fold penalized criteria*

The following result shows an oracle inequality that can be derived for minimizers $\widehat{\lambda}$ of the criteria (3). The proof can be found in [AL14].

Theorem 2. *Assume that $(S_\lambda)_{\lambda\in\Lambda}$ is a finite collection of finite dimensional linear spaces such that, for any $\lambda \in \Lambda$, $\sup_{t\in S_\lambda/\{0\}} \frac{\|t\|_\infty}{\|t\|} \le \sqrt{n}$ and either all the projections $f_\lambda^\star$ of $f^\star$ onto S_λ are uniformly bounded by a constant $A \ge \|f^\star\|_\infty$ or the collection $(S_\lambda)_{\lambda\in\Lambda}$ is nested, that is $S_\lambda \cup S_{\lambda'} \subset \{S_\lambda, S_{\lambda'}\}$ and $A = \|f^\star\|_\infty$. Then, for any $1/2 < C \le 2$, there exists an absolute constant κ such that, for any $x \ge 1$, with probability larger than $1 - e^{-x}$, for any $\theta \in (0,1)$, the minimizer $\widehat{\lambda}$ of (3) satisfies*

$$\left(\frac{1-\delta_-}{1+\delta_+} - \theta\right)\left\|\widehat{f}_{\widehat{\lambda}} - f^\star\right\|^2 \le \inf_{\lambda\in\Lambda}\left\|\widehat{f}_\lambda - f^\star\right\|^2 + \kappa\frac{A(|\Lambda|^2 + x^2)}{\theta^3 n} \ ,$$

where $\delta = 2(C-1)$, $\delta_+ = \delta \vee 0$, $\delta_- = (-\delta) \vee 0$ *and* $|\Lambda|$ *is the cardinality of* Λ. *In particular, integrating this inequality,*

$$\left(\frac{1-\delta_-}{1+\delta_+} - \theta \right) R(\widehat{\lambda}) \leq \inf_{\lambda \in \Lambda} R(\lambda) + \kappa \frac{A|\Lambda|^2}{\theta^3 n}$$

Unbiased criteria, such as corrected V-fold cross-validation, are obtained for $C = 1$, i.e. $\delta = 0$. Theorem 2 shows they are asymptotically optimal. More generally, δ measures the bias of the V-fold penalization as an estimator of the ideal penalty. The result suggests no asymptotically optimal oracle inequality can be obtained when δ does not converge to 0. While this result is not proved in [AL14], the ideas of [Arl08] can easily be adapted to show that it is actually the case in general. In particular, as in [Arl08] or [Cel12], we get that classical V-fold cross-validation, with a fixed V, or leave-p-out selection, with a fixed p, are asymptotically suboptimal.

1.3. **The minimal penalty phenomenon**

The minimal penalty phenomenon was discovered by Birgé and Massart [BM07] in the model selection framework. We were lucky to have a very interesting overview of this topic and its recent developments by P. Massart during the Journées MAS 2014.

1.3.1. *The minimal penalty phenomenon in kernel selection*

The material of this section is borrowed from [ALM14, LMRB14]. We will only consider the projection kernels and Parzen kernels. Recall that a projection kernel is defined by

$$k_\lambda(x,y) = \sum_{i \in \mathcal{I}_\lambda} \varphi_i(x)\varphi_i(y) \ ,$$

where $(\varphi_i)_{i \in \mathcal{I}_\lambda}$ is an orthonormal family. To simplify the exposition, we consider projection kernels onto the regular histograms in $[0,1)$. More precisely, $\Lambda = \{1, \dots, n\}$, for any $\lambda \in \Lambda$, $\mathcal{I}_\lambda = \{1, \dots, \lambda\}$ and, for any $i \in \mathcal{I}_\lambda$, $\varphi_i = \sqrt{\lambda} \mathbf{1}_{[(i-1)/\lambda, i/\lambda)}$. A Parzen kernel is defined by

