Sankhyā A : The Indian Journal of Statistics
DOI 10.1007/s13171-017-0107-5

Estimator Selection: a New Method with Applications to Kernel Density Estimation

Claire Lacour and Pascal Massart
University Paris-Sud, Orsay, France
Vincent Rivoirard
Université Paris Dauphine, Paris, France

Abstract

Estimator selection has become a crucial issue in non parametric estimation. Two widely used methods are penalized empirical risk minimization (such as penalized log-likelihood estimation) or pairwise comparison (such as Lepski's method). Our aim in this paper is twofold. First we explain some general ideas about the calibration issue of estimator selection methods. We review some known results, putting the emphasis on the concept of minimal penalty which is helpful to design data-driven selection criteria. Secondly we present a new method for bandwidth selection within the framework of kernel density density estimation which is in some sense intermediate between these two main methods mentioned above. We provide some theoretical results which lead to some fully data-driven selection strategy.

AMS (2000) *subject classification.* Primary: 62G05; Secondary: 62C20, 62H12.
Keywords and phrases. Concentration inequalities, Kernel density estimation, Penalization methods, Estimator selection, Oracle inequality

1 Introduction

Sincethe beginning of the 80's and the pioneering works of Pinsker (see Pinsker, 1980; Efroimovitch and Pinsker, 1984) many efforts have been made in nonparametric statistics to design and analyze adaptive estimators. Several breakthroughs have been performed in the 90's. Lepski initiated a strategy mainly based on a new bandwidth selection method for kernel estimators (see Lepskii, 1990, 1991) while Donoho, Johnstone, Kerkyacharian and Picard devoted a remarkable series of works to adaptive properties of wavelet thresholding methods (see Donoho and Johnstone, 1994a, b; Donoho et al., 1995 for an overview). In the same time, following a somehow more abstract information-theoretic path Barron and Cover (1991) have built adaptive density estimators by using some minimum complexity principle on discrete

models, paving the way for a more general connection between model selection and adaptive estimation that has been performed in a series of works by Birgé and Massart (see Birge and Massart, 1998; Barron et al., 1999; Massart, 2007 for an overview). Estimator selection is a common denominator between all these works. In other words, if one observes some random variable ξ (which can be a random vector or a random process) with unknown distribution, and one wants to estimate some quantity f (the target) related to the distribution of ξ, a flexible approach to estimate some target f is to consider some collection of preliminar estimators $\left\{\hat{f}_m, m \in \mathcal{M}\right\}$ and then try to design some genuine data-driven procedure $\hat{m}$ to produce some new estimator $\hat{f}_{\hat{m}}$. We have more precisely in mind frameworks in which the observation ξ depends on some parameter n (the case $\xi = (\xi_1, ..., \xi_n)$, where the variables $\xi_1, ..., \xi_n$ are independent and identically distributed is a typical example) and when we shall refer to asymptotic results, this will implicitly mean that n goes to infinity. The model selection framework corresponds to the situation where each estimator $\hat{f}_m$ is linked to some model S_m through some empirical risk minimization procedure (least squares for instance) but bandwidth selection for kernel estimators also falls into this estimator selection framework. Since the beginning of the century, the vigorous developments of high-dimensional data analysis led to the necessity of building new inference methods and accordingly new mathematical analysis of the performance of these methods. Dealing with high-dimensional data also raises new questions related to the implementation of the methods: you do not only need to design estimation methods that performs well from a statistical point of view but you also need fast algorithms. In those situations the estimator selection issue is still relevant but you may encounter situations for which not only the estimators but the collection $\mathcal{M}$ itself depends on the data. The celebrated Lasso algorithm introduced by Tibshirani in Tibshirani (1996) (and mathematically studied in Bickel et al. 2009) perfectly illustrates this fact, since in this case the regularization path of the Lasso naturally leads to some data-dependent ordered collection $\mathcal{M}$ of set of variables. Estimator selection has therefore become a central issue in nonparametric statistics. To go further into the subject one needs to consider some loss function ℓ allowing to measure the quality of each estimator $\hat{f}_m$ by $\ell\left(f, \hat{f}_m\right)$. If the target to be estimated as well as the estimators belong to some Hilbert space $(\mathbb{H}, \|.\|)$, the quadratic loss $\ell\left(f, \hat{f}_m\right) = \left\|f - \hat{f}_m\right\|^2$ is a standard choice, but of course other losses are considered in the literature (typically powers of $\mathbb{L}_p$ norms when f is a function). Given this loss function, the quality of a selection procedure $\hat{m}$ is then measured by

$\ell\left(f, \hat{f}_{\hat{m}}\right)$ and mathematical results on estimator selection are formulated in terms ofupper bounds (either in probability or in expectation) on $\ell\left(f, \hat{f}_{\hat{m}}\right)$ that allow to measure how far this quantity is from what is usually called the *oracle* risk $\inf_{m\in\mathcal{M}} \mathbb{E}\ell\left(f, \hat{f}_m\right)$. These comparison inequalities are called oracle inequalities. A common feature of estimator selection methods such as penalized empirical risk minimization or Lepski's method (or more recently Glodenshluger-Lepski's method) is that they involve some "hyper-parameter" λ say that needs to be chosen by the statistician (typically λ is some multiplicative constant in a penalty term). Positive mathematical results on a given selection procedure typically tells you that if λ remains above some quantity $\lambda_{\min}$ then you are able to prove some oracle inequality. Now the point is that in practice these methods are very sensitive to this choice λ. In particular, too small value of λ should be prohibited since they lead to overfitting. Birgé and Massart have introduced in Birgé and Massart (2007) the concept of *minimal penalty*. They prove that in the Gaussian white noise framework, there is number of model selection problems for which it is possible to prove the existence of a critical value $\lambda_{\min}$ for λ (which precisely corresponds to the minimal penalty) below which the selection procedure dramatically fails in the sense that the risk of the selected estimator is of order $\sup_{m\in\mathcal{M}} \mathbb{E}\ell\left(f, \hat{f}_m\right)$ instead of $\inf_{m\in\mathcal{M}} \mathbb{E}\ell\left(f, \hat{f}_m\right)$! Moreover they also prove some oracle inequalities showing that taking λ as twice this critical value $\lambda_{\min}$ corresponds to the "best" choice for λ (at least asymptotically i.e. when the level of noise goes to zero). Since this critical value may be detected from the data (see the slope heuristics described below), the estimated critical value $\hat{\lambda}_{\min}$ leads to an entirely data-driven choice $\hat{\lambda} = 2\hat{\lambda}_{\min}$ for the hyper parameter λ which can be proved to be (nearly) optimal for some model selection problems. This approach for calibrating the penalty has been implemented, tested on simulation studies and used on real in several papers devoted to model selection (see Lebarbier, 2005 for the first simulation studies and Baudry et al., 2011 for an overview on practical aspects). It has also been mathematically studied in several statistical frameworks that differs from the Gaussian white noise (see Arlot and Massart, 2009; Saumard, 2013; Lerasle, 2012; Lerasle and Takahashi, 2016 for instance). In there very nice paper (Arlot and Bach, 2009), Arlot and Bach have adapted this strategy (based on the concept of minimal penalty) for calibrating a least squares penalized criterion to the context of selection of (linear) estimators which are not necessarily least squares estimators. In the same spirit, Lerasle, Malter-Magalahes and Reynaud-Bouret have computed minimal and optimal penalty formulas for the problem of kernel density

estimators selection via some penalized least square criterion (see Lerasle et al., 2015). Minimal penalties for Lasso type estimates can also be also revealed as illustrated by Bertin et al. (2011) in the framework of density estimation. Very recently, Lacour and Massart have shown in Lacour and Massart (2016) that the concept of minimal penalty also makes sense for the Goldenshluger-Lepski method in the context of kernel estimators bandwidth selection. Unlike penalized least squares (or other risk minimization criteria), the Goldenshluger-Lepski method is a pairwise comparison based estimator selection procedure and the minimal penalty result given in Lacour and Massart (2016) is a first step in the direction of designing entirely data-driven calibration strategies for such pairwise comparison methods. From a mathematical view point, it is quite well known now that the proofs of oracle inequalities heavily rely on right tail deviation inequalities for stochastic quantities such as suprema of empirical (or Gaussian) processes. These deviation inequalities may derive or not from concentration inequalities and as a matter of fact many oracle inequalities have been obtained by using deviation inequalities which are not concentration inequalities per se (see Lepskii, 2013 for instance). The price to pay is that the resulting oracle inequalities typically require that $\lambda > \lambda_0$ for some value λ_0 which can be much larger than $\lambda_{\min}$. Interestingly, concentration inequalities lead to sharper results that allows (at least for some specific situations for which things are computable) to derive the computation of $\lambda_{\min}$. Oracle inequalities are obtained by using right tail concentration while the minimal penalty results are proved by using left tails concentration inequalities. Our purpose in this paper is first to recall some heuristics on the penalty calibration method mentioned above in the context of penalized model selection. Then we briefly explain (following the lines of Birgé and Massart, 2007) why concentration is the adequate tool to prove a minimal penalty result. Finally we introduce some new estimator selection method which lies somewhere between penalized empirical risk minimization and the pairwise comparison methods (such as the Goldenshluger-Lepski's method). We study this new method in the context of bandwidth selection for kernel density estimators with the squared $\mathbb{L}_2$ loss. We provide oracle inequalities and minimal penalty results that allow to compute minimal and optimal values for the penalty.

2 Penalized Model Selection

Viewing model selection as a special instance of estimator selection requires to formulate the model selection problem as an estimation problem of some target quantity f (typically, f is a function) belonging to some space $\mathcal{S}$ say as suggested above. In the model selection framework, the list of estimators

and the loss function are intimately related in the sense that they derive from the same *contrast function* (also called *empirical risk* in the machine learning literature). More precisely, a contrast function L_n is a function on the set $\mathcal{S}$ depending on the observation $\xi^{(n)}$ in such a way that

$$g \to \mathbb{E}\left[L_n\left(g\right)\right]$$

achieves a minimum at point f. Given some collection of subsets $(S_m)_{m\in\mathcal{M}}$ of $\mathcal{S}$ (these subsets are simply called models in what follows), for every $m \in \mathcal{M}$, some estimator $\hat{f}_m$ of f is obtained by minimizing L_n over S_m ($\hat{f}_m$ is called *minimum contrast estimator* or *empirical risk minimizer*). On the other hand some "natural" (non negative) loss function can also be attached to L_n through the simple definition

$$\ell\left(f,g\right) = \mathbb{E}\left[L_n\left(g\right)\right] - \mathbb{E}\left[L_n\left(f\right)\right] \tag{1}$$

for all $g \in \mathcal{S}$. In the case where $\xi^{(n)} = (\xi_1, ..., \xi_n)$, an empirical criterion L_n can be defined as an empirical mean

$$L_n\left(g\right) = P_n\left[L\left(g,.\right)\right] := \frac{1}{n}\sum_{i=1}^{n} L\left(g,\xi_i\right),$$

which justifies the terminology of empirical risk. Empirical risk minimization includes maximum likelihood and least squares estimation as briefly recalled below in the contexts of density estimation and Gaussian white noise.

- **Density estimation**

One observes $\xi_1, ..., \xi_n$ which are i.i.d. random variables with unknown density f with respect to some given measure μ. The choice

$$L\left(g,x\right) = -\ln\left(g\left(x\right)\right)$$

leads to maximum likelihood estimation and the corresponding loss function ℓ is given by

$$\ell\left(f,g\right) = K\left(f,g\right),$$

where $K\left(f,g\right)$ denotes the Kullback-Leibler divergence between the probabilities $f\mu$ and $g\mu$, i.e.

$$K\left(f,g\right) = \int f \ln\left(\frac{f}{g}\right) d\mu$$

if $f\mu$ is absolutely continuous with respect to $g\mu$ and $K(f,g) = +\infty$ otherwise. Assuming that $f \in \mathbb{L}_2(\mu)$, it is also possible to define a least squares density estimation procedure by setting this time

$$L(g,x) = \|g\|^2 - 2g(x)$$

where $\|.\|$ denotes the norm in $\mathbb{L}_2(\mu)$ and the corresponding loss function ℓ is in this case given by

$$\ell(f,g) = \|f-g\|^2,$$

for every $g \in \mathbb{L}_2(\mu)$.

- **Gaussian white noise**

Let us consider the general Gaussian white noise framework as introduced in Birgé and Massart (2001). This means that, given some separable Hilbert space $\mathbb{H}$, one observes

$$\xi(g) = \langle f, g\rangle + \varepsilon W(g), \text{ for all } g \in \mathbb{H}, \tag{2}$$

where W is some isonormal process, i.e. W maps isometrically $\mathbb{H}$ onto some Gaussian subspace of $\mathbb{L}_2(\Omega)$. This framework is convenient to cover both the infinite dimensional white noise model for which $\mathbb{H} = \mathbb{L}_2([0,1])$ and $W(g) = \int_0^1 g(x)\,dB(x)$, where B is a standard Brownian motion, and the finite dimensional linear model for which $\mathbb{H} = \mathbb{R}^n$ and $W(g) = \langle \eta, g\rangle = \frac{1}{n}\sum_{i=1}^n g_i\eta_i$, where η is a standard n-dimensional Gaussian vector. In what follows, we shall write the level of noise ε as $\varepsilon = 1/\sqrt{n}$. The introduction of n here is purely artificial, it just helps in keeping homogenous notations when passing from the density estimation framework to the Gaussian white noise framework (note that this choice is not misleading since in the n-dimensional case mentioned above the level of noise is actually equal to $1/\sqrt{n}$). The least squares criterion is defined by

$$L_n(g) = \|g\|^2 - 2\xi(g),$$

and the corresponding loss function ℓ is simply the quadratic loss,

$$\ell(f,g) = \|f-g\|^2,$$

for every $g \in \mathbb{H}$.

