AN INTRODUCTION TO NONPARAMETRIC ADAPTIVE ESTIMATION

Gaëlle Chagny

AN INTRODUCTION TO NONPARAMETRIC ADAPTIVE ESTIMATION

GAËLLE CHAGNY

LMRS, UMR CNRS 6085, UNIVERSITÉ DE ROUEN NORMANDIE, FRANCE.
GAELLE.CHAGNY@UNIV-ROUEN.FR

Abstract. *Statistical estimation* aims at building procedures to recover unknown parameters by analysing some measured data sampled from a large population. This note deals with the case of infinite dimensional parameters, that is functions, through the example of probability density estimation. After discussing how to quantify the performances of estimation methods, we discuss the limits of accuracy of any estimator for the density (minimax point of view) and present the main two methods of *nonparametric* estimation: *projection and kernel estimators*. Upper-bounds on the accuracy of the defined estimators for a fixed amount of data are derived. They highly depend on smoothing parameters (the model dimension and the bandwidth, respectively for the two methods), which should be carefully chosen. The second part of the text is devoted to *data-driven estimator selection*, for which we provide a brief review: both the *model selection* and the *bandwidth choice* issues are addressed. We describe two methods that permit to obtain so-called *oracle-type inequalities* while being *adaptive*: the selection does not depend on the unknown smoothness of the target density. A large list of references is provided, and numerical experiments illustrate the theoretical results.

1. Introduction

1.1. **Statistical inference.** Statistical inference is the use of probability theory to deduce properties or characteristics of a population which is only partially observed.

The general process can be described as follows. An observed data set $\mathbb{X} = \{x_1, x_2, \dots, x_n\}$, assumed to be sampled from a large population, is available. Typically, the x_i's are assumed to be realisations of independent and identically distributed (*i.i.d.* in the sequel) random variables (or random vectors) $\{X_1, \dots, X_n\}$ on a measurable space $(\Omega, \mathcal{A})$. The first step is to choose a *model*, that is a set of probability distributions $\mathcal{P} = \{\mathbb{P}_\theta, \theta \in \Theta\}$ on $(\Omega, \mathcal{A})$, which should adequatly describe the data, in the following sense: the "true" underlying probability distribution of the X_i's is supposed to be not too far from this set, or ideally, is supposed to be an element $\mathbb{P}_{\theta_0}$ of the model. Then, the goal is to propose some methods to recover from $\mathbb{X}$ the features of this probability distribution which describes at best the data in the model. This can be done through the *estimation* of the parameter θ_0, which is the problem considered in this note, and even if statistical inference cannot be reduced to estimation (it also includes tests, classification...). The next step is thus naturally to check the performance of the proposed estimation methods.

1.2. **Parametric versus nonparametric statistics.** In many situations, there is sufficient motivation for using models that are described by a finite number of finite dimensional parameters (for example if the statistician has prior information about the studied population). This is *parametric statistics*. In such a framework, the assumption is that Θ is a part of an euclidean

Date: September 6, 2016.

space, typically $\Theta \subset \mathbb{R}^d$. For example, setting $\Theta = \mathbb{R} \times \mathbb{R}_+$, and $\mathbb{P}_{(\mu,\sigma^2)} = \mathcal{N}(\mu, \sigma^2)$ for any $(\mu, \sigma^2) \in \Theta$, permits to parametrise the Gaussian family of distribution. If a data set is supposed to come from this model, estimating the two parameters (the mean μ_0 and the variance σ_0^2) of the "true" underlying distribution is sufficient to recover it entirely. On the opposite, if for whatever reason a parametric model is not forthcoming (when, for example, there is no prior opinion about the data or when a parametrised distribution cannot easily fit the data), it can be interesting to choose *nonparametric statistics*. The idea is to make as few assumptions as possible on the underlying probability distribution of the observations, to leave it essentially free: the number of parameters may not be fixed and may grow with the amount of data, or a distribution-free approach can be used (all the possible probability measures are permitted)... The set Θ is thus an infinite dimensional space, *e.g.* a functional space: the space of all densities, the space of all cumulative distribution functions... The objective is thus to estimate a function. The only constraint concerns the smoothness of the target function: we will restrict Θ to a ball of a functional space (Hölder spaces...). Specific methods are required to deal with a nonparametric estimation problem. It should be kept in mind that, since few assumptions are required, the applicability of nonparametric methods is much wider than parametric ones, and they are more robust. In particular, we will consider *adaptive estimation*, which aims at building totally data-driven functional estimators, that do not depend on the unknown smoothness of the function to recover. However, when the choice of the model is correct, parametric methods will produce more accurate and precise estimates: to obtain similar results (similar "convergence rates", see definition in Section 2.2.2 below), a nonparametric estimator requires a larger data set.

1.3. **Overview of the note.** Nonparametric methods in statistical inference are now widely developed for estimation (as presented above) but also for testing and we cannot reasonably make an exhaustive review of the literature in this note. We refer the reader to the monographies of Conover (1980); Wasserman (2006), and Bosq (2012). Our aim is to briefly present the two main classes of functional estimators (kernel and projection methods) in the simple framework of univariate density estimation, and their nonasymptotic theoretical study. The framework and the notations are introduced in Section 2. Section 3 permits to define the estimators and to discuss the theoretical results which can be expected. We then explain how the procedure can be "tuned" to adapt automatically to the unknown smoothness of the function to be estimated, which constitutes the main goal of adaptive estimation (Section 4): a brief overview of adaptive methods is proposed, and two of them are developed with more details (model selection via penalisation and Goldenshluger-Lepski method). Numerical experiments illustrate the methods throughout the text. The exposition is based on the monographies of Tsybakov (2009) and Comte (2015), where most of the results and proofs can be found.

2. Statistical framework

2.1. **Estimation problem and motivation.** In this note, we consider the basic problem of univariate density estimation. Let $\mathbb{X} = \{X_1, \dots, X_n\}$ be an *i.i.d.* sample of a real random variable X on $(\Omega, \mathcal{A})$, with probability density function f with respect to the Lebesgue measure. The function f is considered completely unknown, and the aim is to recover it from the data $X_1, \dots, X_n$ on an interval $I \subset \mathbb{R}$ (for simplifying, we confuse the realisations of the random variables with the random variables themselves, compared to what has been described in the introduction). The model is thus for the moment $\mathcal{P} = \{\mathbb{P}_f, f \in \mathcal{F}\}$, where $\mathcal{F}$ is the set of the nonnegative functions on $\mathbb{R}$ which integrate to one, and $\mathbb{P}_f$ the probability measure with density

f. We will define some *estimators* $\hat{f}(\mathbb{X})$ for the true f, that is some measurable functions of the data (we denote by f and not by f_0 the true density). Any estimator is thus a function: $\hat{f}(\mathbb{X}) : I \to \mathbb{R}$. For the sake of simplicity we denote it only by $\hat{f}$.

Estimating a density is a very classical but important question: this enables to visualise a data set or to recover geometrical properties of a probability distribution (like the number of modes) for example. Applications are obviously various. The distributions of most of quantitative variables (wages, height in a population...) can be represented by density estimators. Let us quote two more specific examples: Wasserman (2006) uses it for galaxy cluster detection purpose (Chapter 4), Efromovich (2008) proposes to analyse a lottery (daily numbers game) with density estimation (Chapter 1)... However simple as the model may seem, it raises numerous questions. Moreover, density estimation is also a starting point to propose new estimation methods which can be then developed for other objectives. This is thus a current research topic. We will not explore a "real" data set in this note, which is devoted to a theoretical point of view. However, all the methods will be illustrated through simulations. We will mainly focus on one example: the estimation of the density f_{Simul} of a mixture of two Gaussian distributions $0.5\mathcal{N}(-2, 0.4)+0.5\mathcal{N}(2, 0.4)$, which can be expressed as

$$f_{\text{Simul}}(x) = \frac{0.5}{\sqrt{0.8\pi}} \left(\exp\left(-\frac{(x+2)^2}{0.8} \right) + \exp\left(-\frac{(x-2)^2}{0.8} \right) \right). \tag{1}$$

2.2. **Evaluation of a functional estimation method.**

2.2.1. *Quadratic risk.* Technically, any measurable function $\hat{f}$ of the data $\mathbb{X}$ is an *estimator* for f. Thus, any estimation procedure has two steps: a step of definition of an estimator, and a step of evaluation of the performances of the estimator. The second step involves being able to compare two functions. Classically, L^p-distances are considered, mostly with $p = 1, 2$ or ∞. We choose $p = 2$ in the sequel: this choice is motivated by the Hilbertian method we will consider, and can be also seen as a compromise between the L^1-norm (natural for density estimation: a density is an L^1-function) and the L^∞-norm (which can be easily affected by peaks). Thus, we assume that $\mathcal{F}$ is the set of probability densities which are squared integrable on I (with respect to the Lebesgue measure): $\mathcal{F} = L^2(I)$. The distance is called the *loss function*, and we define the associated *risk* of an estimate $\hat{f}$ for the estimation of f by

$$\mathcal{R}(\hat{f}, f) = \mathbb{E}_f \left[\left\| \hat{f} - f \right\|^2 \right], \text{ with } \left\| \hat{f} - f \right\|^2 = \int_I \left(\hat{f}(x) - f(x) \right)^2 dx, \tag{2}$$

where $\mathbb{E}_f$ is the expectation under the distribution $\mathbb{P}_f$. This risk is the *Mean Integrated Squared Error* (M.I.S.E.) or quadratic risk. We could also consider the Mean Squared Error (M.S.E.), based on the pointwise loss, and defined by $\mathbb{E}_f[(\hat{f}(x_0) - f(x_0))^2]$, for any fixed $x_0 \in I$ (most of the following results have their analogous versions for the M.S.E.), or other risks based on intrisic probability measures (such as the Hellinger distance or the Kullback divergence).