$$k_{K,h}(x,y) = \frac{1}{h} K\left(\frac{x-y}{h} \right) \ ,$$

where K is a bounded, symmetric, real valued function and $h > 0$ is a bandwidth. In both cases, the estimator $\widehat{f}_k(x) = \frac{1}{n} \sum_{i=1}^n k(X_i, x)$. Elementary algebraic computation show that the criterion minimized by $\widehat{k}$ is equal to

$$\begin{aligned} P_n \gamma(\widehat{f}_k) + \mathrm{pen}(k) = \mathrm{pen}(k) &- \frac{2P\chi_k - P\Xi_k}{n} + \frac{(P_n - P)(\Xi_k - 2\chi_k)}{n} \\ &+ \frac{U_{A,k} - 2U_k}{n^2} + \frac{2(n-1)}{n}(P_n - P)(F_{A,k} - 2f_k^\star) \\ &+ \frac{n-1}{n}\left(\|f_k^\star\|^2 - 2Pf_k^\star \right) \ , \end{aligned}$$

where $\chi_k(x) = k(x,x)$, $\Xi_k(x) = A_k(x,x)$, $A_k(x,y) = \int k(x,z)k(z,y)d\mu(z)$,

$$U_{A,k} = \sum_{1 \leq i < j \leq n} A_k(X_i, X_j) - F_{A,k}(X_i) - F_{A,k}(X_j) + \mathbb{E}\left[A_k(X_i, X_j) \right] \ ,$$

$$U_k = \sum_{1 \leq i < j \leq n} k(X_i, X_j) - f_k^\star(X_i) - f_k^\star(X_j) + \mathbb{E}\left[k(X_i, X_j) \right] \ ,$$

$F_{A,k}(x) = \mathbb{E}\left[A_k(X,x) \right]$, $f_k^\star(x) = \mathbb{E}\left[k(X,x) \right]$. Thus, Bernstein's concentration inequality and concentration bounds for totally degenerate U-statistics of order 2 show that, not only in expectation, but also with large

probability, the criterion is well approximated by

$$\|f_k^\star - f^\star\|^2 - \frac{2P\chi_k - P\Xi_k}{n} + \mathrm{pen}(k) - \|f^\star\|^2 \quad .$$

Hence, if, for some $u > 0$, $\mathrm{pen}(k) = \frac{2P\chi_k - P\Xi_k}{n} - u\frac{P\Xi_k}{n}$, $\widehat{k}$ minimizes essentially

$$\|f_k^\star - f^\star\|^2 - u\frac{P\Xi_k}{n} \quad ,$$

while, if $\mathrm{pen}(k) = \frac{2P\chi_k - P\Xi_k}{n} + u\frac{P\Xi_k}{n}$, for some $u > 0$, $\widehat{k}$ minimizes essentially

$$\|f_k^\star - f^\star\|^2 + u\frac{P\Xi_k}{n} \quad .$$

Let us discuss this result in our examples. For any projection kernel k_λ, it comes from the orthonormality of the system $(\varphi_i)_{i\in\mathcal{I}_\lambda}$ that $A_{k_\lambda}(x,y) = k_\lambda(x,y)$, hence, for the projection kernels onto the histogram spaces,

$$A_{k_\lambda}(x,x) = k_\lambda(x,x) = \lambda\sum_{i=1}^{\lambda}\mathbf{1}_{[(i-1)/\lambda, i/\lambda)} = \lambda\mathbf{1}_{[0,1)} \quad ,$$