2.1. The Model Choice Paradigm. The choice of a model S_m on which the empirical risk minimizer is to be defined is a challenging issue. It is generally difficult to guess what is the right model to consider in order to reflect the nature of data from the real life and one can get into problems whenever the model S_m is false in the sense that the true f is too far from S_m. One could then be tempted to choose S_m as big as possible. Taking S_m as $\mathcal{S}$ itself or as a huge subset of $\mathcal{S}$ is known to lead to inconsistent (see Bahadur, 1958) or suboptimal estimators (see Birge and Massart, 1993). Choosing some model S in advance leads to some difficulties

- If S_m is a *small* model (think of some parametric model, defined by 1 or 2 parameters for instance) the behavior of a minimum contrast estimator on S_m is satisfactory as long as f is close enough to S_m but the model can easily turn to be false.
- On the contrary, if S_m is a *huge* model (think of the set of all continuous functions on $[0,1]$ in the regression framework for instance), the minimization of the empirical criterion leads to a very poor estimator of f even if f truly belongs to S_m.

This is by essence what makes the data-driven choice of a model an attractive issue. Interestingly the risk of the empirical risk minimizer on a given model S_m is a meaningful criterion to measure the quality of the model as illustrated by the following example.

Illustration (white noise) In the white noise framework, if one takes S_m as a linear subspace with dimension D_m of $\mathbb{H}$, one can compute the least squares estimator explicitly. Indeed, if $(\phi_j)_{1\leq j\leq D_m}$ denotes some orthonormal basis of S, the corresponding least squares estimator is merely a projection estimator

$$\hat{f}_m = \sum_{j=1}^{D_m} \xi(\phi_j)\,\phi_j.$$

Since for every $1 \leq j \leq D_m$

$$\xi(\phi_j) = \langle f, \phi_j \rangle + \frac{1}{\sqrt{n}}\eta_j$$

where the variables $\eta_1, ..., \eta_D$ are i.i.d. standard normal variables, the expected quadratic risk of $\hat{f}_m$ can be easily computed

$$\mathbb{E}\left[\left\|f - \hat{f}_m\right\|^2\right] = d^2(f, S_m) + \frac{D_m}{n}.$$

This formula for the quadratic risk perfectly reflects the model choice paradigm since if one wants to choose a model in such a way that the risk of the resulting least square estimator is small, we have to warrant that the bias term $d^2(f, S_m)$ and the variance term D_m/n are small simultaneously. Therefore if $(S_m)_{m\in\mathcal{M}}$ is a list of finite dimensional subspaces of $\mathbb{H}$ and $\left(\hat{f}_m\right)_{m\in\mathcal{M}}$ be the corresponding list of least square estimators, it is relevant to consider that an *ideal* model should minimize $\mathbb{E}\left[\left\|f-\hat{f}_m\right\|^2\right]$ with respect to $m\in\mathcal{M}$, which amounts to consider this model selection problem as an instance of estimator selection.

More generally if we consider some empirical contrast L_n and some collection of models $(S_m)_{m\in\mathcal{M}}$, each model S_m is represented by the corresponding empirical risk minimizer $\hat{f}_m$ related to L_n. The criterion L_n also provides a natural loss function ℓ and the benchmark for this estimator selection problem is merely $\inf_{m\in\mathcal{M}}\mathbb{E}\left[\ell\left(f,\widehat{f}_m\right)\right]$. In other words, one would ideally like to select $m(f)$ minimizing the risk $\mathbb{E}\left[\ell\left(f,\widehat{f}_m\right)\right]$ with respect to $m\in\mathcal{M}$. Note that the resulting *oracle* model $S_{m(f)}$ is not necessarily a *true* model and in this sense this approach to model selection differs from other approaches which are based on the concept of a true model. Nevertheless we stick now to the estimator selection approach presented above and the issue becomes to design data-driven criteria to select an estimator which tends to mimic an oracle, i.e. one would like the risk of the selected estimator $\hat{f}_{\hat{m}}$ to be as close as possible to the benchmark $\inf_{m\in\mathcal{M}}\mathbb{E}\left[\ell\left(f,\widehat{f}_m\right)\right]$.

2.2. Model Selection via Penalization. The penalized empirical risk selection procedure consists in considering some proper penalty function pen: $\mathcal{M}\to\mathbb{R}_+$ and take $\hat{m}$ minimizing

$$L_n\left(\hat{f}_m\right)+\mathrm{pen}(m)$$

over $\mathcal{M}$. We can then define the selected model $S_{\hat{m}}$ and the corresponding selected estimator $\hat{f}_{\hat{m}}$.

Penalized criteria have been proposed in the early seventies by Akaike or Schwarz (see Akaike, 1973; Schwarz, 1978) for penalized maximum log-likelihood in the density estimation framework and Mallows for penalized least squares regression (see Daniel and Wood, 1971; Mallows, 1973). In both cases the penalty functions are proportional to the number of parameters D_m of the corresponding model S_m

- Akaike (AIC): D_m/n

- Schwarz (BIC): $\ln(n) D_m/(2n)$
- Mallows (C_p): $2D_m\sigma^2/n$,

where the variance σ^2 of the errors of the regression framework is assumed to be known by the sake of simplicity.

Akaike's or Schwarz's proposals heavily rely on the assumption that the dimensions and the number of the models are bounded w.r.t. n and n tends to infinity.

Viewing model selection as an instance of estimator selection is much more related to the non asymptotic view of model selection. In other words, since the purpose of the game is to mimic the oracle, it makes sense to consider that n is what it is and to allow the list of models as well as the models themselves to depend on n.

As we shall see, concentration inequalities are deeply involved both in the construction of the penalized criteria and in the study of the performance of the resulting penalized estimator $\widehat{f}_{\widehat{m}}$ but before going to the mathematical heart of this paper, it is interesting to revisit Akaike's heuristics in order to address the main issue: how to calibrate the penalty in order to mimic the oracle?

2.3. Revisiting Akaike's Heuristics. If one wants to better understand how to penalize optimally and the role that concentration inequalities could play in this matter, it is very instructive to come back to the root of the topic of model selection via penalization i.e. to Mallows' and Akaike's heuristics which are both based on the idea of estimating the risk in an unbiased way (at least asymptotically as far as Akaike's heuristics is concerned). The idea is the following.

Let us consider, in each model S_m some minimizer f_m of $t \to \mathbb{E}[L_n(t)]$ over S_m (assuming that such a point does exist). Defining for every $m \in \mathcal{M}$,

$$\widehat{b}_m = L_n(f_m) - L_n(f) \text{ and } \widehat{v}_m = L_n(f_m) - L_n\left(\widehat{f}_m\right),$$

minimizing some penalized criterion

$$L_n\left(\widehat{f}_m\right) + \text{pen}(m)$$

over $\mathcal{M}$ amounts to minimize

$$\widehat{b}_m - \widehat{v}_m + \text{pen}(m).$$

The point is that $\widehat{b}_m$ is an unbiased estimator of the bias term $\ell(f, f_m)$. If we have in mind to use concentration arguments, one can hope that minimizing the quantity above will be approximately equivalent to minimize

$$\ell(f, f_m) - \mathbb{E}[\widehat{v}_m] + \text{pen}(m).$$

Since the purpose of the game is to minimize the risk $\mathbb{E}\left[\ell\left(f,\widehat{f}_m\right)\right]$, an ideal penalty would therefore be

$$\text{pen}\,(m)=\mathbb{E}\left[\widehat{v}_m\right]+\mathbb{E}\left[\ell\left(f_m,\widehat{f}_m\right)\right].$$

In the Mallows' C_p case, the models S_m are linear and $\mathbb{E}\left[\widehat{v}_m\right]=\mathbb{E}\left[\ell\left(f_m,\widehat{f}_m\right)\right]$ are explicitly computable (at least if the level of noise is assumed to be known). For Akaike's penalized log-likelihood criterion, this is similar, at least asymptotically. More precisely, one uses the fact that

$$\mathbb{E}\left[\widehat{v}_m\right]\approx\mathbb{E}\left[\ell\left(f_m,\widehat{f}_m\right)\right]\approx D_m/\left(2n\right),$$

where D_m stands for the number of parameters defining model S_m.

If one wants to design a fully data-driven penalization strategy escaping to asymptotic expansions, a key notion introduced in Birgé and Massart (2007) is the concept of minimal penalty. Intuitively if the penalty is below the critical quantity $\mathbb{E}\left[\widehat{v}_m\right]$, then the penalized model selection criterion should fail in the sense that it systematically selects a model with high dimension (the criterion explodes). Interestingly, this typical behavior helps in estimating the minimal penalty from the data. For instance, if one takes a penalty of the form $\text{pen}\,(m)=\lambda D_m$ one can hope to estimate the minimal value for λ from the data by taking $\hat{\lambda}_{\min}$ as the greatest value λ for which the penalized criterion with penalty λD_m does explode. A complementary idea which is also introduced in Birgé and Massart (2007) is that you may expect the optimal penalty to be related to the minimal penalty. The relationship which is investigated in Birgé and Massart (2007) is especially simple, since it simply amounts to take the optimal penalty as twice the minimal penalty. Heuristically it is based on the guess that the approximation $\mathbb{E}\left[\widehat{v}_m\right]\approx\mathbb{E}\left[\ell\left(f_m,\widehat{f}_m\right)\right]$ remains valid even in a non asymptotic perspective. If this belief is true then the minimal penalty is $\mathbb{E}\left[\widehat{v}_m\right]$ while the optimal penalty should be $\mathbb{E}\left[\widehat{v}_m\right]+\mathbb{E}\left[\ell\left(f_m,\widehat{f}_m\right)\right]$ and their ratio is approximately equal to 2. Hence $2\mathbb{E}\left[\widehat{v}_m\right]$ (which is exactly twice the minimal penalty $\mathbb{E}\left[\widehat{v}_m\right]$!) turns out to be a good approximation of the ideal penalty. Coming back to the typical case when the penalty has the form $\text{pen}\,(m)=\lambda D_m$ this leads to an easy to implement rule of thumb to finally choose the penalty from the data as $\text{pen}\,(m)=2\hat{\lambda}_{\min}D_m$. In some sense explains the rule of thumb which is given in the preceding Section: the minimal penalty is $\widehat{v}_m$ while the optimal penalty should be $\widehat{v}_m+\mathbb{E}\left[\ell\left(f_m,\widehat{f}_m\right)\right]$ and their ratio is approximately equal to 2.

Implicitly we have made as if the empirical losses $\widehat{v}_m$ were close to their expectations for all models at the same time. The role of concentration arguments will be to validate this fact, at least if the list of models is not too rich. In practice this means that starting from a given list of models, one has first to decide to penalize in the same way the models which are defined by the same number of parameters. Then one considers a new list of models $(S_D)_{D\geq 1}$, where for each integer D, S_D is the union of those among the initial models which are defined by D parameters and then apply the preceding heuristics to this new list.

2.4. Concentration Inequalities in Action: a Case Example. Our aim in this section is to explain the role of concentration inequalities in the derivation of minimal penalty results by studying the simple but illustrative example of ordered variable selection within the Gaussian framework. More precisely, let us consider some infinite dimensional separable Hilbert space $\mathbb{H}$ and an orthonormal basis $\{\phi_j, j \in \mathbb{N}^*\}$ of $\mathbb{H}$. Let us assume that one observes the Gaussian white noise process ξ defined by (2). Given some arbitrary integer N, the ordered variable selection problem consists in selecting a proper model $S_{\hat{D}}$ among the collection $\{S_D, 1 \leq D \leq N\}$, where for every D, S_D is defined as the linear span of $\{\phi_j, j \leq D\}$. The penalized least squares procedure consists of selecting $\hat{D}$ minimizing over $D \in [1, N]$, the criterion $\left\|\hat{f}_D\right\|^2 - 2\xi\left(\hat{f}_D\right) + \text{pen}\,(D)$, where $\hat{f}_D$ is merely the projection estimator $\hat{f}_D = \sum_{j=1}^{D} \xi\left(\phi_j\right)\phi_j$. Equivalently $\hat{D}$ minimizes

$$\text{crit}\,(D) = -\sum_{j=1}^{D} \xi^2\left(\phi_j\right) + \text{pen}\,(D)$$

over $D \in [1, N]$. We would like to show why the value $\lambda = 1$ is critical when the penalty is defined as $\text{pen}\,(D) = \lambda D\varepsilon^2$. To make the problem even simpler let us also assume that $f = 0$ (this assumption can be relaxed of course and the interested reader will find in Birgé and Massart (2007) a complete proof of the minimal penalty result under much more realistic assumptions!). In this case the empirical loss can merely be written as

$$\sum_{j=1}^{D} \xi^2\left(\phi_j\right) = \varepsilon^2 \sum_{j=1}^{D} W^2\left(\phi_j\right) = \varepsilon^2 \chi_D^2$$

where the variable χ_D^2 follows a chi-squared distribution with D degrees of freedom. It is easy now to see how fundamental concentration inequalities are for proving a minimal penalty result. Due to the Gaussian framework

that we use here, it's not surprise that the celebrated Gaussian concentration of measure phenomenon can be used. More precisely, we may apply the Gaussian concentration inequality and the Gaussian Poincaré inequality (see Boucheron et al., 2013 for instance) which ensure that on the one hand defining Z_D as either $\chi(D) - \mathbb{E}[\chi(D)]$ or $-\chi(D) + \mathbb{E}[\chi(D)]$,

$$\mathbb{P}\left\{Z_D \geq \sqrt{2x}\right\} \leq e^{-x}, \text{ for all positive } x$$

and on the other hand

$$0 \leq \mathbb{E}\left[\chi_D^2\right] - \left(\mathbb{E}\left[\chi_D\right]\right)^2 \leq 1.$$

Since of course $\mathbb{E}\left[\chi_D^2\right] = D$, this leads to the right tail inequality

$$\mathbb{P}\left\{\chi_D - \sqrt{D} \geq \sqrt{2x}\right\} \leq e^{-x} \tag{3}$$

and to the left tail inequality

$$\mathbb{P}\left\{\chi_D - \sqrt{D-1} \geq \sqrt{2x}\right\} \leq e^{-x}. \tag{4}$$

Taking the penalty as $\text{pen}(D) = \lambda D\varepsilon^2$, here is what can be proved by using these inequalities (The result below is absolutely not new. It is a consequence of much more general results proved in Birgé and Massart (2007). We recall it with its proof for expository reasons).