2.2.2. *Minimax point of view.* What can be expected for the risk of an estimator? A primary requirement is that it is as small as possible: it should go to zero when the number n of observations in the data sample goes to infinity. An estimator $\hat{f}$ reaches the *convergence rate* ψ_n over a functional class $\mathcal{F}$ if the following upper-bound holds:

$$\sup_{f \in \mathcal{F}} \mathcal{R}(\hat{f}, f) \leq C\psi_n^2, \tag{3}$$

where $(\psi_n)_{n\in\mathbb{N}\setminus\{0\}}$ is a decreasing sequence which goes to zero, C a positive constant, and where $\sup_{f\in\mathcal{F}}\mathcal{R}(\hat{f},f)$ is called the *maximal risk* over $\mathcal{F}$. To assess the *optimality* of an estimator, such an upper-bound could be compared to the *minimax risk*, defined by

$$\inf_{\tilde{f}}\sup_{f\in\mathcal{F}}\mathcal{R}(\tilde{f},f),$$

where the infimum is over all possible estimators $\tilde{f}$ for f that can be computed from $\mathbb{X}$. This smallest maximum risk among all estimators measures what happens in the worst case allowed in the problem (what happens when estimating the most difficult function f of the class $\mathcal{F}$). If, for a sequence $(\psi_n)_n$ like above,

$$\inf_{\tilde{f}}\sup_{f\in\mathcal{F}}\mathcal{R}(\tilde{f},f)\geq c\psi_n^2, \tag{4}$$

then ψ_n is called a *minimax rate of convergence* for the estimation of f over $\mathcal{F}$. If the upper-bound (3) is completed by a lower bound of the form (4), with the same sequence $(\psi_n)_n$ (that is if the upper-bound matches with the lower-bound), then $\hat{f}$ satisfying (3) is said to be *minimax optimal*.

Such convergence rates generally depend on the smoothness of the function f: we restrict ourselves to an (infinite dimensional) subset $\mathcal{F}_\alpha$ of $\mathcal{F}$, where $\alpha>0$ is an index that quantifies the smoothness of f. The statisticians consider general spaces, typically Hölder or Nikol'skiĭ space when dealing with kernel estimators, Sobolev or Besov spaces for projection methods. In this note we will not give precise definitions: see Tsybakov (2009) for the first ones and DeVore and Lorentz (1993) for the seconds. Heuristically, $\mathcal{F}_\alpha=\mathcal{C}^\alpha(I)$, the space of $\alpha-$times differentiable functions on I: the larger α, the faster the convergence rates. For the considered univariate density estimation problem, the minimax quadratic risk can be expressed as

$$\psi_n=n^{-\frac{\alpha}{2\alpha+1}} \tag{5}$$

for Hölder spaces (Juditsky and Lambert-Lacroix, 2004), Nikol'skiĭ and Sobolev spaces (Ibragimov and Has′minskiĭ, 1980; Hasminskii and Ibragimov, 1990), Besov spaces (Kerkyacharian and Picard, 1992; Donoho et al., 1996; Reynaud-Bouret et al., 2011). As expected, the rate is slower than the one classically obtained for parametric estimation, which is $\psi_n=1/\sqrt{n}$. The computations of the lower bounds are based on general reduction schemes. A clear account is provided by Tsybakov (2009), chapter 2. The present work addresses the problem of building estimators that reach the minimax risk (5), without using the unknown smoothness index α: since f is unknown, α is probably unknown too and should not be used to build estimators. This is what we call *adaptive estimation*.

Notice finally that the accuracy/optimality of an estimation procedure can be measured through other quality criteria: *efficiency* is for example an additional feature that can be considered (see Efromovich 2008 or Tsybakov 2009). The *maxiset* approach has been introduced by Cohen et al. (2001) as an alternative point of view less pessimistic than the minimax one to assess optimality.

3. Two classical estimation methods

In this section, we introduce the two main methods which are used to estimate some functions: minimum of contrast methods, based on projection on linear subspaces and kernel methods based

on convolution arguments. The M.I.S.E. of the estimators are evaluated, and the results for the two methods compared.

3.1. **Projection estimators.**

3.1.1. *Approximation and minimum contrast estimators.* Projection method is heavily based on the assumption that the true density f to estimate belongs to the Hilbert space $\mathcal{F} = (L^2(I), \|.\|, \langle\cdot,\cdot\rangle)$. The main idea is to approximate f by its orthogonal projection onto finite dimensional subspaces, called the *sieves* or the *models*. This specific terminology and the method described here have been developed by Birgé and Massart (1993, 1998). Let $S_\Lambda = \text{Span}\{\varphi_\lambda,\ \lambda \in \Lambda\}$ be a linear subset of $L^2(I)$, with $1 \leq |\Lambda| < \infty$, and $\{\varphi_\lambda,\ \lambda \in \Lambda\}$ a family of linearly independent functions on I. Then the orthogonal projection $\Pi_{S_\Lambda} f$ can be written

$$\Pi_{S_\Lambda} f = \sum_{\lambda\in\Lambda} a_\lambda \varphi_\lambda, \quad \text{with } a_\lambda = \langle f, \varphi_\lambda\rangle = \int_I f(x)\varphi_\lambda(x)dx.$$

Instead of recovering f, one can rather first estimate $\Pi_{S_\Lambda} f$, which amounts to estimate the finite family of coefficients $(a_\lambda)_{\lambda\in\Lambda}$. This momentarily reduces the problem to a parametric one! Recall that $\Pi_{S_\Lambda} f$ is also defined by

$$\Pi_{S_\Lambda} f = \arg\min_{t\in S_\Lambda} \|f - t\|^2 = \arg\min_{t\in S_\Lambda} \|t\|^2 - 2\langle t, f\rangle, \tag{6}$$

and we use this definition to estimate it. Since $\langle t, f\rangle = \mathbb{E}_f[t(X_1)]$, an empirical counterpart[1] for this quantity is $n^{-1}\sum_{i=1}^n t(X_i)$. We thus define

$$\gamma_n(t) = \|t\|^2 - \frac{2}{n}\sum_{i=1}^n t(X_i), \quad t \in L^2(I), \tag{7}$$

The function γ_n is called a *contrast function* (see Birgé and Massart 1993, p.117 or Birgé and Massart 1998, p.318) or a *least-squares contrast function* (by analogy with the least-squares contrat function that permits to estimate a regression function). We check that

$$\mathbb{E}_f[\gamma_n(t)] = \|t\|^2 - 2\langle t, s\rangle = \|t - s\|^2 - \|s\|^2.$$

Thus, by comparing this with (6), γ_n suits well to estimate $\Pi_{S_\Lambda} f$: minimising it over the linear subspace S_Λ leads to a *minimum contrast estimator* (a kind of $M-estimator$[2]) for f:

$$\hat{f}_\Lambda = \arg\min_{t\in S_\Lambda} \gamma_n(t). \tag{8}$$

The estimator $\hat{f}_\Lambda$ is uniquely defined: we compute

$$\hat{f}_\Lambda = \sum_{\lambda\in\Lambda} \hat{a}_\lambda \varphi_\lambda, \text{ with } \hat{a}_\lambda = n^{-1}\sum_{i=1}^n \varphi_\lambda(X_i).$$

Moreover, it is an *unbiased* estimator for $\Pi_{S_\Lambda} f$ in the sense that $\mathbb{E}_f[\hat{f}_\Lambda] = \Pi_{S_\Lambda} f$.

---

[1] When an unknown quantity appears, the statistician replaces it by an estimator, build from the available data. Here, $n^{-1}\sum_{i=1}^n t(X_i)$ is an unbiased estimator of $\mathbb{E}_f[t(X_1)]$ (this means that $\mathbb{E}_f[n^{-1}\sum_{i=1}^n t(X_i)] = \mathbb{E}_f[t(X_1)]$ which is also consistent (it converges almost surely to $\mathbb{E}_f[t(X_1)]$).

[2] M-estimators is a broad class of estimates, which are obtained as the minima of functions of the data. It covers the minimum contrast estimators, like here, but also the maximum likelihood estimators.

3.1.2. *Models.* The method raises the question of the definition of the model $S_\Lambda \subset (L^2(I), \|.\|)$, that is the definition of its basis $(\varphi_\lambda)_{\lambda\in\Lambda}$. The choice falls to the statistician, and we consider the following assumptions.

$(\mathcal{M}_1)$: $\dim(S_\Lambda) = |\Lambda| \leq n$.
$(\mathcal{M}_2)$: $\{\varphi_\lambda\ \lambda \in \Lambda\}$ is an orthonormal family of functions.
$(\mathcal{M}_3)$: $S_\Lambda \subset (L^\infty(I), \|.\|_\infty)$, the space of essentially bounded measurable functions on I and

$$\exists \Phi_0^2 > 0, \quad \left\| \sum_{\lambda\in\Lambda} \varphi_\lambda \right\|_\infty^2 \leq \Phi_0^2 |\Lambda|,$$

Assumption $(\mathcal{M}_1)$ is reasonable. It states that the dimension of the model is bounded by the number of observations: $|\Lambda|$ is indeed the number of coefficients of the orthogonal projection to recover (see its development in the basis $(\varphi_\lambda)_{\lambda\in\Lambda}$ above), and one cannot hope to have a good estimate if it is larger than the number of data[3]. We only assume $(\mathcal{M}_2)$ for technical purpose: it can be relaxed by only assuming that $(\varphi_\lambda)_{\lambda\in\Lambda}$ is a Riesz basis, which means

$$\exists c, C > 0, \ \ \forall (a_\lambda)_{\lambda\in\Lambda},\ c \sum_{\lambda\in\Lambda} a_\lambda^2 \leq \left\| \sum_{\lambda\in\Lambda} a_\lambda \varphi_\lambda \right\|^2 \leq C \sum_{\lambda\in\Lambda} a_\lambda^2.$$

(see Härdle et al. 1998, definition 6.1). Assumption $(\mathcal{M}_3)$ is a connection between the L^2- and the $L^\infty-$structures of the models, since it is equivalent to

$$\forall t \in S_\Lambda,\ \|t\|_\infty^2 \leq \Phi_0^2 |\Lambda| \|t\|^2,$$

see Birgé and Massart (1998), Lemma 1.