and $P\Xi_{k_\lambda} = \lambda$. Moreover, $f_{k_\lambda}^\star$ is the projection of $f^\star$ onto the histogram space with bin size $1/\lambda$ and the bias is then essentially a non increasing function of λ. Thus, when $\mathrm{pen}(\lambda) = \frac{2P\chi_k - P\Xi_k}{n} - u\frac{P\Xi_k}{n} = \frac{(1-u)\lambda}{n}$, $k_{\widehat{\lambda}}$ minimizes a criterion where both term are non-increasing with λ. It is shown for example in [LMRB14] that, if $f^\star$ is α-Hölderian, $k_{\widehat{\lambda}}$ is a model with $\widehat{\lambda} \geq \square n$ asymptotically for some positive constant $\square > 0$ and $\widehat{f}_{k_{\widehat{\lambda}}}$ is not a consistent estimator of $f^\star$. On the other hand, when $\mathrm{pen}(\lambda) = \frac{(1+u)\lambda}{n}$, $k_{\widehat{\lambda}}$ balances a bias term $\left\|f_{k_\lambda}^\star - f^\star\right\|^2$ and a variance term $u\lambda/n$. Its complexity $\widehat{\lambda}$ is much more reasonable and $\widehat{f}_{k_{\widehat{\lambda}}}$ satisfies an oracle inequality with a constant that depends on u. In particular, it implies that $\widehat{f}_{k_{\widehat{\lambda}}}$ is consistent and converges to $f^\star$ at the minimax rate of convergence. The sharp phase transition in the behavior of $\widehat{f}_{k_{\widehat{\lambda}}}$ shows that $\mathrm{pen}(\lambda) = \frac{\lambda}{n}$ is a minimal penalty in this problem.

For the second example, fix K and consider the problem of the choice of the bandwidth h. Elementary algebra shows that $\chi_{k_{K,h}}(x) = K(0)/h$ and $\Xi_{k_{K,h}}(x) = \|K\|^2/h$. Moreover, as in the previous example, as $f_{k_{K,h}}^\star$ is the convolution of $f^\star$ with the approximation to the identity $K(./h)/h$, the bias term $\left\|f^\star - f_{k_{K,h}}^\star\right\|^2$ is essentially non-increasing with $1/h$. Thus, when

$$\mathrm{pen}(h) = \frac{2P\chi_k - P\Xi_k}{n} - u\frac{P\Xi_k}{n} = \frac{2K(0) - \|K\|^2}{nh} - u\frac{\|K\|^2}{nh} \quad ,$$

$k_{K,\widehat{h}}$ minimizes a criterion where both terms are non-increasing with $1/h$. It is shown in [LMRB14] that, if $f^\star$ is α-Hölderian, $k_{K,\widehat{h}}$ is a model such that $1/n\widehat{h} \geq \square n$ asymptotically for some positive constant $\square > 0$ and $\widehat{f}_{K,\widehat{h}}$ is not a consistent estimator of $f^\star$. On the other hand, when

$$\mathrm{pen}(h) = \frac{2K(0) - \|K\|^2}{n} + u\frac{\|K\|^2}{nh} \quad ,$$

$k_{K,\widehat{h}}$ balances a bias term $\left\|f_{k_{K,h}}^\star - f^\star\right\|^2$ and a variance term $u\|K\|^2/(hn)$. Its complexity $1/\widehat{h}$ is much more reasonable and $\widehat{f}_{k_{K,\widehat{h}}}$ satisfies an oracle inequality with a leading constant that depends on u. In addition, it is

consistent and converges to $f^\star$ at optimal rate of convergence in the minimax sense. As in the previous example, the sharp phase transition in the behavior of $\widehat{f}_{\widehat{K,h}}$ shows that $\mathrm{pen}(h) = \frac{2K(0)-\|K\|^2}{hn}$ is a minimal penalty in this problem.
These examples are typical of the minimal penalty phenomenon of [BM07] in estimator selection. The first point is that there is a sharp phase transition in the behavior of the selected estimator when the penalty goes from $(1-u)\,\mathrm{pen}_{\min}$ to $(1+u)\,\mathrm{pen}_{\min}$. The selected estimator, that was not even consistent, becomes an oracle. The second point is that this phase transition can be observed on the "complexity" of the selected estimator : the complexity is usually a deterministic quantity such as λ in the histogram example or $1/h$ in the regularization example, the one of the selected estimator decreases very rapidly during this phase transition, we refer to [AM09] for more details and to the next section for more references.