PROPOSITION 1. *Given δ and λ in $(0,1)$, if N is large enough depending only on λ and δ, and if the penalty is taken as*

$$\text{pen}(D) = \lambda D\varepsilon^2, \text{ for all } D \in [1, N] \tag{5}$$

then

$$\mathbb{P}\left\{\hat{D} \geq \frac{(1-\lambda)}{2}N\right\} \geq 1-\delta \text{ and } \mathbb{E}\left[\left\|\widehat{f}_{\widehat{m}}\right\|^2\right] \geq \frac{(1-\delta)(1-\lambda)}{4}N\varepsilon^2.$$

PROOF. In order to compare the values of the penalized criterion at points D and N, we write

$$\varepsilon^{-2}[\text{crit}(D) - \text{crit}(N)] \geq \chi_N^2 - \chi_D^2 - \lambda N. \tag{6}$$

Now, recalling that $\lambda < 1$ we set

$$0 < \eta = (1-\lambda)/2 < 1/2;\ \theta = \eta^2/48. \tag{7}$$

Assume that N is large enough (depending on δ and λ) to ensure that the following inequalities hold:

$$e^{-\theta N\eta}\sum_{D\geq 1} e^{-\theta D} \leq \delta; \qquad \theta N\eta \geq 1/6. \tag{8}$$

Let us introduce the event

$$\bar{\Omega} = \left[\bigcap_{D<N\eta}\left\{\chi_D \leq \sqrt{D} + \sqrt{2\theta\left(D+N\eta\right)}\right\}\right] \bigcap \left[\bigcap_{N\eta\leq D}\left\{\chi_D \geq \sqrt{D-1} - \sqrt{2\theta\left(D+N\eta\right)}\right\}\right].$$

Using either (3) if $D < N\eta$ or (4) if $D \geq N\eta$, we get by (8)

$$\mathbb{P}\left(\bar{\Omega}^c\right) \leq \sum_{D\geq 1} e^{-\theta(D+N\eta)} \leq \delta.$$

Moreover, on $\overline{\Omega}$, $\chi_D^2 \leq \left(1+2\sqrt{\theta}\right)^2 N\eta$, for all D such that $D < N\eta$ and, by (7) and (8), $\chi_N \geq \sqrt{N-1} - \sqrt{3\theta N}$ and $\theta N > 1/3$. Therefore $\chi_N^2 \geq N\left(1-2\sqrt{3\theta}\right)$. Hence, on $\bar{\Omega}$, (6) and (7) yield

$$\begin{aligned}\varepsilon^{-2}[\mathrm{crit}\left(D\right) - \mathrm{crit}\left(N\right)] &\geq N\left(1-2\sqrt{3\theta}\right) - \left(1+2\sqrt{\theta}\right)^2 N\eta - \lambda N \\ &> (1-\frac{\eta}{2})N - \frac{3}{2}\eta N - (1-2\eta)N \quad = \quad 0,\end{aligned}$$

for all D such that $D < N\eta$. This immediately implies that $\hat{D}$ cannot be smaller than $N\eta$ on $\overline{\Omega}$ and therefore,

$$\mathbb{P}\left\{\hat{D} \geq N\eta\right\} \geq \mathbb{P}\left(\bar{\Omega}\right) \geq 1-\delta. \tag{9}$$

Moreover, on the same set $\bar{\Omega}$, $\chi_D \geq \sqrt{D-1} - \sqrt{2\theta\left(D+N\eta\right)}$ if D is such that $D \geq N\eta$. Noticing that $N\eta > 32$ and recalling that $\eta \leq 1/2$, we derive that on the set $\overline{\Omega}$ if D is such that $D \geq N\eta$

$$\chi_D \geq \sqrt{N\eta}\left(\sqrt{1-\frac{1}{32}} - \frac{1}{8}\right) > \sqrt{\frac{\eta N}{2}}.$$

Hence, on $\overline{\Omega}$, $\hat{D} \geq N\eta$ and $\chi_D \geq \sqrt{(\eta N)/2}$ for all D such that $D \geq N\eta$ and therefore $\chi_{\hat{D}} \geq \sqrt{(\eta N)/2}$. Finally,

$$\mathbb{E}\left[\left\|\hat{f}_{\hat{D}}\right\|^2\right] = \varepsilon^2 \mathbb{E}\left[\chi^2_{\hat{D}}\right] \geq \varepsilon^2 \frac{\eta}{2} N \mathbb{P}\left\{\chi_{\hat{D}} \geq \sqrt{\frac{\eta N}{2}}\right\} \geq \varepsilon^2 \frac{\eta N}{2} \mathbb{P}\left(\bar{\Omega}\right),$$

which, together with (7) and (9) yields

$$\mathbb{E}\left[\left\|\hat{f}_{\hat{D}}\right\|^2\right] \geq \frac{(1-\delta)(1-\lambda)}{4} N\varepsilon^2.$$

This result on the penalized selection procedure among the collection $\left\{\hat{f}_D, 1 \leq D \leq N\right\}$ tells us that if $\lambda < 1$, the penalized procedure with penalty $\text{pen}(D) = \lambda D\varepsilon^2$ has a tendency to select a high dimensional model and that the risk of the selected estimator $\hat{f}_{\hat{D}}$ is order $N\varepsilon^2$ (up to some multiplicative constant which tends to 0 when λ is close to 1). This means that roughly speaking the selected estimator behaves like the worst estimator of the collection $\hat{f}_N$. Conversely, if $\lambda > 1$, the right tail Gaussian concentration inequality can be used to prove an oracle inequality (see Birgé and Massart, 2001). The proof being in some sense more standard we think that it is useless to recall it here. This oracle inequality can be stated as follows. For some constant $C(\lambda)$ depending only on λ, whatever f

$$\mathbb{E}\left[\left\|\hat{f}_{\hat{D}} - f\right\|^2\right] \leq C(\lambda)\left(\inf_{D \leq N} \mathbb{E}\left[\left\|\hat{f}_D - f\right\|^2\right]\right).$$

In the case where $f = 0$, the best estimator among the collection $\left\{\hat{f}_D, D \leq N\right\}$ is of course $\hat{f}_1$ and the oracle inequality above tells us that when $\lambda > 1$, the penalized estimator $\hat{f}_{\hat{D}}$ has a quadratic risk which is of order of the quadratic risk of $\hat{f}_1$ (i.e. ε^2), up to the multiplicative constant $C(\lambda)$ (which of course tends to infinity when λ is close to 1). Combining this with Proposition 1 shows that the value $\lambda = 1$ is indeed critical since the penalized estimator behaves like $\hat{f}_N$ which is the worst estimator $\hat{f}_N$ when λ is below the critical value and like the best estimator $\hat{f}_1$ when λ is above the critical value. Note that Proposition 1 is asymptotic in the sense that N has to be large enough. The practical consequence of this asymptotic nature of this critical phenomenon is that if you run simulations, you will systematically see that some phase transition occurs that some critical value $\hat{\lambda}_{\min}$ but this value is not necessarily close to 1. Nevertheless the main point is that the phase transition does occur and more importantly, the final estimator that you intend

to choose, which the penalized estimator with penalty $\text{pen}(D) = 2\hat{\lambda}_{\min} D\varepsilon^2$ has a good behavior (even better than the penalized estimator with penalty corresponding to the Mallows C_p proposal, i.e. with $\text{pen}(D) = 2\hat{\lambda}_{\min} D\varepsilon^2$).

3 Bandwidth Selection for Kernel Density Estimation

This section is devoted to another instance of estimator selection, namely kernel estimator selection. Kernel estimators are natural procedures when we consider the problem of density estimation. So, in the sequel, we consider n i.i.d. observed random variables $X_1, \ldots, X_n$ belonging to $\mathbb{R}$ with unknown density f. Given $\mathcal{H}$ a set of positive bandwidths, for any $h \in \mathcal{H}$, we denote $\hat{f}_h$ the classical kernel rule:

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i),$$

with $K_h(\cdot) = \frac{1}{h} K\left(\frac{\cdot}{h}\right)$ and K a fixed integrable bounded kernel. As in Section 2, selecting a kernel estimate $\hat{f}_h$ is equivalent to selecting a parameter, namely here a bandwidth. A huge amount of literature has been devoted to this problem (see Devroye and Lugosi, 2001; Donoho et al. 1996; Goldenshluger and Lepsli, 2014, 2011; Massart, 2007, Reynaud-Bouret et al., 2011; Rigollet, 2006; Silverman, 1986 and references therein) but we point out adaptive minimax approaches proposed by Lepskii (1991) and its recent variation proposed by Goldenshluger and Lepski (2013). As detailed in subsequent Section 3.1, these methods are based on pair by pair comparisons of kernel rules. They enjoy optimal theoretical properties but suffer from tuning issues and high computation costs in particular in the multivariate setting (see Bertin et al., 2016; Doumic et al., 2012). In the sequel, we propose a new penalization methodology for bandwidth selection where concentration inequalities, in particular for U-statistics, play a key role (see Section 5). We first give heuristic arguments, then theoretical results are provided in the oracle setting. In particular, minimal and optimal penalties are derived under mild assumptions. Finally, we generalize our results to the multivariate setting and we derive rates of convergence of our procedure on Nikol'skii classes.

In the sequel, we still denote $\|\cdot\|$ the classical $\mathbb{L}_2$-norm and $\langle \cdot, \cdot \rangle$ the associated scalar product. For any $p \in [1, +\infty]$, we denote $\|\cdot\|_p$ the classical $\mathbb{L}_p$-norm. We also set

$$f_h = \mathbb{E}[\hat{f}_h] = K_h \star f,$$

where $\star$ denotes the standard convolution product.

3.1. A New Selection Method: Heuristics and Definition. Recall that our aim is to find a data-driven method to select the best bandwidth, i.e. a bandwidth h such that the risk $\mathbb{E}\|f - \hat{f}_h\|^2$ is minimum. Starting from the classical bias-variance decomposition

$$\mathbb{E}\|f - \hat{f}_h\|^2 = \|f - f_h\|^2 + \mathbb{E}\|f_h - \hat{f}_h\|^2 =: b_h + v_h,$$

it is natural to consider a criterion of the form

$$\mathrm{Crit}(h) = \hat{b}_h + \hat{v}_h,$$

where $\hat{b}_h$ is an estimator of the bias b_h and $\hat{v}_h$ an estimator of the variance v_h. Minimizing such a criterion is hopefully equivalent to minimizing the risk. Using that v_h is (tightly) bounded by $\|K\|^2/(nh)$, we naturally set $\hat{v}_h = \lambda\|K\|^2/(nh)$, with λ some tuning parameter. The difficulty lies in estimating the bias. Here we assume that $h_{\min}$, the minimum of the bandwidths grid, is very small. In this case $f_{h_{\min}} = K_{h_{\min}} \star f$ is a good approximation of f, so that $\|f_{h_{\min}} - f_h\|^2$ is close to b_h. This is tempting to estimate this term by $\|\hat{f}_{h_{\min}} - \hat{f}_h\|^2$ but doing this introduces a bias. Indeed, since

$$\hat{f}_{h_{\min}} - \hat{f}_h = (\hat{f}_{h_{\min}} - f_{h_{\min}} - \hat{f}_h + f_h) + (f_{h_{\min}} - f_h)$$

we have the decomposition

$$\mathbb{E}\|\hat{f}_{h_{\min}} - \hat{f}_h\|^2 = \|f_{h_{\min}} - f_h\|^2 + \mathbb{E}\|\hat{f}_{h_{\min}} - \hat{f}_h - f_{h_{\min}} + f_h\|^2. \tag{10}$$

But the centered variable $\hat{f}_{h_{\min}} - \hat{f}_h - f_{h_{\min}} + f_h$ can be written

$$\hat{f}_{h_{\min}} - \hat{f}_h - f_{h_{\min}} + f_h = \frac{1}{n}\sum_{i=1}^{n}(K_{h_{\min}} - K_h)(. - X_i) - \mathbb{E}((K_{h_{\min}} - K_h)(. - X_i)).$$

So, the second term in the right hand side of (10) is of order $n^{-1}\int (K_{h_{\min}}(x) - K_h(x))^2 dx$. Hence

$$\mathbb{E}\|\hat{f}_{h_{\min}} - \hat{f}_h\|^2 \approx \|f_{h_{\min}} - f_h\|^2 + \frac{\|K_{h_{\min}} - K_h\|^2}{n}$$

and then

$$b_h \approx \|f_{h_{\min}} - f_h\|^2 \approx \|\hat{f}_{h_{\min}} - \hat{f}_h\|^2 - \frac{\|K_{h_{\min}} - K_h\|^2}{n}.$$

These heuristic arguments lead to the following criterion to be minimized:

$$\mathrm{Crit}(h) = \|\hat{f}_{h_{\min}} - \hat{f}_h\|^2 - \frac{\|K_{h_{\min}} - K_h\|^2}{n} + \lambda\frac{\|K_h\|^2}{n}. \tag{11}$$

Thus, our method consists in comparing every estimator of our collection to the overfitting one, namely $\hat{f}_{h_{\min}}$, before adding the penalty term

$$\mathrm{pen}_\lambda(h) = \frac{\lambda\|K_h\|^2 - \|K_{h_{\min}} - K_h\|^2}{n}.$$

We call it *Penalized Comparison to Overfitting*, abbreviated PCO in the sequel.