In the litterature, there exist benchmark models which satisfy these three assumptions: models based on the Fourier basis, on regular piecewise polynomials with dyadic partition, or on compactly supported wavelets. The last two models have an additional property of "localisation", which is sometimes helpful. If Riesz bases are allowed, B-spline can also be considered. For details, we refer to Birgé and Massart (1998) for general points, Härdle et al. (1998) for wavelets and DeVore and Lorentz (1993) for spline bases. The Laguerre basis is sometimes used for estimation over $I = \mathbb{R}_+$, see Belomestny et al. (2016) *e.g.* For illustration, we only consider here the trigonometric model, defined on $I = [a, b]$ by $\Lambda = \Lambda_m = \{1, \dots, D_m\}$ with $D_m = 2m+1$, $m \geq 0$ and $\varphi_1(x) = \sqrt{b-a}^{-1} \mathbf{1}_I(x)$,

$$\varphi_{2j}(x) = \frac{1}{\sqrt{b-a}} \mathbf{1}_I(x) \sqrt{2} \cos\left(2\pi j \frac{x-a}{b-a}\right) \quad \varphi_{2j+1}(x) = \frac{1}{\sqrt{b-a}} \mathbf{1}_I(x) \sqrt{2} \sin\left(2\pi j \frac{x-a}{b-a}\right),$$

for $j = 1, \dots, m$. The models spanned by this basis are nested: if $m \leq m'$, $S_{\Lambda_m} \subset S_{\Lambda_{m'}}$.

The crucial point for projection estimators is that the model has good approximation properties: it is clear that the estimation method fails if $\Pi_{S_\Lambda} f$ is far from f. Lemma 12 from Barron et al. (1999) established that for reasonable wavelets, piecewise polynomials and for the trigonometric basis defined above,

$$\|f - \Pi_{S_\Lambda} f\| \leq C |\Lambda|^{-\alpha} \tag{9}$$

[3]In this paper, we do not consider the framework of *high dimensional statistics* ("big data" analysis), which deal with the special case $|\Lambda| >> n$ in this context and required specific methods.

if f belongs to a ball of the Besov space $\mathcal{B}^{\alpha}_{2,\infty}(I)$. As explained in Section 2.2.2, we do not give a detailed definition of such functional spaces (that are interpolation spaces lying between Sobolev spaces): the reader should just keep in mind that $\alpha > 0$ is a measure of the smoothness of f. Analogous results exist for spline and Laguerre bases.

3.1.3. *Upper-bound for the risk.* We have heuristically explained the definition of the projection estimator $\hat{f}_\Lambda$ for f, see (8). Let us study its risk. Thanks to the Pythagoras theorem, the M.I.S.E. is splitted into two terms

$$\mathcal{R}\left(\hat{f}_\Lambda, f\right) = \mathbb{E}_f\left[\left\|\hat{f}_\Lambda - f\right\|^2\right] = \|f - \Pi_{S_\Lambda} f\|^2 + \mathbb{E}_f\left[\left\|\hat{f}_\Lambda - \Pi_{S_\Lambda} f\right\|^2\right]. \tag{10}$$

The first term is the *squared-bias term* or the *approximation error*. The second one is the *variance term* of the risk or the *stochastic error*, and can be bounded as follows

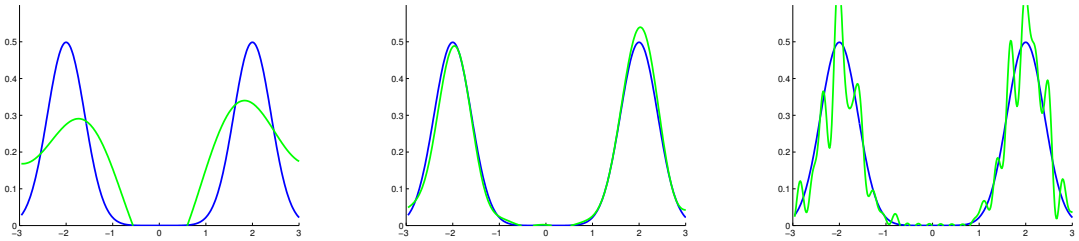
$$\begin{aligned}
\mathbb{E}_f\left[\left\|\hat{f}_\Lambda - \Pi_{S_\Lambda} f\right\|^2\right] &= \mathbb{E}_f\left[\left\|\sum_{\lambda\in\Lambda}(\hat{a}_\lambda - a_\lambda)\varphi_\lambda\right\|^2\right], \\
&= \sum_{\lambda\in\Lambda}\mathrm{Var}_f(\hat{a}_\lambda) = \frac{1}{n}\sum_{\lambda\in\Lambda}\mathrm{Var}_f\left(\varphi_\lambda(X_1)\right) \\
&\leq \frac{1}{n}\sum_{\lambda\in\Lambda}\mathbb{E}_f\left[\varphi_\lambda^2(X_1)\right] \leq \Phi_0^2\frac{|\Lambda|}{n},
\end{aligned}$$

thanks to assumptions $(\mathcal{M}_2)$ and $(\mathcal{M}_3)$. This leads to

$$\mathcal{R}\left(\hat{f}_\Lambda, f\right) \leq \|f - \Pi_{S_\Lambda} f\|^2 + \Phi_0^2\frac{|\Lambda|}{n}. \tag{11}$$

The bias and the variance terms of the risk have thus opposite behaviours with respect to the dimension $|\Lambda|$ of the model: the bias term decreases when $|\Lambda|$ increases (the larger the model, the better the approximation) while the variance term increases with $|\Lambda|$ (since the number of estimated coefficients grows with $|\Lambda|$): in this case, S_Λ is likely to *overfit*. A compromise is thus required to minimise the risk: the largest model is not the best one for estimation purpose! It is the so-called *bias-variance trade-off*, which is illustrated in Figure 1. To implement the estimator, we calibrate the estimation interval $I = [a; b]$ with the data $a = \min\{X_i, i = 1, \ldots, n\}$ and $b = \max\{X_i, i = 1, \ldots, n\}$.

3.1.4. *Towards model selection.* The upper-bound (11) for the risk of a projection estimate shows that the model S_Λ, and precisely its dimension $\dim(S_\Lambda) = |\Lambda|$ play a crucial role in the procedure. It is thus natural to consider a collection of models $(S_{\Lambda_m})_{m\in\mathcal{M}_n}$ with $\mathcal{M}_n \subset \mathbb{N}\backslash\{0\}$ a finite collection of indices (which cardinality may depend on the data sample size n, see the assumptions of Theorem 1). Thus, there are a collection $(\hat{f}_{\Lambda_m})_{m\in\mathcal{M}_n}$ of estimates for f. For our simulation example (see (1)), they are plotted in Figure 2. The statistician has to choose the "best" one in the collection ("best", in the sense of the quality criterion chosen, here the M.I.S.E.). To simplify the notations and exposition, from now on, we assume that the model collection includes only one model per dimension. For any $m \in \mathcal{M}_n$ S_{Λ_m} is thus be denoted by S_m, $\hat{f}_{\Lambda_m} = \hat{f}_m$, and we set $|\Lambda_m| := D_m$.



$|\Lambda| = 7$ $|\Lambda| = 16$ $|\Lambda| = 51$

FIGURE 1. Projection estimators (in the Fourier basis) for f_{Simu}, computed with $n = 500$ observations, for three choices of model dimension D_{Λ_m}. Bold blue line: true function f_{Simu}. Green line: estimator.

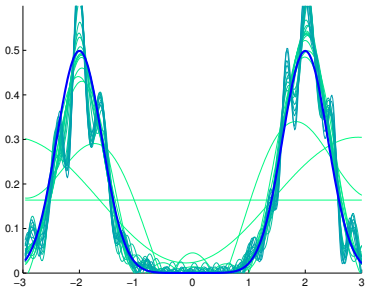


FIGURE 2. Collection of projection estimators (in the Fourier basis) for f_{Simu}, computed with $n = 500$ observations, for different model dimensions. Bold blue line: true function f_{Simu}. Thin lines: estimators $\hat{f}_m$, $m = 1, 3, 5, \ldots, 51$.

The best model in the collection has index m^* satisfying

$$m^* = \arg \min_{m \in \mathcal{M}_n} \mathcal{R}(\hat{f}_m, f) \tag{12}$$

It is called the *oracle* (an "oracle" who knows in advance the collection of the risks - the terminology has been introduced by Donoho and Johnstone 1994), and is not available, since it depends on the true function f to recover: it is not an estimator.

If the density f belongs to a ball of the Besov space $\mathcal{B}^{\alpha}_{2,\infty}(I)$, then the bias term in Equation (11) is upper-bounded by $D_m^{-2\alpha}$, see (9). If the index α is known, model selection is easy to perform: we choose the model $S_{m(\alpha)}$ such that $m(\alpha) = \arg\min_{m \in \mathcal{M}_n} \{D_m^{-2\alpha} + \Phi_0^2 D_m/n\}$. This leads to a dimension $D_{m(\alpha)}$ of order $n^{1/(2\alpha+1)}$, and a convergence rate $n^{-2\alpha/(2\alpha+1)}$ for the maximal risk of $\hat{f}_{m(\alpha)}$ in the sense defined above (see Section 2.2.2). This proves that projection estimates might have good behaviour, since this rate is the minimax one for density estimation from an *i.i.d.* sample (see again Section 2.2.2) However, if f is unknown, α is probably unknown too. The challenge of *adaptive estimation* is to perform *model selection* in a data-driven way, this is the goal of Section 4.

3.2. **Kernel estimators.** Kernel estimators is the second family of function estimators. For simplicity, we consider $I = \mathbb{R}$ in this section (if this is not the case, one can replace f by $f\mathbf{1}_I$).