1.3.2. *Learning from bad behavior: the slope algorithm*

Birgé and Massart introduced the slope heuristics that states that an optimal penalty is equal to $2 \times \mathrm{pen}_{\min}$. The idea, as described in [AM09] is the following. A penalized criterion is equal to

$$P_n\gamma(\widehat{f}_\lambda) + \mathrm{pen}(\lambda) \ .$$

Since we would like to minimize $P\gamma(\widehat{f}_\lambda)$, an ideal penalty is given by

$$(P-P_n)\gamma(\widehat{f}_\lambda) = P\left(\gamma(\widehat{f}_\lambda) - \gamma(f_\lambda^\star)\right) + P_n\left(\gamma(f_\lambda^\star) - \gamma(\widehat{f}_\lambda)\right) + (P-P_n)\gamma(f_\lambda^\star) \ ,$$

where $f_\lambda^\star$ denotes, for example the orthogonal projection of $f^\star$ onto S_λ in the model selection framework, or the convolution $f_k^\star$ in the regularization kernel example. Arlot and Massart [AM09] called respectively $p_1(\lambda)$, $p_2(\lambda)$ and $\delta(\lambda)$ the three terms in the right hand side of this equality. $\delta(\lambda)$ is centered and well concentrated thanks to Bernstein's inequality, so let's forget about it. It is then clear that $p_1(\lambda)+p_2(\lambda)$ is an ideal penalty. Moreover, $p_2(\lambda)$ is a minimal penalty, since, if $\mathrm{pen}(\lambda) = p_2(\lambda)$, the criterion is equal to $P_n(\gamma(f_\lambda^\star))$, which concentrates around $P\gamma(f_\lambda^\star)$ by Bernstein's inequality, which is the bias part of the risk. As discussed in the example, this bias is usually minimized for large values of the complexity. On the other hand, if $\mathrm{pen}(\lambda) = p_2(\lambda) + up_1(\lambda)$ the criterion balances the same bias and a term proportional to the variance. Consequently, the selected estimator will satisfy an oracle inequality.
The slope heuristic then easily follows if $p_1(\lambda) \simeq p_2(\lambda)$. This is usually the case because the expectations

$$\mathbb{E}\left[p_1(\lambda)\right] \simeq \mathbb{E}\left[p_2(\lambda)\right] \ ,$$

and because both quantities concentrate around their expectation. For projection kernels, we have

$$p_1(\lambda) = \left\|\widehat{f}_\lambda\right\|^2 - \|f_\lambda^\star\|^2 - 2\langle \widehat{f}_\lambda - f_\lambda^\star, f_\lambda^\star\rangle = \left\|\widehat{f}_\lambda\right\|^2 - \|f_\lambda^\star\|^2 - 2\langle \widehat{f}_\lambda, f_\lambda^\star\rangle = \left\|\widehat{f}_\lambda - f_\lambda^\star\right\|^2 \ .$$
$$p_2(\lambda) = \|f_\lambda^\star\|^2 - \left\|\widehat{f}_\lambda\right\|^2 - 2\langle f_\lambda^\star - \widehat{f}_\lambda, \widehat{f}_\lambda\rangle = \|f_\lambda^\star\|^2 + \left\|\widehat{f}_\lambda\right\|^2 - 2\langle f_\lambda^\star, \widehat{f}_\lambda\rangle = \left\|\widehat{f}_\lambda - f_\lambda^\star\right\|^2 \ .$$

Hence, $p_1(\lambda) = p_2(\lambda)$ and the slope heuristic holds, see also [Ler12] for a rigorous proof in the density estimation framework. On the other hand, for regularization kernel, [LMRB14] prove that p_1 and p_2 concentrate around their expectations that are respectively given by

$$\mathbb{E}\left[p_1(K,h)\right] = \frac{2K(0)-\|K\|^2}{hn}, \qquad \mathbb{E}\left[p_2(K,h)\right] = \frac{\|K\|^2}{hn} \ .$$