Let us now compare this new method to existing ones. Actually, since $\hat{f}_{h_{\min}}$ is close to $n^{-1}\sum_{i=1}^n \delta_{X_i}$, our criterion is very close to the classical penalized least squares contrast which is used in regression context. More precisely, if $h_{\min} \to 0$, $\langle \hat{f}_{h_{\min}}, \hat{f}_h \rangle \to n^{-1}\sum_{i=1}^n \hat{f}_h(X_i)$ and then, using (11),

$$\begin{aligned}\mathrm{Crit}(h) &\approx \|\hat{f}_{h_{\min}}\|^2 + \|\hat{f}_h\|^2 - \frac{2}{n}\sum_{i=1}^n \hat{f}_h(X_i) + \mathrm{pen}_\lambda(h)\\ &\approx \|\hat{f}_{h_{\min}}\|^2 + L_n(\hat{f}_h) + \mathrm{pen}_\lambda(h)\end{aligned}$$

where $L_n(g) = \|g\|^2 - 2P_n(g)$. Since the first term does not play any role in the minimization, this is the criterion studied by Lerasle et al. (2015).

The classical Lespki's method (Lepskii, 1990, 1991) consists in selecting a bandwidth $\hat{h}$ by using the rule

$$\hat{h} = \max\left\{h \in \mathcal{H} : \quad \|\hat{f}_{h'} - \hat{f}_h\|^2 \le V_n(h') \ \text{ for any } h' \in \mathcal{H} \text{ s.t. } h' \le h\right\},$$

for some well-chosen bandwidth-dependent sequence $V_n(\cdot)$. Introduced in Goldenshluger and Lepski (2008), the Goldenshluger-Lepski's methodology is a variation of the Lepski's procedure still based on pair-by-pair comparisons between estimators. More precisely, Goldenshluger and Lepski suggest to use the selection rule

$$\hat{h} = \arg\min_{h\in\mathcal{H}} \{A(h) + V_2(h)\},$$

with

$$A(h) = \sup_{h'\in\mathcal{H}} \left\{\|\hat{f}_{h'} - \hat{f}_{h\vee h'}\|^2 - V_1(h')\right\}_+,$$

where x_+ denotes the positive part $\max(x, 0)$, $h \vee h' = \max(h, h')$ and $V_1(\cdot)$ and $V_2(\cdot)$ are penalties to be suitably chosen (Goldenshluger and Lepski essentially consider $V_2 = V_1$ or $V_2 = 2V_1$ in Goldenshluger and Lepski 2008, 2009, 2011, 2013). The authors establish the minimax optimality of their method when V_1 and V_2 are large enough. However, observe that if $V_1 = 0$, then, under mild assumptions,

$$A(h) = \sup_{h'\in\mathcal{H}} \|\hat{f}_{h'} - \hat{f}_{h\vee h'}\|^2 \approx \|\hat{f}_{h_{\min}} - \hat{f}_h\|^2$$

so that our method turns out to be exactly some degenerate case of the Goldenshluger-Lespki's method. We study its performances in the oracle setting.

3.2. Oracle Inequality. As explained in Section 3.1, we study the performances of $\hat{f} := \hat{f}_{\hat{h}}$ with

$$\hat{h} = \arg\min_{h\in\mathcal{H}} \left\{ \|\hat{f}_h - \hat{f}_{h_{\min}}\|^2 + \mathrm{pen}_\lambda(h) \right\},$$

where $h_{min} = \min \mathcal{H}$ and $\mathrm{pen}_\lambda(h)$ is the penalty suggested by heuristic arguments of Section 3.1. We assume that $\max \mathcal{H}$ is smaller than an absolute constant. We have the following result.

THEOREM 2. *Assume that K is symmetric and $\int K(u)du = 1$. Assume also that $h_{\min} \geq \|K\|_\infty \|K\|_1/n$ and $\|f\|_\infty < \infty$. Let $x \geq 1$ and $\varepsilon \in (0,1)$. If*

$$\mathrm{pen}_\lambda(h) = \frac{\lambda\|K_h\|^2 - \|K_{h_{\min}} - K_h\|^2}{n}, \qquad \textit{with } \lambda > 0,$$

then, with probability larger than $1 - C_1|\mathcal{H}|e^{-x}$,

$$\begin{aligned}\|\hat{f}_{\hat{h}} - f\|^2 \;\leq\; & C_0(\varepsilon) \min_{h\in\mathcal{H}} \|\hat{f}_h - f\|^2 \\ & + C_2(\varepsilon,\lambda)\|f_{h_{\min}} - f\|^2 + C_3(\varepsilon, K, \lambda)\left(\frac{\|f\|_\infty x^2}{n} + \frac{x^3}{n^2 h_{\min}}\right),\end{aligned}$$

where C_1 is an absolute constant and $C_0(\varepsilon) = \lambda + \varepsilon$ if $\lambda \geq 1$, $C_0(\varepsilon) = 1/\lambda + \varepsilon$ if $0 < \lambda < 1$. The constant $C_2(\varepsilon,\lambda)$ only depends on ε and λ and $C_3(\varepsilon, K, \lambda)$ only depends on ε, K and λ.

The proof of Theorem 2 can be found in Section 5.2 which shows that $C_2(\varepsilon,\lambda)$ and $C_3(\varepsilon, K, \lambda)$ blow up when ε goes to 0. We thus have established an oracle inequality, provided that the tuning parameter λ is positive. The two last terms are remainder terms and are negligible under mild assumptions, as proved in the minimax setting in Corollary 7.

Note that when the tuning parameter λ is fixed by taking $\lambda = 1$, the leading constant $C_0(\varepsilon)$ is minimum and is equal to $1 + \varepsilon$, so can be as close to 1 as desired, and in this case the penalty is

$$\mathrm{pen}_{\lambda=1}(h) = \frac{\|K_h\|^2 - \|K_{h_{\min}} - K_h\|^2}{n} = \frac{2\langle K_h, K_{h_{\min}}\rangle - \|K_{h_{\min}}\|^2}{n}.$$

Since the last term does not depend on h, it can be omitted and we obtain an optimal penalty by taking

$$\mathrm{pen}_{opt}(h) := \frac{2\langle K_h, K_{h_{\min}}\rangle}{n}.$$

The procedure associated with pen_{opt} (or associated with any penalty pen such that $\text{pen}(h) - \text{pen}_{opt}(h)$ does not depend on h) is then optimal in the oracle approach. For density estimation, leading constants for oracle inequalities achieved by Goldenshluger and Lepski's procedure are $1 + \sqrt{2}$ (if $\|K\|_1 = 1$) in Lacour and Massart (2016) and $1 + 3\|K\|_1$ in Goldenshluger and Lepski (2011). To our kwnoledge, if the Goldenshluger and Lepski's procedure can achieve leading oracle constants of the form $1 + \varepsilon$, for any $\varepsilon > 0$, remains an open question.

3.3. Minimal Penalty. In this section, we study the behavior of selected bandwidths when λ, the tuning parameter of our procedure, is too small. We assume that the bias $\|f_{h_{\min}} - f\|^2$ is negligible with respect to the integrated variance, which is equivalent to

$$nh_{\min}\|f_{h_{\min}} - f\|^2 = o(1). \tag{12}$$

The following result is proved in Section 5.3.

THEOREM 3. *Assume that K is symmetric and $\int K(u)du = 1$. Assume also that $\|f\|_\infty < \infty$ and $h_{\min}$ satisfies (12) and*

$$\frac{\|K\|_\infty \|K\|_1}{n} \le h_{\min} \le \frac{(\log n)^\beta}{n}$$

for some real β. If

$$\text{pen}_\lambda(h) = \frac{\lambda\|K_h\|^2 - \|K_{h_{\min}} - K_h\|^2}{n}, \qquad \textit{with } \lambda < 0,$$

then, for n large enough, with probability larger than $1 - C_1|\mathcal{H}|\exp(-(n/\log n)^{1/3})$,

$$\hat{h} \le C(\lambda)h_{\min} \le C(\lambda)\frac{(\log n)^\beta}{n}$$

where C_1 is an absolute constant and $C(\lambda) = 2.1 - 1/\lambda$.

Thus, if the penalty is too small (i.e. $\lambda < 0$), our method selects $\hat{h}$ close to $h_{\min}$ with high probability. This leads to an overfitting estimator. In the same spirit as in Section 2.4, we derive a minimal penalty, which is (up to additive constants)

$$\text{pen}_{min}(h) := \frac{2\langle K_h, K_{h_{\min}}\rangle - \|K_h\|^2}{n}.$$

The interest of this result is not purely theoretical since it can be used to provide a data-driven choice of the tuning parameter λ (see the discussion in Section 2.4).

3.4. The Multivariate Case and Adaptive Minimax Rates. We now deal with the multivariate setting and we consider n i.i.d. observed random vectors $X_1, \dots, X_n$ belonging to $\mathbb{R}^d$ with unknown density f. Given $\mathcal{H} = \mathcal{H}_1 \times \dots \mathcal{H}_d$ a set of multivariate positive bandwidths, for $h = (h_1, \dots, h_d) \in \mathcal{H}$, we now denote $\hat{f}_h$ the kernel rule:

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i),$$

with for any $x = (x_1, \dots, x_d) \in \mathbb{R}^d$,

$$K_h(x) = \frac{1}{h_1 \dots h_d} K\left(\frac{x_1}{h_1}, \dots, \frac{x_d}{h_d}\right).$$

We consider $h_{\min} = (h_{1,min}, \dots, h_{d,\min})$ with $h_{j,\min} = \min \mathcal{H}_j$ for any $j = 1, \dots, d$ and we study $\hat{f} := \hat{f}_{\hat{h}}$ with

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \left\{ \|\hat{f}_h - \hat{f}_{h_{\min}}\|^2 + \mathrm{pen}_\lambda(h) \right\}$$

and $\mathrm{pen}_\lambda(h)$ is the penalty similar to the one chosen in the univariate setting. We still assume that for $j = 1, \dots, d$, $\max \mathcal{H}_j$ is smaller than an absolute constant. We obtain in the multivariate setting, results analog to Theorems 2 and 3.

THEOREM 4. *Assume that K is symmetric and $\int K(u)du = 1$. Assume also that $\|f\|_\infty < \infty$ and $h_{\min}$ satisfies*

$$n \prod_{j=1}^{d} h_{j,\min} \times \|f_{h_{\min}} - f\|^2 = o(1)$$

and

$$\frac{\|K\|_\infty \|K\|_1}{n} \le \prod_{j=1}^{d} h_{j,\min} \le \frac{(\log n)^\beta}{n}$$

for some real β. If

$$\mathrm{pen}_\lambda(h) = \frac{\lambda \|K_h\|^2 - \|K_{h_{\min}} - K_h\|^2}{n}, \qquad \text{with } \lambda < 0,$$

then, for n large enough, with probability larger than $1 - C_1|\mathcal{H}| \exp(-(n/\log n)^{1/3})$,

$$\prod_{j=1}^{d} \hat{h}_j \le C(\lambda) \prod_{j=1}^{d} h_{j,\min} \le C(\lambda) \frac{(\log n)^\beta}{n}$$

where C_1 is an absolute constant and $C(\lambda) = 2.1 - 1/\lambda$.

As previously, if the penalty is too small, with high probability, each component of the selected bandwidth is (up to constants) the minimum one in each direction. Let us state the analog of Theorem 2.