3.2.1. *Kernel and approximation.* A *kernel* is an integrable function $K : \mathbb{R} \to \mathbb{R}$ which satisfies $\int_{\mathbb{R}} K(u)du = 1$. For any real-number $h > 0$, let $K_h : u \in \mathbb{R} \mapsto K(u/h)/h$. The basic property which makes kernel interesting for estimation purpose is the following: the family $(K_h)_{h\geq 0}$ is an approximate identity for the convolution product. This means that the convolution $K_h \star f : x \mapsto \int_{\mathbb{R}} K_h(x-x')f(x')dx'$ goes to f (in $L^2(\mathbb{R})$) when h goes to zero, and the convergence rate is all the faster that f is smooth: if f belongs to a ball of a Nikol'skiĭ space $\mathcal{N}_2^\alpha(\mathbb{R})$ (Nikol′skiĭ, 1975) then, for any $h > 0$,

$$\|K_h \star f - f\|_2 \leq Ch^\alpha, \tag{13}$$

for a constant $C > 0$ which does not depend on the parameter h, as soon as the kernel K satisfies $\int_{\mathbb{R}} |x|^\alpha |K(x)|dx < +\infty$ and has *order* $l = \lfloor \alpha \rfloor$ ($\lfloor \alpha \rfloor$ is the greatest integer strictly less than α): this means that for any $j = 0, \ldots, l$, the functions $x \mapsto x^j K(x)$ are integrable and $\int_{\mathbb{R}} x^j K(x)dx = 0$. A proof of (13) can be found in Proposition 1.5 of Tsybakov (2009).

The true density f can thus be approximated by $K_h \star f$, which satisfies $K_h \star f(x) = \mathbb{E}_f[K_h(x - X_1)]$. The *kernel estimator* for f with fixed parameter $h > 0$ is thus the empirical counterpart of the expectation

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i), \quad x \in \mathbb{R}. \tag{14}$$

It has been introduced by Rosenblatt (1956), for the "rectangular" kernel $K = \mathbf{1}_{]-1;1[}/2$: in this case, it can be seen has a kind of "derivative" of the empirical cumulative distribution function (see Tsybakov 2009 for details). Then Parzen (1962) has generalised the definition for any kernel function. Tsybakov (2009) has listed six usual kernels that are plotted in Figure 3. Beta kernels can also be considered, see Bertin and Klutchnikoff (2011).

Figure 4 shows the functions K_h when h is getting smaller (close to 0), for the rectangular kernel and the Gaussian one. The parameter h in Definition (14) is a smoothing parameter called the *bandwidth*: the main challenge is to choose a good value for it (see sections 3.2.2 and 4.3). Notice also that kernels of a given order l, as defined above, can be built with at least two methods. A first construction is proposed by Kerkyacharian et al. (2001) (see also Comte 2015, p.53) and a second way to build them is to take advantage of the Legendre polynomials, see Tsybakov (2009), p.10.

3.2.2. *Risk and bandwidth selection problem.* The decomposition of the M.I.S.E. of the kernel estimate is similar to that of projection estimator. Keeping in mind that $\mathbb{E}_f[\hat{f}_h] = K_h \star f$, the analogous of (10) is

$$\mathcal{R}\left(\hat{f}_h, f\right) = \mathbb{E}_f\left[\left\|\hat{f}_h - f\right\|^2\right] = \|f - K_h \star f\|^2 + \mathbb{E}_f\left[\left\|\hat{f}_h - K_h \star f\right\|^2\right].$$

Since $\|K_h\|^2 = \|K\|^2/h$, we obtain for the variance term of the risk

$$\mathbb{E}_f\left[\left\|\hat{f}_h - K_h \star f\right\|^2\right] = \int_{\mathbb{R}} \mathrm{Var}_f(\hat{f}_h(x))dx \leq \frac{1}{n}\int_{\mathbb{R}} \mathbb{E}_f\left[K_h^2(x - X_1)\right]dx \leq \frac{\|K\|^2}{nh},$$

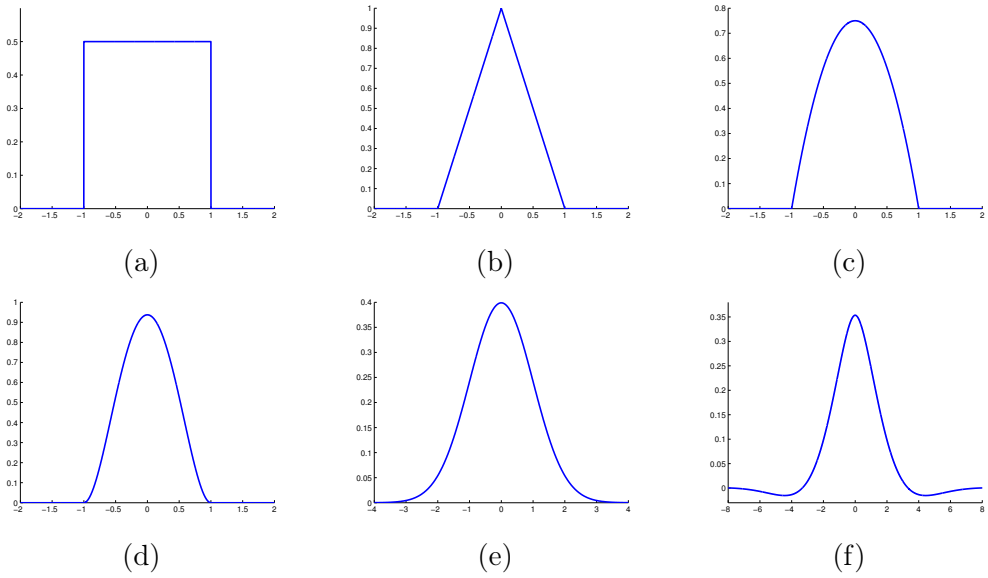


FIGURE 3. Usual kernels in statistics. (a) rectangular kernel, (b) triangular kernel, (c) Epanechnikov kernel, (d) "biweight" kernel, (e) Gaussian kernel, (f) Silverman kernel.

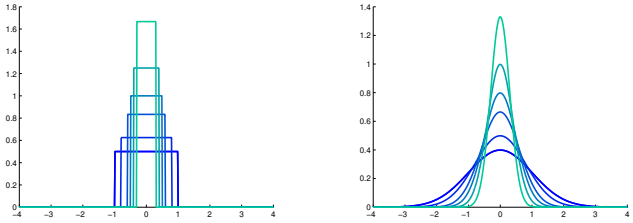
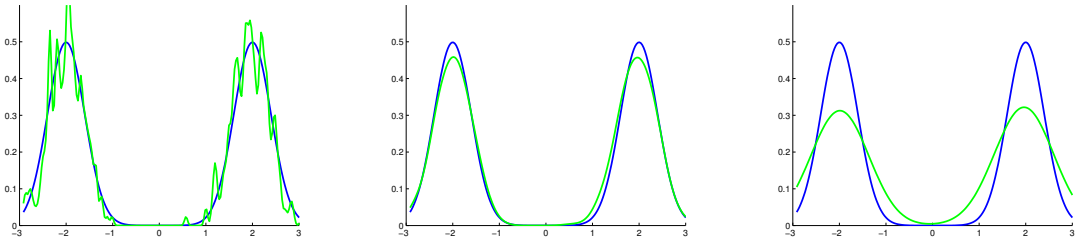


FIGURE 4. Examples of functions $(K_h)_{h>0}$, which form an approximate identity for the convolution product when h goes to zero.

which leads to

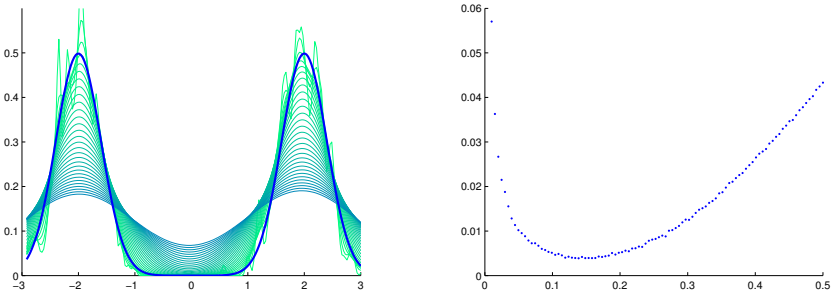
$$\mathcal{R}\left(\hat{f}_h, f\right) \leq \|f - K_h \star f\|^2 + \|K\|^2 \frac{1}{nh}. \tag{15}$$

Here again the two terms of the upper-bound must be balanced, to minimise the risk: the bias term goes to zero with the bandwidth h (and thus is too large if h is large), while the variance term explodes when h is too small (overfitting), see Figures 5 and 6. In the last one, the M.I.S.E., defined as an expectation (see (2)), is approximated by a Monte-Carlo method. It is obtained by averaging the following approximations ISE_j of the Integrated Squared Error, for $j \in \{1 \ldots, J = 100\}$, computed with $J = 100$ replications:



$h = 0.033$ $h = 0.04$ $h = 1$

FIGURE 5. Kernel estimators (Gaussian kernel) for f_{Simu} (defined by (1)), computed with $n = 500$ observations, for three choices of bandwidth h. Bold blue line: true function f_{Simu}. Green line: estimator.



(a) (b)

FIGURE 6. (a) Collection of kernel estimators (Gaussian kernel) for f_{Simu}, computed with $n = 500$ observations, for different bandwidths. Bold blue line: true function f_{Simu}. Thin lines: estimators $\hat{f}_h$, $h = 1/30, 2/30, \ldots, 1$. (b) Plot of the M.I.S.E. $\mathcal{R}(\hat{f}_h, f_{\text{Simu}})$, with $\hat{f}_h$ computed from $n = 500$ observations, as a function of the bandwidth h.

$$ISE_j = \frac{b-a}{N} \sum_{k=1}^{N} \left(\hat{f}_h^{(j)}(x_k) - f_{\text{Simu}}(x_k) \right)^2, \tag{16}$$

where the $(x_k)_k$ are a regular grid of $N = 50$ points over $[a, b]$, where $a = \min\{X_i^{(j)}, i = 1, \ldots, n\}$ and $b = \max\{X_i^{(j)}, i = 1, \ldots, n\}$ for the $j-$th simulated sample $(X_i^{(j)})_{i=1,\ldots,n}$.

Notice that the choice of the kernel may also be discussed: the Epanechnikov kernel has been shown to be an optimal choice in some cases (minimisation of the asymptotic M.I.S.E. over nonnegative kernels) but can also be considered as "inadmissible" for other criteria and the requirement $K \geq 0$ might be dropped (a clear discussion can be found in Sections 1.2 and 1.3 of Tsybakov 2009). We would not address this issue here: for an introduction, we may say that the choice of K is less crucial for the quality of $\hat{f}_h$ as an estimator of f than the choice of h.