There are several unusual facts that can be noticed here. First, the relation optimal penalty equals 2 times the minimal one is only true in the particular case where $\|K\|^2 = K(0)$ which is not verified by many classical

kernels, as the Gaussian kernel $K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ or the Epanechnikov kernel $K(x) = \frac{3}{4}(1-x^2)_+$ for example. However, it can easily be corrected when we only want to select h. Actually, if $\|K\|^2 \neq 2K(0)$, the optimal penalty $\mathrm{pen}_{\mathrm{opt}}(k) = 2K(0)/(nh)$ and the minimal one $\mathrm{pen}_{\mathrm{min}}(k) = (2K(0) - \|K\|^2)/(nh)$ satisfy

$$\mathrm{pen}_{\mathrm{opt}}(k) = \frac{2K(0)}{2K(0) - \|K\|^2} \mathrm{pen}_{\mathrm{min}}(k) \ .$$

This type of non trivial relationship between optimal and minimal penalty has already been underlined in [AB09] in regression framework for selecting linear estimators. However, note that if one allows two kernel functions K_1 and K_2 in the family of kernels such that $2K_1(0) \neq \|K_1\|^2$, $2K_2(0) \neq \|K_2\|^2$ and

$$\frac{2K_1(0)}{2K_1(0) - \|K_1\|^2} \neq \frac{2K_2(0)}{2K_2(0) - \|K_2\|^2} \ ,$$

there is no simple relationship between the minimal penalty and the optimal one. Next, the minimal penalty $p_1(K,h)$ can be negative, if $\|K\|^2 > 2K(0)$, in that case, a minimizer of the empirical risk $P_n\gamma(\widehat{f}_{K,h})$ satisfy an oracle inequality! All these facts are illustrated in the paper [LMRB14].
A very interesting feature of the slope phenomenon is it can be used to calibrate a penalty. Actually, [AM09] propose to detect the minimal penalty by the explosion of the complexity of the selected estimator and then to use that the optimal penalty equals twice the minimal one (or sometimes another constant) to calibrate the final estimator. They show that this "slope algorithm" actually has very nice theoretical properties. The practical implementation of the slope algorithm is extensively discussed in [BMM10]. It is particularly useful in practical situations where no reasonable constants can be proposed from the theory, as in the simulation study in [LT11, LT14]. Another interesting example of practical application of the slope algorithm in a segmentation problem was presented by Alice Cleynen in the Journées MAS 2014. She presented a result obtained in her paper [CL14].

2. Selection with test

This section presents recent developments due to Baraud [Bar11] for estimator selection based on ideas of Birgé that are summarized in [Bir06]. The setting is very general and the L^2-loss that was used in the previous sections is replaced here by the Hellinger loss h to avoid unnecessary assumptions on the target $f^\star$.

2.1. **Setting**

The set of probability densities is endowed with the Hellinger distance h defined for any densities f and g with respect to μ by

$$h^2(f,g) = \frac{1}{2}\int_{\mathbb{R}} (\sqrt{f} - \sqrt{g})^2 d\mu = \left\| \sqrt{f} - \sqrt{g} \right\|^2 = 1 - \rho(f,g) \ ,$$

where the Hellinger affinity $\rho(f,g) = \int \sqrt{fg} d\mu$. This distance has many advantages over the L^2-distance, in particular, it can be defined for any density and it does not depend on the measure μ. As in the previous section, a set of estimators $(\widehat{f}_\lambda)_{\lambda\in\Lambda}$ is given and we want to select one with a small Hellinger risk, that is,

$$R(\lambda) = \mathbb{E}\left[h^2(f^\star, \widehat{f}_\lambda) \right] \ .$$

To simplify the presentation, we assume furthermore that the candidates $\widehat{f}_\lambda$ are fixed functions and we denote $\widehat{f}_\lambda = f_\lambda$ to emphasize this point. Remark though that this condition is not necessary since a generalization of the following results to random estimators can be found in Baraud [Bar11].