THEOREM 5. *Assume that K is symmetric and $\int K(u)du = 1$. Assume also that $\prod_{j=1}^{d} h_{j,\min} \geq \|K\|_\infty \|K\|_1/n$ and $\|f\|_\infty < \infty$. Let $x \geq 1$ and $\varepsilon \in (0,1)$. If*

$$\mathrm{pen}_\lambda(h) = \frac{\lambda\|K_h\|^2 - \|K_{h_{\min}} - K_h\|^2}{n}, \qquad \text{with } \lambda > 0,$$

then, with probability larger than $1 - C_1|\mathcal{H}|e^{-x}$,

$$\begin{aligned}\|\hat{f}_{\hat{h}} - f\|^2 \quad \leq \quad & C_0(\varepsilon) \min_{h\in\mathcal{H}} \|\hat{f}_h - f\|^2 \\ & + C_2(\varepsilon,\lambda)\|f_{h_{\min}} - f\|^2 \\ & + C_3(\varepsilon,K,\lambda)\left(\frac{\|f\|_\infty x^2}{n} + \frac{x^3}{n^2\prod_{j=1}^{d} h_{j,\min}}\right),\end{aligned}$$

where C_1 is an absolute constant and $C_0(\varepsilon) = \lambda + \varepsilon$ if $\lambda \geq 1$, $C_0(\varepsilon) = 1/\lambda + \varepsilon$ if $0 < \lambda < 1$. The constant $C_2(\varepsilon,\lambda)$ only depends on ε and λ and $C_3(\varepsilon,K,\lambda)$ only depends on ε, K and λ.

From the previous result, classical adaptive minimax anisotropic rates of convergence can be obtained. To be more specific, let us consider anisotropic Nikol'skii classes (see Nikol'skii, 1977 or Goldenshluger and Lepski, 2014 for a clear exposition). For this purpose, let $(e_1,\ldots,e_d)$ denote the canonical basis of $\mathbb{R}^d$. For any function $g : \mathbb{R}^d \mapsto \mathbb{R}$ and any $u \in \mathbb{R}$, we define the first order difference operator with step size u in the j-th direction by

$$\Delta_{u,j} g(x) = g(x + ue_j) - g(x), \quad j = 1,\ldots,d.$$

By induction, the k-th order difference operator with step size u in the j-th direction is defined as

$$\Delta_{u,j}^k g(x) = \Delta_{u,j}\Delta_{u,j}^{k-1} g(x) = \sum_{\ell=1}^{k} (-1)^{\ell+k} \binom{k}{\ell} \Delta_{u\ell,j} g(x).$$

We then set

DEFINITION 6. *For any given vectors $\boldsymbol{r} = (r_1,\ldots,r_d)$, $r_j \in [1,+\infty]$, $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_d)$, $\beta_j > 0$, and $\boldsymbol{L} = (L_1,\ldots,L_d)$, $L_j > 0$, $j = 1,\ldots,d$, we say that the function $g : \mathbb{R}^d \mapsto \mathbb{R}$ belongs to the anisotropic Nikol'skii class $\mathcal{N}_{\boldsymbol{r},d}(\boldsymbol{\beta},\boldsymbol{L})$ if*

(i) $\|g\|_{r_j} \leq L_j$ for all $j = 1,\ldots,d$

(ii) for every $j = 1, \dots, d$, there exists a natural number $k_j > \beta_j$ such that

$$\|\Delta_{u,j}^{k_j} g\|_{r_j} \leq L_j |u_j|^{\beta_j}, \quad \forall u \in \mathbb{R}^d, \quad \forall j = 1, \dots, d.$$

Note that the anisotropic Nikol'skii class is a specific class of the anisotropic Besov class (Kerkyacharian et al., 2008; Nikol'skii, 1977) :

$$\mathcal{N}_{\mathbf{r},d}(\boldsymbol{\beta}, .) = \mathcal{B}^{\boldsymbol{\beta}}_{\mathbf{r},\infty}(\cdot).$$

We consider the construction of the classical kernel K proposed in Section 3.2 of Goldenshluger and Lepski (2014) such that assumptions of Theorem 5 are satisfied and

$$\int K(u) u^k du = 0, \quad \forall |k| = 1, \dots, \ell - 1,$$

where $k = (k_1, \dots, k_d)$ is the multi-index, $k_i \geq 0$, $|k| = k_1 + \cdots + k_d$ and $u^k = u_1^{k_1} \cdots u_d^{k_d}$ for $u = (u_1, \dots, u_d)$. In this case, Lemma 3 of Goldenshluger and Lepski (2014) states that if $f \in \mathcal{N}_{\mathbf{2},d}(\boldsymbol{\beta}, \mathbf{L})$ with $\ell > \max_{j=1,\dots,d} \beta_j$ then

$$f_h - f = \sum_{j=1}^{d} B_{j,h},$$

with the B_j's satisfying

$$\|B_{j,h}\| \leq M L_j h_j^{\beta_j}, \tag{13}$$

for M a positive constant depending on K and $\boldsymbol{\beta}$. Finally, we consider $\mathcal{H}$ the following set of bandwidths:

$$\mathcal{H} = \left\{ h = (h_1, \dots, h_d) : \quad \frac{\|K\|_\infty \|K\|_1}{n} \leq \prod_{j=1}^{d} h_j, \right.$$
$$\left. h_j \leq 1 \text{ and } h_j^{-1} \text{ is an integer } \forall\, j = 1, \dots, d \right\}.$$

Note that $|\mathcal{H}| \leq \tilde{C}_1(K, d) n^d$ where $\tilde{C}_1(K, d)$ only depends on K and d. By combining Theorem 5 and Proposition 8, standard computations lead to the following corollary (see Section 5.4).

Corollary 7. *Consider the previous construction of the kernel K (depending on a given integer ℓ) and the previous specific set of bandwidths $\mathcal{H}$. For any $h \in \mathcal{H}$, we set*

$$\mathrm{pen}_\lambda(h) = \frac{\lambda \|K_h\|^2 - \|K_{h_{\min}} - K_h\|^2}{n}, \qquad \textit{with } \lambda > 0.$$

For given $\boldsymbol{L} = (L_1, \ldots, L_d)$ *and* $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)$ *such that* $\ell > \max_{j=1,\ldots,d} \beta_j$ *and for* $B > 0$, *we denote* $\tilde{\mathcal{N}}_{\mathbf{2},d}(\boldsymbol{\beta}, \boldsymbol{L}, B)$ *the set of densities bounded by* B *and belonging to* $\mathcal{N}_{\mathbf{2},d}(\boldsymbol{\beta}, \boldsymbol{L})$. *Then,*

$$\sup_{f \in \tilde{\mathcal{N}}_{\mathbf{2},d}(\boldsymbol{\beta},\boldsymbol{L},B)} \mathbb{E}\left[\|\hat{f}_{\hat{h}} - f\|^2\right] \leq \tilde{C}_2(\boldsymbol{\beta}, K, B, d, \lambda) \left(\prod_{j=1}^{d} L_j^{\frac{1}{\beta_j}}\right)^{\frac{2\bar{\beta}}{2\bar{\beta}+1}} n^{-\frac{2\bar{\beta}}{2\bar{\beta}+1}},$$

where $\tilde{C}_2(\boldsymbol{\beta}, K, B, d, \lambda)$ *is a constant only depending on* $\boldsymbol{\beta}$, K, B, d *and* λ *and*

$$\frac{1}{\bar{\beta}} = \sum_{j=1}^{d} \frac{1}{\beta_j}.$$

Theorem 3 of Goldenshluger and Lepski (2014) states that up to the constant $\tilde{C}_2(\boldsymbol{\beta}, K, B, d, \lambda)$, we cannot improve the rate achieved by our procedure. So, the latter achieves the adaptive minimax rate over the class $\tilde{\mathcal{N}}_{\mathbf{2},d}(\boldsymbol{\beta}, \mathbf{L}, B)$. These minimax adaptive properties can be extended to the case of multivariate Hölder spaces if we restrict estimation to compactly supported densities (see arguments of Bertin et al., 2016) and, in the univariate case, to the class of Sobolev spaces.

4 Conclusion

To conclude, our method shares all the optimal theoretical properties of Lepski's and Goldenshluger-Lepski's methodologies in oracle and minimax approaches but its computational cost is much slower. Indeed, for any $h \in \mathcal{H}$, we only need to compare $\hat{f}_h$ with respect to $\hat{f}_{h_{\min}}$, and not to all $\hat{f}_{h'}$ for $h' \in \mathcal{H}$ (or at least for a large subset of $\mathcal{H}$). It is a crucial point in the case where $|\mathcal{H}|$ is of order n^d, as previously in the d-dimensional case. And last but not least, calibration issues of our methodology are also addressed with clear identifications of minimal and optimal penalties with simple relationships between them: for $\mathrm{Crit}(h) = \|\hat{f}_h - \hat{f}_{h_{\min}}\|^2 + \mathrm{pen}_\lambda(h)$, we have

$$\mathrm{pen}_{min}(h) = \frac{2\langle K_h, K_{h_{\min}}\rangle - \|K_h\|^2}{n} \quad \text{and} \quad \mathrm{pen}_{opt}(h) = \frac{2\langle K_h, K_{h_{\min}}\rangle}{n}.$$

Numerical performances of PCO will be studied in details in a further work. Extending our method to other loss functions than the quadratic loss if of course very tempting. Some preliminary numerical experiments with the Hellinger loss seem to indicate that it could work event though some theory remains to be built. In the same spirit one can wonder whether the PCO

could be used for more general selection estimator selection issues than bandwidth selection for kernel density estimation.

5 Proofs of Section 3

In this section, we denote for any function g, for any t,

$$\tilde{g}(t) = g(-t).$$

The notation $\square$ denotes an absolute constant that may change from line to line. The proofs use the following lower bound (15) established in Proposition 4.1 of Lerasle et al. (2015) combined with their Proposition 3.3.

PROPOSITION 8. *Assume that K is symmetric and $\int K(u)du = 1$. Assume also that $h_{\min} \geq \|K\|_\infty \|K\|_1/n$. Let $\Upsilon \geq (1 + 2\|f\|_\infty \|K\|_1^2)\|K\|_\infty / \|K\|^2$. For all $x \geq 1$ and for all $\eta \in (0,1)$, with probability larger than $1 - \square|\mathcal{H}|e^{-x}$, for all $h \in \mathcal{H}$, each of the following inequalities hold*

$$\|f - \hat{f}_h\|^2 \leq (1+\eta)\left(\|f - f_h\|^2 + \frac{\|K_h\|^2}{n}\right) + \square\frac{\Upsilon x^2}{\eta^3 n}, \tag{14}$$

$$\|f - f_h\|^2 + \frac{\|K_h\|^2}{n} \leq (1+\eta)\|f - \hat{f}_h\|^2 + \square\frac{\Upsilon x^2}{\eta^3 n}. \tag{15}$$

We first give a general result for the study of $\hat{f} := \hat{f}_{\hat{h}}$.

THEOREM 9. *Assume that K is symmetric and $\int K(u)du = 1$. Assume also that $h_{\min} \geq \|K\|_\infty \|K\|_1/n$ and $\|f\|_\infty < \infty$. Let $x \geq 1$ and $\theta \in (0,1)$. With probability larger than $1 - C_1|\mathcal{H}|\exp(-x)$, for any $h \in \mathcal{H}$,*

$$\begin{aligned}(1-\theta)\|\hat{f}_{\hat{h}} - f\|^2 \leq\ & (1+\theta)\|\hat{f}_h - f\|^2 + \left(\mathrm{pen}_\lambda(h) - 2\frac{\langle K_h, K_{h_{\min}}\rangle}{n}\right) \\ & - \left(\mathrm{pen}_\lambda(\hat{h}) - 2\frac{\langle K_{\hat{h}}, K_{h_{\min}}\rangle}{n}\right) \\ & + \frac{C_2}{\theta}\|f_{h_{\min}} - f\|^2 + \frac{C(K)}{\theta}\left(\frac{\|f\|_\infty x^2}{n} + \frac{x^3}{n^2 h_{\min}}\right)\end{aligned}$$

where C_1 and C_2 are absolute constants and $C(K)$ only depends on K.

We prove in the sequel Theorem 9, then Theorems 2 and 3. Theorems 4 and 5 are proved similarly by replacing h by $\prod_{j=1}^d h_j$ and $h_{\min}$ by $\prod_{j=1}^d h_{j,\min}$, so we omit their proof. Section 5.4 gives the proof of Corollary 7.