Thus, the kernel estimator might be a good estimator, if its bandwidth is carefully chosen: the M.I.S.E. is small only if both the variance and the squared bias term are small. Given a

finite collection $\mathcal{H}_n$ of possible bandwidths, an oracle can be defined, like above for the model, but cannot be computed from the data.

If f belongs to a ball of a Nikol'skiĭ space $\mathcal{N}_2^\alpha(\mathbb{R})$ and if the kernel has order $l = \lfloor \alpha \rfloor$ (with $\int_{\mathbb{R}} |x|^\alpha |K(x)| dx < +\infty$), gathering (13) and (15) permits to deduce that

$$\mathcal{R}\left(\hat{f}_{h(\alpha)}, f\right) := \min_{h \in \mathcal{H}_n} \mathcal{R}\left(\hat{f}_h, f\right) \leq C \min_{h \in \mathcal{H}_n} \left\{ h^{-2\alpha} + \frac{1}{nh} \right\} = C n^{\frac{-2\alpha}{\alpha+1}}.$$

The rate of decrease of kernel estimators can thus be the minimax one. However, the optimal $h(\alpha)$ is not an estimator, since α is unknown. Like model selection, the problem is to derive a data-driven procedure which permits to automatically choose h in the collection.

4. Adaptive methods

4.1. **Introduction to adaptation.**

4.1.1. *Main issue.* Projection and kernel methods for density estimation have now been introduced. Both of the methods provide estimates which depends on a smoothing parameter. The role of the model dimension for projection estimators can be compared to the role of (the inverse of) the bandwidth for kernel estimates. The framework can thus be summed up as follows: in any case, we have defined a finite collection $(\hat{f}_b)_{b \in \mathcal{B}_n}$ of estimators for f which depends on a smoothing parameter b, $b = D_m$ for projection estimators, $b = 1/h$ for kernel estimators. For any estimator of the two collections, we have proved that

$$\mathcal{R}(\hat{f}_b, f) \leq \left\| \mathbb{E}_f[\hat{f}_b] - f \right\|^2 + c\frac{b}{n}, \quad b \in \mathcal{B}_n, \tag{17}$$

see (11) and (15). The best function (for the M.I.S.E.) in the collection is not an estimator, it is the so-called oracle (see also (12)),

$$b^* = \arg\min_{b \in \mathcal{B}_n} \mathcal{R}\left(\hat{f}_b, f\right). \tag{18}$$

We have shown that the minimax rate $n^{-2\alpha/(2\alpha+1)}$ for the estimation of a density with smoothness index α can be achieved for one of the function of each collection, by choosing a parameter b which depends on α, see the end of sections 3.1.4 and 3.2.2.

The problem that we now want to address is the following. Starting from the collection $(\hat{f}_b)_{b \in \mathcal{B}_n}$, how can we build an estimator that achieves the same optimal rate for functions of smoothness α, but in a data-driven way? Its definition should not depend on the smoothness index α of the target function f to estimate. Such an estimator is said to be *adaptive*. It realizes the best bias-variance compromise. Moreover, it makes possible *adaptation* to the unknown α. We focus below (sections 4.2 and 4.3 on methods for which nonasymptotic theoretical results can be provided: we do not assume that the sample size n tends to infinity while all other parameters of the problem stay fixed. Thus, the family $(\hat{f}_b)_{b \in \mathcal{B}_n}$ may vary with n: when more data are available, it can be reasonable to assume that more estimators can be considered. While asymptotic results sometimes hide inside $o(\cdot)$ some terms that can modify the behaviour of the methods in practice, in the nonasymptotic approach, all parameters can appear explicitely in the bounds, even if the idea is not to analyse very small samples.

4.1.2. *Expected results.* A distinction may be drawn between two kinds of methods that are proved (or that will be proved) to be adaptive: *aggregation* or *estimator selection*.

The aim of *aggregation*, initiated by Nemirovski (2000), is to combine the estimators $(\hat{f}_b)_{b\in\mathcal{B}_n}$ of the collection, called a *dictionnary*, to define a new estimator. Typically, the new estimator is a linear combination of the previous ones with form $\tilde{f} = \sum_{b\in\mathcal{B}_n} \theta_b \hat{f}_b$, and the coefficients $(\theta_b)_b$ are selected from the data (optimisation of constraint problems). Comprehensive descriptions of the method can be found in Rigollet and Tsybakov (2007); Tsybakov (2008).

When the estimators of the collection $(\hat{f}_b)_{b\in\mathcal{B}_n}$ are all similar (all kernel estimates, or all projection estimates for example), the objective is to choose "the best" of them, that is to select $\hat{b} \in \mathcal{B}_n$ such that $\mathcal{R}(\hat{f}_{\hat{b}}, f)$ is as small as possible: $\hat{f}_{\hat{b}}$ should mimic the oracle $\hat{f}_{b^*}$, which means that it should satisfy an *oracle inequality*[4]

$$\mathcal{R}\left(\hat{f}_{\hat{b}}, f\right) \leq C \inf_{b\in\mathcal{B}_n} \mathcal{R}\left(\hat{f}_b, f\right) + R_n.$$

The leading constant is $C \geq 1$, and R_n is a remainder term, negligible in front of $\inf_{b\in\mathcal{B}_n} \mathcal{R}(\hat{f}_b, f)$. The closer to 1 the constant C, the better the inequality: the inequality is said to be *sharp* if $C = 1+\delta_n$, with $\delta_n \to 0$ as $n \to \infty$. With the methods described below, we prove slightly weaker results, called *oracle-type inequality*,

$$\mathcal{R}\left(\hat{f}_{\hat{b}}, f\right) \leq C \inf_{b\in\mathcal{B}_n} \left\{ \left\| \mathbb{E}_f\left[\hat{f}_b\right] - f \right\|^2 + c\frac{b}{n} \right\} + R_n, \tag{19}$$

where c is the constant involved in (17). An estimator which satisfies an oracle-type inequality is an estimator that achieves the best bias-variance trade-off. It avoids both *overfitting* (large variance term in the risk, estimators which follow the data too closely) and *underfitting* (large bias term, too simple estimators) that can occur if the model is not well chosen, as it can be seen in Figures 1 and 5. This is the main challenge of estimator selection, and this permits to build *minimax adaptive estimators*, provided the family $(\hat{f}_b)_{b\in\mathcal{B}_n}$ is well-chosen, which is the case here (for kernel and projection methods): in a second step, by assuming that f belongs to a functional space of smooth functions, one obtain the best convergence rate in the collection $(\hat{f}_b)_{b\in\mathcal{B}_n}$ by computing the right-hand-side of (19). This best rate is the minimax one if the collection is the kernel or the projection ones described above.

4.1.3. *Brief overview of the methods.* Several methods of estimator selection have been investigated, from practical and/or theoretical purposes: *coefficient thresholding* for wavelet projection estimators, *cross-validation*, *model selection via penalisation*, *Lepski's methods* for bandwidth selection... For wavelet thresholding, we refer to Donoho and Johnstone (1994) or Härdle et al. (1998): starting from a collection of projection estimates (in a wavelet basis), it is about finding the coefficients that are interesting to keep to define the final estimates: too small coefficients are suppressed by introducing a threshold. The last three methods are based on the same following principle. The ideal selection would be the oracle (18), which depends on the unknown underlying distribution, through the computation of the risk involved in (18). Validation, penalisation

[4]Here, we consider oracle inequalities which hold in expectation. An other possibility is to prove oracle inequalities of form $\|\hat{f}_{\hat{b}} - f\|^2 \leq C \inf_{b\in\mathcal{B}_n} \|\hat{f}_b - f\|^2 + R_n$, which holds with large probability (that is, a probability larger than $1 - \varepsilon(n)$, with $\lim_{n\to+\infty} \varepsilon(n) = 0$).

and Lepski's method replace this unknown risk by an empirical criterion denoted by Crit in the sequel, and propose to select

$$\hat{b} \in \arg\min_{b\in\mathcal{B}_n} \mathrm{Crit}(b). \tag{20}$$

Cross-validation is a classical method introduced by Allen (1974); Stone (1974); Geisser (1975). The idea is to split the data in two subsets, a training set $(X_i)_{i\in E}$ from which the different estimators of the collection are computed, and a validation set $(X_i)_{i\in E^c}$ (where $E \subset \{1,\dots,n\}$ and $E^c = \{1,\dots,n\}\backslash E$), which permits to define Crit and estimates the risk of each of the estimators. A collection $\mathcal{E}$ of training sets E is generally used to repeat the procedure: depending on $\mathcal{E}$, we refer to hold-out, leave-p-out, Monte-Carlo or V-fold cross-validation estimators. A comprehensive overview has been written by Arlot and Celisse (2010). Although the asymptotic properties of cross-validation estimators have been widely studied, very few nonasymptotic results exist in the litterature: earlier bounds have been obtained by Arlot (2008), recently extended by Arlot and Lerasle (2016) and Arlot et al. (2015). The following two sections focus on penalisation and Lespki's methods.

4.2. **Model selection for projection estimators.** Model selection theory originates in the works of Akaike (1973) and Mallows (1973), and has been formalised by Birgé and Massart (1997) and Barron et al. (1999) (see also Massart 2007).