2.2. **Selection between two candidates**

To choose between two estimators f_λ and $f_{\lambda'}$ the closest to the target $f^\star$, we would like to define the "ideal" function ψ_{id} such that

$$\begin{aligned}\psi_{id}(\lambda,\lambda') &= \lambda, \qquad \text{if} \qquad h(f^\star,f_\lambda) < h(f^\star,f_{\lambda'}) \\ \psi_{id}(\lambda,\lambda') &= \lambda', \qquad \text{if} \qquad h(f^\star,f_\lambda) > h(f^\star,f_{\lambda'}) \ .\end{aligned}$$

Since ψ_{id} takes two values, it is called a test. The problem is that the estimation of the Hellinger loss of a function f_λ is not easy in general. Baraud [Bar11] proposes the following heuristic. Since $h(f^\star,f_{\lambda'})^2 = 1-\rho(f^\star,f_{\lambda'})$ the Hellinger affinity can be used equivalently to define ψ_{id}. Moreover,

$$2\rho(f^\star,f_{\lambda'}) \le \int f^\star \sqrt{\frac{f_{\lambda'}}{r_{\lambda,\lambda'}}}d\mu + \int \sqrt{r_{\lambda,\lambda'}f_{\lambda'}}d\mu =: \rho_{r_{\lambda,\lambda'}}(f_{\lambda'}) \ ,$$

for $r_{\lambda,\lambda'} = (f_\lambda + f_{\lambda'})/2$. Hence, an alternative to the ideal test is given by

$$\begin{aligned}\psi'_{id}(\lambda,\lambda') &= \lambda, \qquad \text{if} \qquad \rho_{r_{\lambda,\lambda'}}(f_\lambda) > \rho_{r_{\lambda,\lambda'}}(f_{\lambda'}) \\ \psi'_{id}(\lambda,\lambda') &= \lambda', \qquad \text{if} \qquad \rho_{r_{\lambda,\lambda'}}(f_\lambda) > \rho_{r_{\lambda,\lambda'}}(f_{\lambda'}) \ .\end{aligned}$$

The great advantage of the function $\rho_{r_{\lambda,\lambda'}}$ is it can be estimated easily by

$$\widehat{\rho}_{r_{\lambda,\lambda'}}(f_{\lambda'}) = \frac{1}{n}\sum_{i=1}^n \sqrt{\frac{f_{\lambda'}(X_i)}{r_{\lambda,\lambda'}(X_i)}} + \int \sqrt{r_{\lambda,\lambda'}f_{\lambda'}}d\mu \ .$$

Baraud's test is then based on the functional

$$\begin{aligned}T(\lambda,\lambda') &= \widehat{\rho}_{r_{\lambda,\lambda'}}(f_{\lambda'}) - \widehat{\rho}_{r_{\lambda,\lambda'}}(f_\lambda) \\ &= \frac{1}{n}\sum_{i=1}^n \frac{\sqrt{f_{\lambda'}(X_i)} - \sqrt{f_\lambda(X_i)}}{\sqrt{f_\lambda(X_i)+f_{\lambda'}(X_i)}} + \frac{1}{2}\int \sqrt{f_\lambda(x)+f_{\lambda'}(x)}\left(\sqrt{f_{\lambda'}(x)} - \sqrt{f_\lambda(x)}\right)dx \ ,\end{aligned}$$

with the convention $0/0 = 0$. The decision rule is given by

$$\psi(\lambda,\lambda') = \begin{cases} \lambda' & \text{if } T(\lambda,\lambda') > 0 \\ \lambda & \text{if } T(\lambda,\lambda') < 0 \end{cases} \ .$$

When $T(\lambda,\lambda') = 0$, define $\psi(\lambda,\lambda')$ arbitrarily in $\{\lambda,\lambda'\}$. Define $\widehat{\lambda} \in \{\lambda,\lambda'\}$ the parameter selected by this procedure. The following proposition is proved in [Bar11].