5.1. Proof of Theorem 9. Let $\theta' \in (0,1)$ be fixed and chosen later. Using the definition of $\hat{h}$, we can write, for any $h \in \mathcal{H}$,

$$\begin{aligned}
\|\hat{f}_{\hat{h}} - f\|^2 + \mathrm{pen}_\lambda(\hat{h}) &= \|\hat{f}_{\hat{h}} - \hat{f}_{h_{\min}}\|^2 + \mathrm{pen}_\lambda(\hat{h}) + \|\hat{f}_{h_{\min}} - f\|^2 \\
&\quad +2\langle \hat{f}_{\hat{h}} - \hat{f}_{h_{\min}}, \hat{f}_{h_{\min}} - f\rangle \\
&\leq \|\hat{f}_h - \hat{f}_{h_{\min}}\|^2 + \mathrm{pen}_\lambda(h) + \|\hat{f}_{h_{\min}} - f\|^2 \\
&\quad +2\langle \hat{f}_{\hat{h}} - \hat{f}_{h_{\min}}, \hat{f}_{h_{\min}} - f\rangle \\
&\leq \|\hat{f}_h - f\|^2 + 2\|f - \hat{f}_{h_{\min}}\|^2 + 2\langle \hat{f}_h - f, f - \hat{f}_{h_{\min}}\rangle \\
&\quad +\mathrm{pen}_\lambda(h) + 2\langle \hat{f}_{\hat{h}} - \hat{f}_{h_{\min}}, \hat{f}_{h_{\min}} - f\rangle.
\end{aligned}$$

Consequently,

$$\begin{aligned}
\|\hat{f}_{\hat{h}} - f\|^2 &\leq \|\hat{f}_h - f\|^2 + \Big(\mathrm{pen}_\lambda(h) - 2\langle \hat{f}_h - f, \hat{f}_{h_{\min}} - f\rangle\Big) \\
&\quad - \Big(\mathrm{pen}_\lambda(\hat{h}) - 2\langle \hat{f}_{\hat{h}} - f, \hat{f}_{h_{\min}} - f\rangle\Big). \qquad (16)
\end{aligned}$$

Then, for a given h, we study the term

$$2\langle \hat{f}_h - f, \hat{f}_{h_{\min}} - f\rangle$$

that can be viewed as an ideal penalty. Let us introduce the degenerate U-statistic

$$U(h, h_{\min}) = \sum_{i \neq j} \langle K_h(. - X_i) - f_h, K_{h_{\min}}(. - X_j) - f_{h_{\min}}\rangle$$

and the following centered variable

$$V(h, h') = < \hat{f}_h - f_h, f_{h'} - f > .$$

We first center the terms in the following way

$$\begin{aligned}
\langle \hat{f}_h - f, \hat{f}_{h_{\min}} - f\rangle &= \langle \hat{f}_h - f_h + f_h - f, \hat{f}_{h_{\min}} - f_{h_{\min}} + f_{h_{\min}} - f\rangle \\
&= \langle \hat{f}_h - f_h, \hat{f}_{h_{\min}} - f_{h_{\min}}\rangle + \langle \hat{f}_h - f_h, f_{h_{\min}} - f\rangle \\
&\quad +\langle f_h - f, \hat{f}_{h_{\min}} - f_{h_{\min}}\rangle + \langle f_h - f, f_{h_{\min}} - f\rangle \\
&= \langle \hat{f}_h - f_h, \hat{f}_{h_{\min}} - f_{h_{\min}}\rangle + V(h, h_{\min}) + V(h_{\min}, h) \\
&\quad +\langle f_h - f, f_{h_{\min}} - f\rangle.
\end{aligned}$$

Now,

$$\begin{aligned}\langle \hat{f}_h - f_h, \hat{f}_{h_{\min}} - f_{h_{\min}}\rangle &= \frac{1}{n^2}\sum_{i,j}\langle K_h(.-X_i) - f_h, K_{h_{\min}}(.-X_j) - f_{h_{\min}}\rangle \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\langle K_h(.-X_i) - f_h, K_{h_{\min}}(.-X_i) - f_{h_{\min}}\rangle \\ &\quad + \frac{U(h,h_{\min})}{n^2}.\end{aligned}$$

Then

$$\begin{aligned}\langle \hat{f}_h - f_h, \hat{f}_{h_{\min}} - f_{h_{\min}}\rangle &= \frac{\langle K_h, K_{h_{\min}}\rangle}{n} - \frac{1}{n}\langle \hat{f}_h, f_{h_{\min}}\rangle - \frac{1}{n}\langle f_h, \hat{f}_{h_{\min}}\rangle \\ &\quad + \frac{1}{n}\langle f_h, f_{h_{\min}}\rangle + \frac{U(h,h_{\min})}{n^2}.\end{aligned}$$

Finally, we have the following decomposition of $\langle \hat{f}_h - f, \hat{f}_{h_{\min}} - f\rangle$:

$$\begin{aligned}\langle \hat{f}_h - f, \hat{f}_{h_{\min}} - f\rangle &= \frac{\langle K_h, K_{h_{\min}}\rangle}{n} + \frac{U(h,h_{\min})}{n^2} && (17)\\ &\quad - \frac{1}{n}\langle \hat{f}_h, f_{h_{\min}}\rangle - \frac{1}{n}\langle f_h, \hat{f}_{h_{\min}}\rangle + \frac{1}{n}\langle f_h, f_{h_{\min}}\rangle && (18)\\ &\quad + V(h,h_{\min}) + V(h_{\min},h) + \langle f_h - f, f_{h_{\min}} - f\rangle. && (19)\end{aligned}$$

We first control the last term of the first line and we obtain the following lemma.

LEMMA 10. *With probability larger than* $1 - 5.54|\mathcal{H}|\exp(-x)$, *for any* $h \in \mathcal{H}$,

$$\frac{|U(h,h_{\min})|}{n^2} \le \theta' \frac{\|K\|^2}{nh} + \frac{\square \|K\|_1^2 \|f\|_\infty x^2}{\theta' n} + \frac{\square \|K\|_\infty \|K\|_1 x^3}{\theta' n^2 h_{\min}}$$

PROOF. We have:

$$\begin{aligned}U(h,h_{\min}) &= \sum_{i\neq j}\langle K_h(.-X_i) - f_h, K_{h_{\min}}(.-X_j) - f_{h_{\min}}\rangle \\ &= \sum_{i=2}^{n}\sum_{j<i} G_{h,h_{\min}}(X_i,X_j) + G_{h_{\min},h}(X_i,X_j),\end{aligned}$$

with

$$G_{h,h'}(s,t) = \langle K_h(.-s) - f_h, K_{h'}(.-t) - f_{h'}\rangle.$$

Therefore, we can apply Theorem 3.4 of Houdré and Reynaud-Bouret (2003):

$$\mathbb{P}\left(|U(h,h_{\min})| \geq \square\left(C\sqrt{x}+Dx+Bx^{3/2}+Ax^2\right)\right) \leq 5.54\exp(-x),$$

with A, B, C and D defined subsequently. Since

$$\|f_{h_{\min}}\|_\infty = \|K_{h_{\min}} \star f\|_\infty \leq \|K_{h_{\min}}\|_\infty = \frac{\|K\|_\infty}{h_{\min}},$$

we have

$$\begin{aligned} A &:= \|G_{h,h_{\min}} + G_{h_{\min},h}\|_\infty \\ &\leq \|G_{h,h_{\min}}\|_\infty + \|G_{h_{\min},h}\|_\infty \\ &\leq 4\|K_{h_{\min}}\|_\infty \times (\|K\|_1 + \|K_h \star f\|_1) \\ &\leq \frac{8\|K\|_\infty\|K\|_1}{h_{\min}} \end{aligned}$$

and

$$\frac{Ax^2}{n^2} \leq \frac{8x^2\|K\|_\infty\|K\|_1}{n^2 h_{\min}}.$$

We have

$$B^2 := (n-1)\sup_t \mathbb{E}[(G_{h,h_{\min}}(t,X_2) + G_{h_{\min},h}(t,X_2))^2].$$

and for any t,

$$\begin{aligned} \mathbb{E}[G^2_{h,h_{\min}}(t,X_2)] &= \mathbb{E}\left[\left(\int (K_h(u-t) - f_h(u))(K_{h_{\min}}(u-X_2)\right.\right. \\ &\qquad \left.\left. - \mathbb{E}[K_{h_{\min}}(u-X_2)])du\right)^2\right] \\ &\leq \mathbb{E}\left[\int (K_h(u-t) - f_h(u))^2 du \times \int (K_{h_{\min}}(u-X_2)\right. \\ &\qquad \left. - \mathbb{E}[K_{h_{\min}}(u-X_2)])^2 du\right] \\ &\leq 2 \times \int \left(K_h^2(u-t) + (K_h \star f)^2(u)\right) du \\ &\qquad \times \int \mathbb{E}[K^2_{h_{\min}}(u-X_2)]du \\ &\leq 4\|K_h\|^2 \times \|K_{h_{\min}}\|^2. \end{aligned}$$

Therefore,

$$\frac{Bx^{3/2}}{n^2} \leq \frac{\theta'}{3}\frac{\|K\|^2}{nh} + \frac{6}{\theta'}\frac{\|K\|^2 x^3}{n^2 h_{\min}}.$$

Now,

$$\begin{aligned}
C^2 &:= \sum_{i=2}^{n}\sum_{j=1}^{i-1}\mathbb{E}[(G_{h,h_{\min}}(X_i,X_j)+G_{h_{\min},h}(X_i,X_j))^2] \\
&\leq \square \times n^2\mathbb{E}[G^2_{h,h_{\min}}(X_1,X_2)] \\
&= \square \\
&\quad \times n^2\mathbb{E}\left[\left(\int (K_h(u-X_1)-f_h(u))(K_{h_{\min}}(u-X_2)-f_{h_{\min}}(u))du\right)^2\right] \\
&= \square \times n^2\mathbb{E}\left[\left(\int K_h(u-X_1)K_{h_{\min}}(u-X_2)du\right.\right. \\
&\qquad - \int K_h(u-X_1)(K_{h_{\min}}\star f)(u)du \\
&\qquad - \int K_{h_{\min}}(u-X_2)(K_h\star f)(u)du \\
&\qquad \left.\left. - + \int (K_h\star f)(u)(K_{h_{\min}}\star f)(u)du\right)^2\right].
\end{aligned}$$

Since

$$\|K_h\star f\|_\infty \leq \|K\|_1\|f\|_\infty, \quad \|K_h\star f\|_1 \leq \|K\|_1,$$

we have:

$$C^2 \leq \square \times n^2\mathbb{E}\left[\left(\int K_h(u-X_1)K_{h_{\min}}(u-X_2)du\right)^2\right] + \square\|K\|_1^4\|f\|_\infty^2 \times n^2.$$

So, we just have to deal with the first term.

$$\begin{aligned}
\mathbb{E}\left[\left(\int K_h(u-X_1)K_{h_{\min}}(u-X_2)du\right)^2\right] &= \mathbb{E}\left[\left((\tilde{K}_h\star K_{h_{\min}})(X_1-X_2)\right)^2\right] \\
&= \iint (\tilde{K}_h\star K_{h_{\min}})^2(u-v)f(u) \\
&\quad \times f(v)dudv \\
&\leq \|f\|_\infty\|\tilde{K}_h\star K_{h_{\min}}\|^2 \\
&\leq \|K\|_1^2\|f\|_\infty\|K_h\|^2.
\end{aligned}$$

Finally

$$C \leq \square \times n\|K\|_1\|f\|_\infty^{1/2}\|K_h\| + \square\|K\|_1^2\|f\|_\infty \times n.$$

So, since $x \geq 1$,

$$\frac{C\sqrt{x}}{n^2} \leq \frac{\theta'}{3}\frac{\|K\|^2}{nh} + \frac{\square\|K\|_1^2\|f\|_\infty\, x}{\theta' n}.$$

Now, let us consider

$$\mathcal{S} := \left\{ a=(a_i)_{2\leq i\leq n},\, b=(b_i)_{1\leq i\leq n-1} :\ \sum_{i=2}^{n}\mathbb{E}[a_i^2(X_i)] \leq 1,\ \sum_{i=1}^{n-1}\mathbb{E}[b_i^2(X_i)] \leq 1 \right\}.$$

We have

$$D := \sup_{(a,b)\in\mathcal{S}} \left\{ \sum_{i=2}^{n}\sum_{j=1}^{i-1}\mathbb{E}[(G_{h,h_{\min}}(X_i,X_j)+G_{h_{\min},h}(X_i,X_j))a_i(X_i)b_j(X_j)] \right\}.$$

We have for $(a,b)\in\mathcal{S}$,

$$\begin{aligned}
&\sum_{i=2}^{n}\sum_{j=1}^{i-1}\mathbb{E}[G_{h,h_{\min}}(X_i,X_j)a_i(X_i)b_j(X_j)] \\
\leq\ & \sum_{i=2}^{n}\sum_{j=1}^{i-1}\mathbb{E}\int |K_h(u-X_i)-(K_h\star f)(u)||a_i(X_i)| \\
&\qquad \times |K_{h_{\min}}(u-X_j)-(K_{h_{\min}}\star f)(u)||b_j(X_j)|du \\
\leq\ & \sum_{i=2}^{n}\sum_{j=1}^{n-1}\int \mathbb{E}\left[|K_h(u-X_i)-(K_h\star f)(u)||a_i(X_i)|\right] \\
&\qquad \times \mathbb{E}\left[|K_{h_{\min}}(u-X_j)-(K_{h_{\min}}\star f)(u)||b_j(X_j)|\right] du
\end{aligned}$$

and for any u,

$$\begin{aligned}
&\sum_{i=2}^{n}\mathbb{E}|K_h(u-X_i)-(K_h\star f)(u)||a_i(X_i)| \\
\leq\ & \sqrt{n}\sqrt{\sum_{i=2}^{n}\mathbb{E}^2\left[K_h(u-X_i)-(K_h\star f)(u)||a_i(X_i)|\right]} \\
\leq\ & \sqrt{n}\sqrt{\sum_{i=2}^{n}\mathbb{E}\left[|K_h(u-X_i)-(K_h\star f)(u)|^2\right]\mathbb{E}\left[a_i^2(X_i)\right]}
\end{aligned}$$

$$\begin{aligned}
&\leq \sqrt{n}\sqrt{\sum_{i=2}^{n}\mathbb{E}\left[K_h^2(u-X_i)\right]\mathbb{E}\left[a_i^2(X_i)\right]} \\
&\leq \sqrt{n}\sqrt{\|f\|_\infty\|K_h\|^2\sum_{i=2}^{n}\mathbb{E}\left[a_i^2(X_i)\right]} \\
&\leq \|K_h\|\sqrt{n\|f\|_\infty}.
\end{aligned}$$

Straightforward computations lead to

$$\begin{aligned}
&\sum_{j=1}^{n-1}\int\mathbb{E}\left[|K_{h_{\min}}(u-X_j)-(K_{h_{\min}}\star f)(u)||b_j(X_j)|\right]du \\
&\leq 2\|K\|_1\sum_{j=1}^{n-1}\mathbb{E}[|b_j(X_j)|] \\
&\leq 2\|K\|_1\sqrt{n}\sqrt{\sum_{j=1}^{n-1}\mathbb{E}^2[|b_j(X_j)|]} \\
&\leq 2\|K\|_1\sqrt{n}.
\end{aligned}$$

Finally,

$$\sum_{i=2}^{n}\sum_{j=1}^{i-1}\mathbb{E}[G_{h,h_{\min}}(X_i,X_j)a_i(X_i)b_j(X_j)] \leq 2n\|K\|_1\sqrt{\|f\|_\infty}\|K_h\|$$

and

$$\begin{aligned}
\frac{Dx}{n^2} &\leq \frac{4\|K\|_1\sqrt{\|f\|_\infty}\|K_h\|x}{n} \\
&\leq \frac{\theta'}{3}\frac{\|K\|^2}{nh}+\frac{12\|K\|_1^2\|f\|_\infty\,x^2}{\theta' n}.
\end{aligned}$$

From Lemma 10, we obtain the following result.