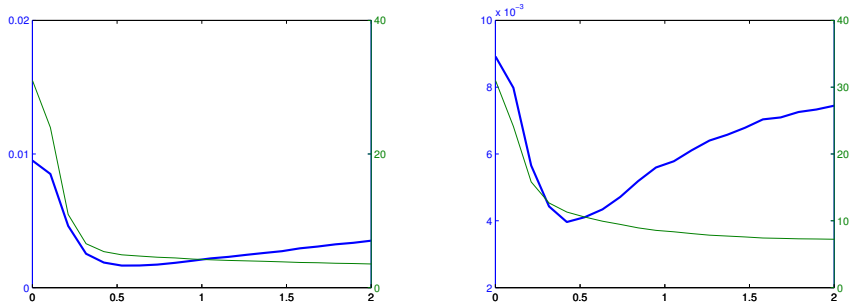
Considering the family of estimators $(\hat{f}_m)_{m\in\mathcal{M}_n}$ defined by (8), the issue is the choice of m from the data. The optimal m is the oracle m^* that minimises $\mathcal{R}(\hat{f}_m, f) = \mathbb{E}_f[\|\hat{f}_m - f\|^2]$, see (18). Since the contrast γ_n introduced in (7) is an empirical equivalent for the loss function $\|\cdot\|$ involved in the risk $\mathcal{R}$, one can be tempted to select $\hat{m}$ that minimises $\gamma_n(\hat{f}_m)$ over all possible m. However, one can see that the quantity $\mathbb{E}_f[\gamma_n(\hat{f}_m)] = \mathbb{E}_f[-\|\hat{f}_m\|^2]$ underestimates the loss $\|f - \hat{f}_m\|^2$, and need to be corrected. Assume for a moment that the models are nested: $m \le m' \;\Rightarrow\; S_m \subset S_{m'}$. Then, if $m \le m'$, $\hat{f}_m \in S_{m'}$, and $\gamma_n(\hat{f}_m) \le \gamma_n(\hat{f}_{m'})$. In this case, $m \mapsto \gamma_n(\hat{f}_m)$ decreases with m, and thus with the dimension $|\Lambda_m|$. Coming back to the general framework, this justifies the introduction of a penalty function $\mathrm{pen} : \mathcal{M}_n \to \mathbb{R}_+$ which measures the complexity of the model S_m, and the selection

$$\hat{m} = \arg\min_{m\in\mathcal{M}_n} \mathrm{Crit}_{BM}(m), \text{ with } \mathrm{Crit}_{BM}(m) = \gamma_n(\hat{f}_m) + \mathrm{pen}(m). \tag{21}$$

Here, the penalty only depends on the dimension of each model, since there is a uniq model per dimension: an appropriate choice is

$$\mathrm{pen}(m) = \kappa\Phi_0^2\frac{D_m}{n}, \tag{22}$$

for a constant $\kappa > 0$. If the model collection is more complicated, the penalty should depend on a measure of the "complexity" of the collection. The order of magnitude of the penalty is also heuristically justified as follows: the criterion to minimise, $\mathrm{Crit}_{BM}(m)$, estimates the risk which is splitted in a bias term and a variance term. A (biased) estimator for the bias $\|\Pi_{S_m}f - f\|^2$ is $-\|\hat{f}_m\|^2 = \gamma_n(\hat{f}_m)$. Indeed, $\|f - \Pi_{S_m}f\|^2 = \|f\|^2 - \|\Pi_{S_m}f\|^2$, with $\|f\|$ independent on m, and $\Pi_{S_m}f$ is estimated by $\hat{f}_m$. Thus, the penalty should estimate the variance term of the risk, and has thus its order (see(11)). The following result can be proved for the *penalised contrast estimator* $\hat{f}_{\hat{m}}$, see *e.g.* Theorem 5.2 in Comte (2015).



(a) (b)

FIGURE 7. Plot of the M.I.S.E. $\mathcal{R}(\hat{f}_{\hat{m}}, f)$ (averaged over 100 samples) (labeling on the left of each graph) and the selected model dimension $D_{\hat{m}}$ (labeling on the right), computed from $n = 1000$ observations, with respect to the value of the constant κ (axis label). (a) f density of the standard Gaussian distribution $\mathcal{N}(0, 1)$. (*b*) $f = f_{\text{Simu}}$ defined in (1). Bold blue line: M.I.S.E. Green thin line: selected model dimension.

Theorem 1. *Suppose that the cardinality of the collection $\mathcal{M}_n$ is bounded by n, and that the models S_m, $m \in \mathcal{M}_n$ satisfy Assumptions $(\mathcal{M}_l)$, $l = 1, 2, 3$, and are nested. Assume also that the true density f is bounded. Then, there exists some constant $\kappa > 0$ such that*

$$\mathcal{R}\left(\hat{f}_{\hat{m}}, f\right) \leq C \inf_{m \in \mathcal{M}_n} \left\{ \left\| \mathbb{E}_f\left[\hat{f}_m\right] - f \right\|^2 + \text{pen}(m) \right\} + \frac{C'}{n},$$

for $C, C' > 0$, C a numerical constant and C' depending on f and Φ_0^2.

This is the oracle type inequality announced in (19): $\hat{f}_{\hat{m}}$ performs as well as the best estimator in the collection, up to the multiplicative constant C, and up to a remainding term of order $1/n$, which is negligible. The implementation of the method raises the question of the tuning of the constant κ involved in (22). From the theoretical point of view, this is a universal constant in the sense that it does not depend on the model parameters or on the estimation parameters. A lower bound is obtained in the proof: it is unfortunately very rough and useless in practice. However, the choice is crucial for the quality of estimation. If κ is too small, the most influencial term in (21) is $\gamma_n(\hat{f}_m)$, and large models are selected. If κ is too large, the reverse occurs: models with too low dimension are selected. The problem of optimal/minimal calibration of the penalties has aroused considerable interest: the first results were obtained by Birgé and Massart (2007). A data-driven procedure, the *slope heuristic*, exists, and has been implemented by Baudry et al. (2012) in a package called C.A.P.U.S.HE (both for MatLab and R). We can also decide to calibrate κ once and for all: the risk of the selected estimator (obtained from simulated data) can be plotted with respect to the value of the constants, and a value leading to reasonable risk and complexity of the selected model can be chosen. Examples are plotted in Figure 7. We choose $\kappa = 0.25$. It should be kept in mind that is more secure to choose the constant too large than too small, since small penalties lead to explosive risks. Figure 8 displays a collection of estimators $(\hat{f}_m)_{m \in \mathcal{M}_n}$ and the selected one $\hat{f}_{\hat{m}}$.

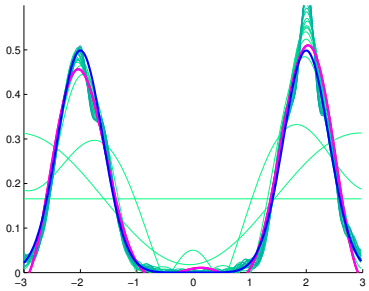


FIGURE 8. Collection of projection estimators (in the Fourier basis) for f_{Simu}, computed with $n = 500$ observations, for different model dimensions and selected estimator $\hat{f}_{\hat{m}}$ obtained with the penalisation method. Bold blue line: true function f_{Simu}. Thin lines: estimators $\hat{f}_m$, $m = 1, 3, 5, \dots, 51$. Bold pink line: $\hat{f}_{\hat{m}}$.

Model selection via penalisation is specifically designed for estimation by contrast minimisation and for the study of quadratic risks. The proof of Theorem 1 is based on the study of the concentration of the supremum of the squared of an empirical process around its mean, and required an order between the indices m of the collection. This makes the method sometimes difficult to extend for multivariate function estimation: dealing with anisotropic functions would necessitate different indices of the models in the different directions, *e.g.* models indexed by two indices $m = (m_1, m_2)$ when estimating a function with two variables. But an order between such m and m' is not easily defined. This is a motivation to describe the Goldenshluger-Lepski method.

4.3. **The Goldenshluger-Lepski method.** Lepski's methods have been introduced to select among kernel estimators with different bandwidths. The main idea is to estimate the bias term of the risk of the estimators by pairwise comparison of the estimators with fixed bandwidths. The procedure originates in earlier works of Lepskiĭ (1991, 1992a,b). We focus on a recent version, which aims at handling the possible anisotropy of multivariate functions. It was first used in the white noise model (Goldenshluger and Lepski, 2008, 2009), then for multivariate density estimation (Goldenshluger and Lepski, 2011) and general frameworks (Goldenshluger and Lepski, 2013).

We present it in our simpler problem of univariate density estimation with quadratic risk to select an estimator among $(\hat{f}_h)_{h\in\mathcal{H}_n}$ (see (14)), keeping in mind that it can be used in many models and for various risks (pointwise, L^p...). Moreover, contrary to (Goldenshluger and Lepski, 2011), we do not consider the case where $\mathcal{H}_n$ is an interval and restrict ourselves to a finite collection, which is more reasonable from the practical point of view. These constraints permit to derive theoretical results for the Goldenshluger-Lepski method we describe below (Section 4.3) through the usual tools of model selection (mainly concentration of empirical processes).

The starting point is the same as for model selection: we want to automatically choose a bandwidth $\hat{h} \in \mathcal{H}_n$ such that $\hat{f}_{\hat{h}}$ mimics the oracle. Since the oracle minimises the risk over all possible estimators (see (18)), and since the risk is upper-bounded by the sum of the stochastic and the approximation errors (see (15)), one can define empirical counterparts for these two terms, and select the bandwidth which minimises the sum of the two empirical terms. This leads

to

$$\hat{h} = \arg\min_{h\in\mathcal{H}_n} \mathrm{Crit}_{GL}(h), \text{ with } \mathrm{Crit}_{GL}(h) = A(h) + V(h), \tag{23}$$

where V is the analogous of the penalty term (22) above, which estimates the stochastic error, and A is the counterpart for the bias term in (15). The definitions, in the spirit of Goldenshluger and Lepski (2011), are the following:

$$V(h) = \kappa' \frac{\|K\|_1^2 \|K\|^2}{nh}, \quad A(h) = \max_{h'\in\mathcal{H}_n} \left(\left\| \hat{f}_{h'} - \hat{f}_{h,h'} \right\|^2 - V(h') \right)_+, \tag{24}$$

where $\|K\|_1^2 = \int_{\mathbb{R}} |K(u)|du$, $\kappa' > 0$ is a constant to be calibrated (like κ in (22)), $x_+ = \max(x, 0)$ is the positive part of x, and $\hat{f}_{h,h'}$ are oversmoothed auxiliary estimators. For density estimation, one can choose $\hat{f}_{h,h'} = K_h \star \hat{f}_{h'}$, but in other frameworks, $\hat{f}_{h,h'} = \hat{f}_{h \sup h'}$ can be more suitable. The specific feature of the method is the estimation of the bias $\|\mathbb{E}_f[\hat{f}_h] - f\|^2$ by $A(h)$. Let us give a short heuristic: since the bias $\|\mathbb{E}_f[\hat{f}_h] - f\|^2$ is equal to $\|K_h \star f - f\|^2$, where f is unknown, we replace it by an estimator with fixed bandwidth $\hat{f}_{h'}$. But such a method introduces variability, which can be canceled by substracting $V(h')$. We thus obtain $(\|\hat{f}_{h'} - \hat{f}_{h,h'}\|^2 - V(h'))_+$ (since the bias is nonnegative). The last step is to remark that we have no reason to choose a h' or another: this justifies the "max" in (24). Obviously, a full proof is required to show that $A(h) \leq C\|\mathbb{E}_f[\hat{f}_h] - f\|^2 + C/n$ for a constant C. The reader may refer to Comte (2015), p.60. The complete result, also proved p.60, can now be stated.