Proposition 3. *There exists an universal constant C such that, for any $x > 0$,*

$$\mathbb{P}\left(Ch^2(f^\star,f_{\widehat{\lambda}}) \ge \inf\left\{h^2(f^\star,f_\lambda), h^2(f^\star,f_{\lambda'})\right\} + x\right) \le e^{-nx} \ .$$

Remark 4. The log-likelihood ratio test does not satisfy this proposition without assumptions on the distributions $f^\star, f_\lambda, f_{\lambda'}$. However, Le Cam [LC75] and Birgé [Bir84], see also Proposition 6 in [Bir06] showed that it can be modified to achieve this goal. The idea is to make a log-likelihood ratio test between the closest points in the balls centered at f_λ and $f_{\lambda'}$ of radius $h(f_\lambda,f_{\lambda'})/4$.

2.3. **Selection among a finite collection**

To select among a larger family of estimators $(f_\lambda)_{\lambda\in\Lambda}$, Birgé proposes to use the tests in the following procedure. For any $\lambda \in \Lambda$, define

$$\mathcal{R}(\lambda) = \{\, \lambda' \in \Lambda,\ \text{s.t.}\ \ \lambda' \neq \lambda \text{ and } \psi(\lambda,\lambda') = \lambda' \,\}\,, \quad \text{and} \quad \mathcal{D}(\lambda) = \sup_{\lambda'\in\mathcal{R}(\lambda)} h^2(f_\lambda, f_{\lambda'})\ ,$$

with the convention $\sup_\emptyset = 0$. The idea is that, when $\mathcal{D}(\lambda)$ is small, any density $f_{\lambda'}$ that is preferred to f_λ is close to f_λ, hence, f_λ must have a risk that is close to the optimal. This is why the final estimator is defined by

$$\widehat{\lambda} \in \arg\min_{\lambda\in\Lambda} \mathcal{D}(\lambda)\ .$$

By Theorem 3 in [Bir06] or Theorem 2 in [Bar11], this estimator satisfies the following proposition.

Proposition 4. *There exists universal constants a and C such that, if $\frac{\log(|\Lambda|)}{n} \leq a$,*

$$\forall x > 0, \qquad \mathbb{P}\left(Ch^2(f^\star, f_{\widehat{\lambda}}) \geq \inf_{\lambda\in\Lambda} h^2(f^\star, f_\lambda) + \frac{\log|\Lambda|}{n} + x \right) \leq e^{-nx}\ .$$

The theory of estimation with tests has been considered a powerful theoretical tool for a long time. Actually, it does not require many assumptions on the observations and very general optimal risk bounds can be shown for the selected estimator, see for example Birgé [Bir06] and Baraud [Bar11]. In particular, the assumption that Λ is finite can be refined using the more general notion of D-model, see [Bir06]. It allows to work with infinite collection of estimators and to avoid the logarithmic loss in some examples in Proposition 4 . However, as noticed by the authors, the practical implementation of the estimators remained intractable in these works, and was only possible in a very particular framework of Gaussian regression [BGH14]. Recently Mathieu Sart [Sar14] and Nelo Magalhães, who presented his results in the Journées MAS 2014 got interested in the practical implementation of the estimators. It was quite a surprise to remark that these estimators, that were designed to achieve risks bounds in general frameworks, show very good performances in practice too, at least in their preliminary experiments.
Once "robust" estimators are built on a single collection $(f_\lambda)_{\lambda\in\Lambda}$, penalization technics can be used to select between various Λ, as we did in Section 1. This was the idea in [Bir06, Bar11, Sar14]. Nelo Magalhães presented an alternative approach based on the V-fold cross-validation principle that we introduced in Section 1.2.2.

Acknowledgments I would like to thank gratefuly both referees for their helpful remarks and comments.