LEMMA 11. *With probability larger than* $1-9.54|\mathcal{H}|\exp(-x)$, *for any* $h\in\mathcal{H}$,

$$\begin{aligned}
|\langle\hat{f}_h-f,\hat{f}_{h_{\min}}-f\rangle-\frac{\langle K_h,K_{h_{\min}}\rangle}{n}| \leq\ & \theta'\|f_h-f\|^2+\theta'\frac{\|K\|^2}{nh} \\
&+\left(\frac{\theta'}{2}+\frac{1}{2\theta'}\right)\|f_{h_{\min}}-f\|^2 \\
&+\frac{C_1(K)}{\theta'}\left(\frac{\|f\|_\infty x^2}{n}+\frac{x^3}{n^2h_{\min}}\right),
\end{aligned}$$

where $C_1(K)$ *is a constant only depending on* K.

PROOF. We have to control (18) and (19). Let h and h' be fixed. First, we have

$$\begin{aligned}\langle \hat{f}_h, f_{h'}\rangle &= \frac{1}{n}\sum_{i=1}^{n}\int K_h(x-X_i)f_{h'}(x)dx \\ &= \frac{1}{n}\sum_{i=1}^{n}(\tilde{K}_h \star f_{h'})(X_i).\end{aligned}$$

Therefore,

$$|\langle \hat{f}_h, f_{h'}\rangle| \le \|\tilde{K}_h \star f_{h'}\|_\infty \le \|f_{h'}\|_\infty \|K\|_1 \le \|f\|_\infty \|K\|_1^2,$$

which gives the following control of (18):

$$\left|-\frac{1}{n}\langle \hat{f}_h, f_{h_{\min}}\rangle - \frac{1}{n}\langle f_h, \hat{f}_{h_{\min}}\rangle + \frac{1}{n}\langle f_h, f_{h_{\min}}\rangle\right| \le \frac{3\|f\|_\infty \|K\|_1^2}{n}.$$

So it remains to bound the three terms of (19). Note that

$$\begin{aligned}V(h,h') &= < \hat{f}_h - f_h, f_{h'} - f > \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(g_{h,h'}(X_i) - \mathbb{E}[g_{h,h'}(X_i)]\right)\end{aligned}$$

with

$$g_{h,h'}(x) = \langle K_h(.-x), f_{h'} - f\rangle = (\tilde{K}_h \star (f_{h'} - f))(x).$$

Since

$$\|g_{h,h'}\|_\infty \le \|K\|_1 \|f_{h'} - f\|_\infty \le \|K\|_1(1+\|K\|_1)\|f\|_\infty$$

and

$$\mathbb{E}[g_{h,h'}^2(X_1)] \le \|f\|_\infty \|\tilde{K}_h \star (f_{h'} - f)\|^2 \le \|f\|_\infty \|K\|_1^2 \|f_{h'} - f\|^2$$

then, with probability larger than $1 - 2e^{-x}$, Bernstein's inequality leads to

$$\begin{aligned}|V(h,h')| &\le \sqrt{\frac{2x}{n}\|f\|_\infty \|K\|_1^2 \|f_{h'} - f\|^2} + \frac{x\|f\|_\infty \|K\|_1(1+\|K\|_1)}{3n} \\ &\le \frac{\theta'}{2}\|f_{h'} - f\|^2 + \frac{C_V(K)\|f\|_\infty x}{\theta' n},\end{aligned}$$

where $C_V(K)$ is a constant only depending on the kernel norm $\|K\|_1$. Previous inequalities are first applied with $h' = h_{\min}$ and then we invert the roles of h and $h_{\min}$. To conclude the proof of the lemma, we use

$$|\langle f_h - f, f_{h_{\min}} - f\rangle| \le \frac{\theta'}{2}\|f_h - f\|^2 + \frac{1}{2\theta'}\|f_{h_{\min}} - f\|^2.$$

Now, Proposition 8 gives, with probability larger than $1-\square|\mathcal{H}|\exp(-x)$, for any $h \in \mathcal{H}$,

$$\|f_h - f\|^2 + \frac{\|K\|^2}{nh} \leq 2\|\hat{f}_h - f\|^2 + C_2(K)\|f\|_\infty \frac{x^2}{n},$$

where $C_2(K)$ only depends on K. Hence, by applying Lemma 11, with probability larger than $1-\square|\mathcal{H}|\exp(-x)$, for any $h \in \mathcal{H}$,

$$\begin{aligned} &\left|\langle \hat{f}_h - f, \hat{f}_{h_{\min}} - f\rangle - \frac{\langle K_h, K_{h_{\min}}\rangle}{n} - \langle \hat{f}_{\hat{h}} - f, \hat{f}_{h_{\min}} - f\rangle + \frac{\langle K_{\hat{h}}, K_{h_{\min}}\rangle}{n}\right| \\ \leq\ & 2\theta'\|\hat{f}_h - f\|^2 \\ & + 2\theta'\|\hat{f}_{\hat{h}} - f\|^2 + \left(\theta' + \frac{1}{\theta'}\right)\|f_{h_{\min}} - f\|^2 + \frac{\tilde{C}(K)}{\theta'}\left(\frac{\|f\|_\infty x^2}{n} + \frac{x^3}{n^2 h_{\min}}\right), \end{aligned}$$

where $\tilde{C}(K)$ is a constant only depending on K. It remains to use (16) and to choose $\theta' = \frac{\theta}{4}$ to conclude.

5.2. Proof of Theorem 2. We set $\tau = \lambda - 1$. Let $\varepsilon \in (0,1)$, and $\theta \in (0,1)$ depending on ε to be specified later. Using Theorem 9 with the chosen penalty, we obtain, with probability larger than $1-\square|\mathcal{H}|\exp(-x)$, for any $h \in \mathcal{H}$,

$$\begin{aligned} (1-\theta)\|\hat{f}_{\hat{h}} - f\|^2 + \tau\frac{\|K_{\hat{h}}\|^2}{n} \leq\ & (1+\theta)\|\hat{f}_h - f\|^2 + \tau\frac{\|K_h\|^2}{n} + \frac{C_2}{\theta}\|f_{h_{\min}} - f\|^2 \\ & + \frac{C(K)}{\theta}\left(\frac{\|f\|_\infty x^2}{n} + \frac{x^3}{n^2 h_{\min}}\right). \end{aligned} \tag{20}$$

We first consider the case where $\tau \geq 0$. Then $\tau\|K_{\hat{h}}\|^2/n \geq 0$, and, using Proposition 8, with probability $1-\square|\mathcal{H}|\exp(-x)$

$$\tau\frac{\|K_h\|^2}{n} \leq \tau(1+\theta)\|f - \hat{f}_h\|^2 + \tau\frac{C'(K)\|f\|_\infty x^2}{\theta^3 n},$$

where $C'(K)$ is a constant only depending on the kernel K. Hence, with probability $1-\square|\mathcal{H}|\exp(-x)$

$$\begin{aligned} (1-\theta)\|\hat{f}_{\hat{h}} - f\|^2 \leq\ & (1+\theta+\tau(1+\theta))\|\hat{f}_h - f\|^2 + \frac{C_2}{\theta}\|f_{h_{\min}} - f\|^2 \\ & + \frac{C(K)}{\theta}\left(\frac{\|f\|_\infty x^2}{n} + \frac{x^3}{n^2 h_{\min}}\right) + \tau\frac{C'(K)\|f\|_\infty x^2}{\theta^3 n}. \end{aligned}$$

With $\theta = \varepsilon/(\varepsilon + 2 + 2\tau)$, we obtain

$$\|\hat{f}_{\hat{h}} - f\|^2 \leq (1 + \tau + \varepsilon)\|\hat{f}_h - f\|^2 + \frac{C_2(\varepsilon + 2 + 2\tau)^2}{(2 + 2\tau)\varepsilon}\|f_{h_{\min}} - f\|^2 + C''(K, \varepsilon, \tau)\left(\frac{\|f\|_\infty x^2}{n} + \frac{x^3}{n^2 h_{\min}}\right)$$

where $C''(K, \varepsilon, \tau)$ is a constant only depending on K, ε, τ.

Let us now study the case $-1 < \tau \leq 0$. In this case $\tau\|K_h\|^2/n \leq 0$, and, using Proposition 8, with probability $1 - \square|\mathcal{H}|\exp(-x)$

$$\tau\frac{\|K_{\hat{h}}\|^2}{n} \geq \tau(1 + \theta)\|f - \hat{f}_{\hat{h}}\|^2 + \tau\frac{C'(K)\|f\|_\infty x^2}{\theta^3 n},$$

where $C'(K)$ is a constant only depending on the kernel K. Hence, (20) becomes

$$\begin{aligned}(1 - \theta + \tau(1 + \theta))\|\hat{f}_{\hat{h}} - f\|^2 \leq\ & (1 + \theta)\|\hat{f}_h - f\|^2 + \frac{C_2}{\theta}\|f_{h_{\min}} - f\|^2 \\ & + \frac{C(K)}{\theta}\left(\frac{\|f\|_\infty x^2}{n} + \frac{x^3}{n^2 h_{\min}}\right) \\ & - \tau\frac{C'(K)\|f\|_\infty x^2}{\theta^3 n}.\end{aligned}$$

With $\theta = (\varepsilon(\tau + 1)^2)/(2 + \varepsilon(1 - \tau^2)) < 1$, we obtain, with probability $1 - \square|\mathcal{H}|\exp(-x)$,

$$\begin{aligned}\|\hat{f}_{\hat{h}} - f\|^2 \leq\ & \left(\frac{1}{1 + \tau} + \varepsilon\right)\|\hat{f}_h - f\|^2 + C''(\varepsilon, \tau)\|f_{h_{\min}} - f\|^2 \\ & + C'''(K, \varepsilon, \tau)\left(\frac{\|f\|_\infty x^2}{n} + \frac{x^3}{n^2 h_{\min}}\right).\end{aligned}$$

where $C''(\varepsilon, \tau)$ depends on ε and τ and $C'''(K, \varepsilon, \tau)$ depends on $K\varepsilon, \tau$.