Theorem 2. *Suppose that the cardinality of $\mathcal{H}_n$ is bounded by n, that for any $h \in \mathcal{H}_n$, $h \geq 1/n$, and that $\sum_{h\in\mathcal{H}_n} h^{-1} \leq c_0 n$, for a constant c_0. Assume also that the true density f is bounded. Then, there exists some constant $\kappa' > 0$ such that*

$$\mathcal{R}\left(\hat{f}_{\hat{h}}, f\right) \leq C \inf_{h\in\mathbb{H}_n} \left\{ \left\| \mathbb{E}_f\left[\hat{f}_h\right] - f \right\|^2 + V(h) \right\} + \frac{C'}{n},$$

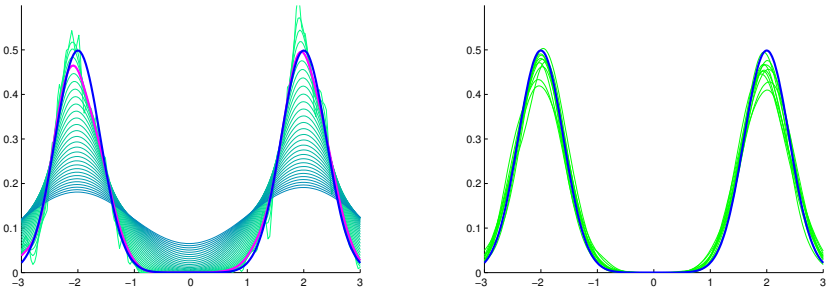
for $C, C' > 0$, C depending only on $\int_{\mathbb{R}} |K(u)|du$, and C' depending on f, $\int_{\mathbb{R}} |K(u)|du$, and $\|K\|$.

The assumptions on the bandwidth collection $\mathcal{H}_n$ are very mild. For example, there are satisfied by the following two collections:

$$\mathcal{H}_{n,1} = \left\{ 2^{-k},\ k = 1, \ldots, [\log_2(n)] \right\}, \quad \mathcal{H}_{n,2} = \left\{ k^{-1},\ k = 1, \ldots, [\sqrt{n}] \right\},$$

with $c_0 = 2$ and $c_0 = 1$ respectively. Until recently, no systematic study had been undertaken to tune the constant κ' involved in the penalty term V. We proceed as for the constant κ of the penalisation method above, see Section 4.2, and choose $\kappa' = 1$. A recent study of Lacour and Massart (2016) is devoted to the problem. It is shown that the procedure fails if κ' is chosen smaller than some critical value which leads to a minimal penalty, like for model selection. A numerical result of the selection method is plotted in Figure 9.

4.4. **Comparison of the methods.** The strength of Lepski-type methods is based on their ability to be applied for several risks and several estimation problems (references can be founded in the introduction of Lacour and Massart 2016, for example), while projection methods are mainly specific to the quadratic risk. The idea of using pairwise comparison of estimators is not only worthwile for bandwidth selection purpose but also for model selection: earlier references are Laurent et al. (2008) who proposed an adaptation of the penalisation method for linear functional,



(a) (b)

FIGURE 9. (a) Collection of kernel estimators (Gaussian) for f_{Simu}, computed with $n = 500$ observations, for different bandwidth, and selected estimator $\hat{f}_{\hat{h}}$ obtained with the Goldenshluger-Lepski method. Bold blue line: true function f_{Simu}. Thin lines: estimators $\hat{f}_h$, $h = 1/30, 2/30, \dots, 1$. Bold pink line: $\hat{f}_{\hat{h}}$. (b) Beams of estimators $\hat{f}_{\hat{h}}$, computed from independent samples of size $n = 500$. Bold blue line: true function f_{Simu}. Thin green lines: estimators $\hat{f}_{\hat{h}}$.

and Plancade (2009) whose objective is pointwise model selection. More recent examples are Comte and Johannes (2012); Chagny (2013); Bertin et al. (2016), who adapted the last version of the Goldenshluger-Lepski method to select the dimension of projection space for various estimation problems. For density estimation, the alternative is to choose the estimator $\hat{f}_{\hat{m}^b}$ in the collection $(\hat{f}_m)_{m\in\mathcal{M}_n}$ defined in (8) by

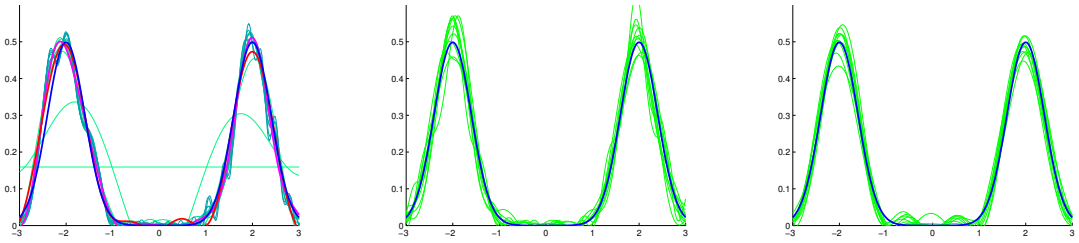
$$\hat{m}^b = \arg\min_{m\in\mathcal{M}_n} \text{Crit}_{GL,b}(m), \quad \text{with } \text{Crit}_{GL,b}(m) = \left\{A^b(m) + V^b(m)\right\},$$

with

$$V^b(m) = \kappa^b \Phi_0^2 \frac{D_m}{n}, \quad A^b(m) = \max_{m'\in\mathcal{M}_n} \left(\left\| \hat{f}_{m'} - \hat{f}_{m\wedge m'} \right\|^2 - V(m') \right)_+ .$$

The term $V^b(m)$ estimates the variance term of the risk of the projection estimator, and is the same as the penalty (22) in classical model selection. The second term A^b of $\text{Crit}_{GL,b}(m)$ is in the spirit of bandwidth selection with Lepski's method. Details can be found in the references above. We conclude with a practical comparison of the method. We plot in Figure 10, Part (a), an example of selected projection estimators $\hat{f}_{\hat{m}}$ (see (21)) and $\hat{f}_{\hat{m}^b}$. Beams of selected estimators, computed from independent samples are plotted on parts (b) and (c).

We also compare the risks of the three methods in Figure 11: projection estimation with model selection via penalisation or via the method in the spirit of Goldenshluger-Lepski, and kernel estimation with Goldenshluger and Lepski bandwidth selection. To that aim, boxplots of the Integrated Squared Errors, computed like in (16) are plotted. The result, obtained here to recover f_{Simu} defined in (1) is quite representative of what could be obtain for other simulation settings.



(a) (b) (c)

FIGURE 10. (a) Collection of projection estimators (in the Fourier basis) for f_{Simu}, computed with $n = 500$ observations, for different model dimension, and selected estimators $\hat{f}_{\hat{m}}$ (penalisation method) and $\hat{f}_{\hat{m}^b}$ (method in the spirit of Goldenshluger and Lepski). Bold blue line: true function f_{Simu}. Thin lines: estimators $\hat{f}_m$, $m = 1, 3, 5, \ldots, 51$. (b) and (c) Beams of estimators $\hat{f}_{\hat{m}}$ and $\hat{f}_{\hat{m}^b}$ respectively, computed from independent samples of size $n = 500$. Bold blue line: true function f_{Simu}. Thin green lines: selected estimators $\hat{f}_{\hat{m}}$ and $\hat{f}_{\hat{m}^b}$.

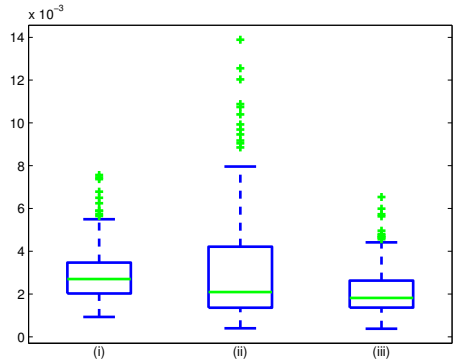


FIGURE 11. Boxplots of the Integrated Squared Error $\|\hat{f} - f_{\text{Simu}}\|^2$, approximated like in (16) for (i) $\hat{f} = \hat{f}_{\hat{h}}$ (kernel estimates with Goldenshluger and Lepski bandwidth selection) (ii) $\hat{f} = \hat{f}_{\hat{m}}$ (penalised projection estimators) (iii) $\hat{f} = \hat{f}_{\hat{m}^b}$ (projection estimators with Goldenshluger-Lepski type model selection). Same setting as before (Gaussian kernel, Fourier basis, $n = 500$).

5. Perspective for adaptive nonparametric estimation

The aim of the note was to introduce adaptive nonparametric estimation, from a theoretical point of view. For the sake of simplicity, we focus on the simple but important problem of univariate density estimation. This should not make one lose sight of the importance of such methods in very various applied problems. Let us quote in no particular order: regression estimation, inference from dependent data, from censored data, functional data analysis...