References

[AB09] S. Arlot and F. Bach. Data-driven calibration of linear estimators with minimal penalties. *Advances in Neural Information Processing Systems (NIPS)*, 22:46–54, 2009.
[AL14] S. Arlot and M. Lerasle. Why $v = 5$ is enough in v-fold cross-validation. *Submitted*, 2014.
[ALM14] S. Arlot, M. Lerasle, and N. Magalhães. v-fold selection of kernel estimators. *Preprint*, 2014.
[AM09] S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *Journal of Machine Learning Research*, 10:245–279, February 2009.
[Arl08] Sylvain Arlot. V-fold cross-validation improved: V-fold penalization, February 2008. arXiv:0802.0566v2.
[Arl09] S. Arlot. Model selection by resampling penalization. *Electron. J. Stat.*, 3:557–624, 2009.
[Bar11] Y. Baraud. Estimator selection with respect to Hellinger-type risks. *Probab. Theory Related Fields*, 151(1-2):353–401, 2011.
[BBM99] Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.

[BGH14] Y. Baraud, C. Giraud, and S. Huet. Estimator selection in the Gaussian setting. *Ann. Inst. Henri Poincaré Probab. Stat.*, 50(3):1092–1119, 2014.

[Bir84] L. Birgé. Sur un théorème de minimax pour des variables indépendantes équidistribuées. *Probab. Math. Statist.*, 3:259–282, 1984.

[Bir06] L. Birgé. Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *Ann. Inst. H. Poincaré Probab. Statist.*, 42(3):273–325, 2006.

[BM07] L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73, 2007.

[BMM10] D. Baudry, C. Maugis, and B. Michel. Slope heuristics: overview and implementation. *INRIA report, available at http://hal.archives-ouvertes.fr/hal-00461639/fr/*, 2010.

[BTWB10] F. Bunea, A. B. Tsybakov, M. H. Wegkamp, and A. Barbu. Spades and mixture models. *Ann. Statist.*, 38(4):2525–2558, 2010.

[BvdG11] P. Bühlmann and S. van de Geer. *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg, 2011.

[Cat04] O. Catoni. *Statistical learning theory and stochastic optimization*. Springer, New-York, 2004.

[Cel12] A. Celisse. Optimal cross-validation in density estimation. *preprint*, 2012.

[CL14] A. Cleynen and E. Lebarbier. Segmentation of the poisson and negative binomial rate models: a penalized estimator. *ESAIM P&S*, 2014.

[DJ94] David L. Donoho and Iain M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

[DJKP96] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Density estimation by wavelet thresholding. *Ann. Statist.*, 24(2):508–539, 1996.

[DT08] A. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp pac-bayesian bounds and sparcity. *Mach. Learn.*, 72(1-2):39–61, 2008.

[Efr83] B. Efron. Estimating the error rate of a prediction rule: im- provement on cross-validation. *J. Amer. Statist. Assoc.*, 78(382):316–331, 1983.

[Fro04] M. Fromont. Model selection by bootstrap penalization for classification. *Learning theory*, 3120:285–299, 2004.

[LC75] L. Le Cam. On local and global properties in the theory of of asymptotic normality experiments. *Stochastic Processes and related topics*, 1:13–54, 1975.

[Ler12] M. Lerasle. Optimal model selection in density estimation. *Ann. Inst. Henri Poincaré Probab. Stat.*, 48(3):884–908, 2012.

[LMRB14] M. Lerasle, N. Magalhães, and P. Reynaud-Bouret. Optimal kernel selection for density estimation. *Preprint*, 2014.

[LT11] M. Lerasle and D.Y. Takahashi. An oracle approach for interaction neighborhood estimation in random fields. *Electron. J. Stat.*, 5:534–571, 2011.

[LT14] M. Lerasle and D.Y. Takahashi. Sharp oracle inequalities and slope heuristic for specification probabilities estimation in discrete random fields. *Accepted for publication in Bernoulli*, 2014.

[San14] L. Sansonnet. Wavelet thresholding estimation in a poissonian interactions model with application to genomic data. *Scandinavian Journal of Statistics*, 41(1):200–226, March 2014.

[Sar14] M. Sart. Estimation of the transition density of a Markov chain. *Ann. Inst. Henri Poincaré Probab. Stat.*, 50(3):1028–1068, 2014.

[Tib96] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc., Ser. B*, 58:267–288, 1996.

[ZH05] H. Zou and T Hastie. Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B*, 67:301–320, 2005.