5.3. Proof of Theorem 3. We still set $\tau = \lambda - 1$. Set $\theta \in (0, 1)$ such that $\theta < -(1 + \tau)/5$. Let us first write the result of Theorem 9 with the chosen penalty and $h = h_{\min}$:

$$\begin{aligned}(1 - \theta)\|\hat{f}_{\hat{h}} - f\|^2 + \tau\frac{\|K_{\hat{h}}\|^2}{n} \leq\ & (1 + \theta)\|\hat{f}_{h_{\min}} - f\|^2 + \tau\frac{\|K_{h_{\min}}\|^2}{n} \\ & + \frac{C_2}{\theta}\|f_{h_{\min}} - f\|^2 \\ & + \frac{C(K)}{\theta}\left(\frac{\|f\|_\infty x^2}{n} + \frac{x^3}{n^2 h_{\min}}\right).\end{aligned}$$

Then inequality (14) with $h = h_{\min}$ gives, with probability $1 - \square|\mathcal{H}| \exp(-x)$

$$\|f - \hat{f}_{h_{\min}}\|^2 \leq (1+\theta)\left(\|f - f_{h_{\min}}\|^2 + \frac{\|K\|^2}{nh_{\min}}\right) + \frac{C'(K)\|f\|_\infty x^2}{\theta^3 n}$$

where $C'(K)$ is a constant only depending on the kernel K. That entails

$$\begin{aligned}(1-\theta)\|\hat{f}_{\hat{h}} - f\|^2 + \tau\frac{\|K_{\hat{h}}\|^2}{n} \quad \leq \quad & ((1+\theta)^2 + C_2/\theta)\|f - f_{h_{\min}}\|^2 \\ & + (\tau + (1+\theta)^2)\frac{\|K_{h_{\min}}\|^2}{n} \\ & + \frac{C(K)}{\theta}\left(\frac{\|f\|_\infty x^2}{n} + \frac{x^3}{n^2 h_{\min}}\right) \\ & + \square(1+\theta)\frac{C'(K)\|f\|_\infty x^2}{\theta^3 n}.\end{aligned}$$

We denote $u_n := \frac{\|f_{h_{\min}} - f\|^2}{\|K\|^2/nh_{\min}}$. The previous inequality can then be rewritten

$$\begin{aligned}(1-\theta)\|\hat{f}_{\hat{h}} - f\|^2 + \tau\frac{\|K_{\hat{h}}\|^2}{n} \leq & [((1+\theta)^2 + C_2/\theta)u_n + \tau + (1+\theta)^2]\frac{\|K_{h_{\min}}\|^2}{n} \\ & + C(K,\theta)\left(\frac{\|f\|_\infty x^2}{n} + \frac{x^3}{n^2 h_{\min}}\right)\end{aligned}$$

where $C(K,\theta)$ is a constant only depending on K and θ. Next we apply inequality (15) to $h = \hat{h}$: with probability $1 - \square|\mathcal{H}| \exp(-x)$

$$\frac{\|K\|^2}{n\hat{h}} \leq 2\|f - \hat{f}_{\hat{h}}\|^2 + \frac{C'(K)\|f\|_\infty x^2}{n},$$

with $C'(K)$ only depending on K. We obtain that with probability $1 - \square|\mathcal{H}| \exp(-x)$

$$\begin{aligned}[(1-\theta)/2 + \tau]\frac{\|K\|^2}{n\hat{h}} \quad \leq \quad & [((1+\theta)^2 + C_2/\theta)u_n + \tau + (1+\theta)^2]\frac{\|K\|^2}{nh_{\min}} \\ & + C'(K,\theta)\left(\frac{\|f\|_\infty x^2}{n} + \frac{x^3}{n^2 h_{\min}}\right)\end{aligned}$$

where $C'(K,\theta)$ is a constant only depending on K and θ. By assumption, $u_n = o(1)$, then there exists N such that for $n \geq N$, $((1+\theta)^2 + C_2/\theta)u_n \leq \theta$.

Let us now choose $x = (n/\log n)^{1/3}$. Using the bound on $h_{\min}$, the remainder term is bounded in the following way: For $C'(K,\theta)$ only depending on K and θ,

$$\begin{aligned}
&\left(\frac{nh_{\min}}{\|K\|^2}\right) C'(K,\theta)\left(\frac{\|f\|_\infty x^2}{n} + \frac{x^3}{n^2 h_{\min}}\right) \\
\leq\ & C''(K,\theta,\|f\|_\infty)\left(x^2 h_{\min} + \frac{x^3}{n}\right) \\
\leq\ & C''(K,\theta,\|f\|_\infty)\left(\frac{(\log n)^{\beta-2/3}}{n^{1/3}} + \frac{1}{\log n}\right) = o(1).
\end{aligned}$$

Then for n large enough, this term is bounded by θ. Thus, for n large enough, with probability $1 - \square|\mathcal{H}|\exp(-(n/\log n)^{1/3})$

$$[(1-\theta)/2 + \tau]\frac{\|K\|^2}{n\hat{h}} \leq [\theta + \tau + (1+\theta)^2 + \theta]\frac{\|K\|^2}{nh_{\min}}$$

and then

$$\frac{\hat{h}}{(1-\theta)/2 + \tau} \geq \frac{h_{\min}}{\tau + 1 + 5\theta}.$$

Note that $(1-\theta)/2 + \tau < 1 + \tau < 0$, and we have chosen θ such that $\tau + 1 + 5\theta < 0$. Then

$$\hat{h} \leq \frac{\tau + (1-\theta)/2}{\tau + 1 + 5\theta} h_{\min}.$$

One can choose for example $\theta = -(\tau+1)/10$.

5.4. Proof of Corollary 7. Let $f \in \tilde{\mathcal{N}}_{\mathbf{2},d}(\boldsymbol{\beta}, \mathbf{L}, B)$ and $\mathcal{E}$ the event corresponding to the intersection of events considered in Theorem 5 and Proposition 8. For any $A > 0$, by taking x proportional to $\log n$, $\mathbb{P}(\mathcal{E}) \geq 1 - n^A$. Therefore,

$$\begin{aligned}
\mathbb{E}\left[\|\hat{f}_{\hat{h}} - f\|^2\right] &\leq \mathbb{E}\left[\|\hat{f}_{\hat{h}} - f\|^2 1_{\mathcal{E}}\right] + \mathbb{E}\left[\|\hat{f}_{\hat{h}} - f\|^2 1_{\mathcal{E}^c}\right] \\
&\leq \tilde{C}_2(\boldsymbol{\beta}, K, B, d, \lambda)\left(\prod_{j=1}^d L_j^{\frac{1}{\bar{\beta}_j}}\right)^{\frac{2\bar{\beta}}{2\bar{\beta}+1}} n^{-\frac{2\bar{\beta}}{2\bar{\beta}+1}},
\end{aligned}$$

where $\tilde{C}_2(\boldsymbol{\beta}, K, B, d, \lambda)$ is a constant only depending on $\boldsymbol{\beta}$, K, B, d and λ. We have used (13) on $\mathcal{E}$ and for any $h \in \mathcal{H}$,

$$\|\hat{f}_h - f\|^2 \leq 2\|f\|^2 + \frac{2n\|K\|^2}{\|K\|_\infty \|K\|_1},$$

on $\mathcal{E}^c$.

References

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings 2nd International Symposium on Information Theory.* (P. N. Petrov and F. Csaki, eds.). Akademia Kiado, Budapest, pp. 267–281.

ARLOT, S. and BACH, F. (2009). Data-driven calibration of linear estimators with minimal penalties. In *Advances in Neural Information Processing Systems.* (Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams and A. Culotta, eds.). vol. **22**, pp. 46–54.

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

BAHADUR, R.R. (1958). Examples of inconsistency of maximum likelihood estimates. *Sankhya Ser. A* **20**, 207–210.

BARRON, A.R. and COVER, T.M. (1991). Minimum complexity density estimation. *IEEE Trans. Inform. Theory* **37**, 1034–1054.

BARRON, A.R., BIRGÉ, L. and MASSART, P. (1999). Risk bounds for model selection via penalization. *Probab. Th. Rel. Fields* **113**, 301–415.

BAUDRY, J.-P., MAUGIS, C. and MICHEL, B. (2011) Slope heuristics: overview and implementation. *Stat. Comput.*, 1–16.

BERTIN, K., LACOUR, C. and RIVOIRARD, V. (2016). Adaptive pointwise estimation of conditional density function. *Ann. Inst. Henri Poincaré Probab. Stat.* **52**, 939–980.

BERTIN, K., LE PENNEC, E. and RIVOIRARD, V. (2011). Adaptive Dantzig density estimation. *Ann. Inst. Henri Poincaré Probab. Stat.* **47**, 43–74.

BICKEL, P.J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Stat.* **37**, 1705–1732.

BIRGÉ, L. and MASSART, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Th. Relat. Fields* **97**, 113–150.

BIRGÉ, L. and MASSART, P. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli* **4**, 329–375.

BIRGÉ, L. and MASSART, P. (2001) Gaussian model selection. *J. Eur. Math. Soc.*, 203–268.

BIRGÉ, L. and MASSART, P. (2007). Minimal penalties for Gaussian model selection. *Probab. Th. Rel. Fields* **138**, 33–73.

BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013) *Concentration inequalities.* Oxford University Press.

DANIEL, C. and WOOD, F.S. (1971). *Fitting Equations to Data.* Wiley, New York.

DEVROYE, L. and LUGOSI, G. (2001). *Combinatorial methods in density estimation*, Springer Series in Statistics. Springer, New York.

DONOHO, D.L. and JOHNSTONE, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–455.

DONOHO, D.L. and JOHNSTONE, I.M. (1994). Ideal denoising in an orthonormal basis chosen from a library of bases. *C. R. Acad. Sc. Paris Sér. I Math.* **319**, 1317–1322.

DONOHO, D.L., JOHNSTONE, I.M., KERKYACHARIAN, G. and PICARD, D. (1995). Wavelet shrinkage:Asymptopia? *J. R. Statist. Soc. B* **57**, 301–369.

DONOHO, D.L., JOHNSTONE, I.M., KERKYACHARIAN, G. and PICARD, D. (1996). Density estimation by wavelet thresholding. *Ann. Statist.* **24**, 508–539.

DOUMIC, M., HOFFMANN, M., REYNAUD-BOURET, P. and RIVOIRARD, V. (2012). Nonparametric estimation of the division rate of a size-structured population. *SIAM J. Numer. Anal.* **50**, 925–950.

EFROIMOVITCH, S.YU. and PINSKER, M.S. (1984). Learning algorithm for nonparametric filtering. *Automat. Remote Control* **11**, 1434–1440. translated from Avtomatika i Telemekhanika **11**, 58–65.

GOLDENSHLUGER, A. and LEPSKI, O. (2008) Universal pointwise selection rule in multivariate function estimation, vol. 14.

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

GOLDENSHLUGER, A. and LEPSKI, O. (2011). Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *Ann. Statist.* **39**, 1608–1632.

GOLDENSHLUGER, A. and LEPSKI, O. (2013). General selection rule from a family of linear estimators. *Theory Probab. Appl.* **57**, 209–226.

GOLDENSHLUGER, A. and LEPSKI, O. (2014). On adaptive minimax density estimation on $\mathbb{R}^d$. *Theory Probab. Appl.* **159**, 479–543.

KERKYACHARIAN, G., LEPSKI, O. and PICARD, D. (2008). Nonlinear estimation in anisotropic multiindex denoising. Sparse case. *Theory Probab. Appl.* **52**, 58–77.

LACOUR, C. and MASSART, P.P. (2016) Minimal penalty for Goldenschluger-Lepski method < hal-01121989v2 >. To appear in *Stoch. Proc. Appl.*

LEBARBIER, E. (2005). Detecting multiple change points in the mean of Gaussian process by model selection. *Signal Process.* **85**, 717–736.

LEPSKII, O.V. (1990). On a problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **36**, 454–466.

LEPSKII, O.V. (1991). Asymptotically minimax adaptive estimation I: Upper bounds. Optimally adaptive estimates. *Theory Probab. Appl.* **36**, 682–697.

LEPSKII, O.V. (2013). Upper functions for positive random functionals. II. Application to the empirical processes theory, Part 1. *Math. Methods Statist.* **22**, 83–99.

LERASLE, M. (2012). Optimal model selection in density estimation. *Ann. Inst. Henri Poincaré Probab. Stat.* **48**, 884–908.

LERASLE, M., MALTER-MAGALAHES, N. and REYNAUD-BOURET, P. (2015) Optimal kernel selection for density estimation. To appear in High dimensional probabilities VII: The Cargese Volume.

LERASLE, M. and TAKAHASHI, D.Y. (2016). Sharp oracle inequalities and slope heuristic for specification probabilities estimation in general random fields. *Bernoulli* **22**, 1.

MALLOWS, C.L. (1973). Some comments on C_p. *Technometrics* **15**, 661–675.

MASSART, P. (2007). *Concentration inequalities and model selection. Ecole d'été de Probabilités de Saint-Flour 2003. Lecture Notes in Mathematics 1896.* Springer, Berlin/Heidelberg.

NIKOL'SKII, S. M. (1977) Priblizhenie funktsii mnogikh peremennykh i teoremy vlozheniya. (Russian) [Approximation of functions of several variables and imbedding theorems] Second edition, revised and supplemented. "Nauka", Moscow.

PINSKER, M.S. (1980). Optimal filtration of square-integrable signals in Gaussian noise. *Probl. Inf. Transm.* **16**, 120–133.

REYNAUD-BOURET, P., RIVOIRARD, V. and TULEAU-MALOT, C. (2011). Adaptive density estimation: a curse of support? *J. Statist. Plann. Inference* **141**, 115–139.

RIGOLLET, P. (2006). Adaptive density estimation using the blockwise Stein method. *Bernoulli* **12**, 351–370.

SAUMARD, A. (2013). Optimal model selection in heteroscedastic regression using piecewise polynomial functions. *Electron. J. Stat.* **7**, 1184–1223.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464.
SILVERMAN, B. W. (1986). *Density estimation for statistics and data analysis. Monographs on Statistics and Applied Probability.* Chapman & Hall, London.
TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.* **58**, 267–288.

CLAIRE LACOUR
PASCAL MASSART
LABORATOIRE DE MATHÉMATIQUES
D'ORSAY, UNIVERSITY PARIS-SUD,
UMR8628, ORSAY 91405, FRANCE
E-mail: claire.lacour@u-psud.fr
pascal.massart@u-psud.fr

VINCENT RIVOIRARD
CEREMADE, CNRS, UMR 7534,
UNIVERSITÉ PARIS DAUPHINE,
PSL RESEARCH UNIVERSITY,
75016 PARIS, FRANCE
E-mail: rivoirard@ceremade.dauphine.fr

Paper received: 7 July 2016.