We have presented the two main classes of nonparametric estimators for the density of a real random variable, built from a sample of independent data. We have also explained how the

selection of the smoothness parameters (the model for projection estimates and the bandwidth for kernel estimators) can be performed in a data-driven way, to obtain final estimators that mimic the unknown oracle (that is the best function of the collection for the quadratic risk), and that reach the minimax optimal risk. We have not provided the proofs of the main results, Theorems 2 and 1, this was not the purpose of the note. Let us point out that they extensively involve a probabilistic tool, concentration of measure. Concentration inequalities play a crucial role to prove oracle bounds: it can be simple Bernstein inequality (Birgé and Massart, 1998, p.366), or less well known results, like the Talagrand inequality (see Klein and Rio 2005). Our aim was also to stress the links between projection and kernel density estimated: we have already remarked that the bandwidth of kernel estimators plays the role of the inverse of the model dimension for projection method, and the two selection rules we studied could be compared and modified by drawing inspiration from each other (see references at the end of Section 4.3). This is in the spirit that research is ongoing in the aera of adaptive nonparametric estimation, and we conclude this note by quoting three recent studies that show that estimator selection is still a dynamic topic!

- The first one goes further in the comparison between the methods: it is in fact possible to classify under the term of *linear estimator* both of the methods: each estimate can be written

$$\hat{f}_b(x) = \frac{1}{n}\sum_{i=1}^{n} m_b(X_i, x), \quad x \in I$$

  for a given function $m_b : I^2 \to \mathbb{R}$ ($b = D_m$ and $m_b(x,y) = \sum_{j=1}^{D_m} \varphi_j(x)\varphi_j(y)$ for projection methods, and $b = h^{-1}$ and $m_b(x,y) = K_h(x-y)$ for kernel estimation). The class also includes weighted estimators (like Pinsker's estimators, see Efroĭmovich 1985). Linear estimators were introduced under the name of *delta-sequences* by Walter and Blum (1979) and called *additive estimators* by Devroye and Lugosi (2001). The study we have in mind is the one of Lerasle et al. (2016), who address the problem of optimal selection among linear estimators: the question of optimal and minimal penalty (in the sense of the tuning of the constant κ, see the end of Section 4.2, and of the proof of sharp oracle inequalities) is solved in a very general way.
- Then, Lacour et al. (2016) deepen the link between model selection and Goldenshluger-Lepski methods by defining a new selection rule that seems very promising.
- Lepski (2016) proposes to use his methodology to solve new problems: the objective is to find hypotheses under which some elements of the solution of a statistical problem could be used to define minimax adaptive estimators for another more difficult problem. He tackles the question of smoothness parameter selection in the new problem by using the one selected in the (simpler) first problem. Conditions which ensure that the substitution is reasonable are established.

These recent references permit to conjecture that lots of theroretical studies, both on adaptive nonparametric statistics and on density estimation, will be developed in future years.

Acknowledgement

I am very thankful to Antoine Channarond and Fabienne Comte for stimulating discussions and careful readings of the manuscript.

References

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

D. M. Allen. The relationship between variable selection and data agumentation and a method for prediction. *Technometrics*, 16(1):125–127, 1974.

S. Arlot. V-fold cross-validation improved: V-fold penalization. preprint, 2008. URL `http://arxiv.org/abs/0802.0566`.

S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79, 2010.

S. Arlot and M. Lerasle. Choice of V for V-fold cross-validation in least-squares density estimation. *The Journal of Machine Learning Research*, 2016. To appear.

S. Arlot, M. Lerasle, and N. Magalhães. Selection of kernel estimators by cross- validation. preprint available in Chapter 3 of N. Magalhães PhD dissertation <tel-01164581>, 2015.

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

J.-P. Baudry, C. Maugis, and B. Michel. Slope heuristics: overview and implementation. *Stat. Comput.*, 22(2):455–470, 2012.

D. Belomestny, F. Comte, and V. Genon-Catalot. Nonparametric laguerre estimation in the multiplicative censoring model. preprint, 2016. URL `https://hal.archives-ouvertes.fr/hal-01252143v3`.

K. Bertin and N. Klutchnikoff. Minimax properties of beta kernel estimators. *J. Statist. Plann. Inference*, 141(7):2287–2297, 2011.

K. Bertin, C. Lacour, and V. Rivoirard. Adaptive pointwise estimation of conditional density function. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 52(2):939–980, 2016.

L. Birgé and P. Massart. Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields*, 97(1-2):113–150, 1993.

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

L. Birgé and P. Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998.

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

D. Bosq. *Nonparametric statistics for stochastic processes: estimation and prediction*, volume 110. Springer Science & Business Media, 2012.

G. Chagny. Penalization versus Goldenshluger-Lepski strategies in warped bases regression. *ESAIM: Probability and Statistics*, 17:328–358, 2013.

A. Cohen, R. DeVore, G. Kerkyacharian, and D. Picard. Maximal spaces with given rate of convergence for thresholding algorithms. *Appl. Comput. Harmon. Anal.*, 11(2):167–191, 2001.

F. Comte. *Estimation non-paramétrique*. Spartacus IDH, 2015.

F. Comte and J. Johannes. Adaptive functional linear regression. *Ann. Statist.*, 40(6):2765–2797, 2012.

W. J. Conover. *Practical nonparametric statistics*. Wiley New York, 1980.

R. A. DeVore and G. G. Lorentz. *Constructive approximation*, volume 303 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1993.

L. Devroye and G. Lugosi. *Combinatorial methods in density estimation*. Springer Series in Statistics. Springer-Verlag, New York, 2001.

D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

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

S. Y. Efroĭmovich. Nonparametric estimation of a density of unknown smoothness. *Teor. Veroyatnost. i Primenen.*, 30(3):524–534, 1985.

S. Efromovich. *Nonparametric curve estimation: methods, theory, and applications*. Springer Science & Business Media, 2008.

S. Geisser. The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70(350):320–328, 1975.

A. Goldenshluger and O. Lepski. Universal pointwise selection rule in multivariate function estimation. *Bernoulli*, 14(4):1150–1190, 2008.

A. Goldenshluger and O. Lepski. Structural adaptation via L_p-norm oracle inequalities. *Probab. Theory Related Fields*, 143(1-2):41–71, 2009.

A. Goldenshluger and O. Lepski. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *Ann. Statist.*, 39(3):1608–1632, 2011.

A. Goldenshluger and O. Lepski. General selection rule from a family of linear estimators. *Theory of Probability & Its Applications*, 57(2):209–226, 2013.

W. Härdle, G. Kerkyacharian, D. Picard, and A. Tsybakov. *Wavelets, approximation, and statistical applications*, volume 129 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1998.

R. Hasminskii and I. Ibragimov. On density estimation in the view of Kolmogorov's ideas in approximation theory. *Ann. Statist.*, 18(3):999–1010, 1990.

I. A. Ibragimov and R. Z. Has′minskiĭ. An estimate of the density of a distribution. *Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI)*, 98:61–85, 161–162, 166, 1980. Studies in mathematical statistics, IV.

A. Juditsky and S. Lambert-Lacroix. On minimax density estimation on $\mathbb{R}$. *Bernoulli*, 10(2): 187–220, 2004.

G. Kerkyacharian and D. Picard. Density estimation in Besov spaces. *Statist. Probab. Lett.*, 13 (1):15–24, 1992.

G. Kerkyacharian, O. Lepski, and D. Picard. Nonlinear estimation in anisotropic multi-index denoising. *Probab. Theory Related Fields*, 121(2):137–170, 2001.

T. Klein and E. Rio. Concentration around the mean for maxima of empirical processes. *Ann. Probab.*, 33(3):1060–1077, 2005.

C. Lacour and P. Massart. Minimal penalty for Goldenshluger-Lepski method. *Stochastic Process. Appl.*, 2016. to appear.

C. Lacour, P. Massart, and V. Rivoirard. Estimator selection: a new method with applications to kernel density estimation. arXiv preprint, 2016. URL `http://arxiv.org/abs/1607.05091`.

B. Laurent, C. Ludeña, and C. Prieur. Adaptive estimation of linear functionals by model selection. *Electronic journal of statistics*, 2:993–1020, 2008.

O. Lepski. Some new ideas in nonparametric estimation. arXiv preprint, 2016. URL `http://arxiv.org/abs/1603.03934`.

O. V. Lepskiĭ. Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Teor. Veroyatnost. i Primenen.*, 36(4):645–659, 1991.

O. V. Lepskiĭ. Asymptotically minimax adaptive estimation. II. Schemes without optimal adaptation. Adaptive estimates. *Teor. Veroyatnost. i Primenen.*, 37(3):468–481, 1992a.

O. V. Lepskiĭ. On problems of adaptive estimation in white Gaussian noise. In *Topics in nonparametric estimation*, volume 12 of *Adv. Soviet Math.*, pages 87–106. Amer. Math. Soc., Providence, RI, 1992b.

M. Lerasle, N. Magalhães, and P. Reynaud-Bouret. Optimal kernel selection for density estimation. In *High dimensional probability VII: The Cargèse Volume*, Progr. Probab. Birkhäuser/Springer, Basel, 2016. to appear.

C. L. Mallows. Comments on C_p. *Technometrics*, 15:661–675, 1973.

P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

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

S. M. Nikol′skiĭ. *Approximation of functions of several variables and imbedding theorems*. Springer-Verlag, New York, 1975. Translated from the Russian by John M. Danskin, Jr., Die Grundlehren der Mathematischen Wissenschaften, Band 205.

E. Parzen. On estimation of a probability density function and mode. *Ann. Math. Statist.*, 33: 1065–1076, 1962.

S. Plancade. Estimation of the density of regression errors by pointwise model selection. *Mathematical Methods of Statistics*, 18(4):341–374, 2009.

P. Reynaud-Bouret, V. Rivoirard, and C. Tuleau-Malot. Adaptive density estimation: a curse of support? *J. Statist. Plann. Inference*, 141(1):115–139, 2011.

P. Rigollet and A. B. Tsybakov. Linear and convex aggregation of density estimators. *Math. Methods Statist.*, 16(3):260–280, 2007.

M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.*, 27:832–837, 1956.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society. Series B (Methodological)*, pages 111–147, 1974.

A. B. Tsybakov. Agrégation d'estimateurs et optimisation stochastique. *J. Soc. Fr. Stat. & Rev. Stat. Appl.*, 149(1):3–26, 2008.

A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

G. Walter and J. Blum. Probability density estimation using delta sequences. *Ann. Statist.*, 7 (2):328–340, 1979.

L. Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006.