Contributions à l'estimation non-paramétrique adaptative : estimation de loi conditionnelle et déconvolution

Claire Lacour

Université Paris-Sud

Faculté des sciences d'Orsay
École doctorale de mathématiques Hadamard (ED 574)
Laboratoire de mathématique d'Orsay (UMR 8628 CNRS)

Mémoire présenté pour l'obtention du

Diplôme d'habilitation à diriger les recherches

Discipline : Mathématiques

par

Claire LACOUR

Contributions à l'estimation non-paramétrique adaptative : estimation de loi conditionnelle et déconvolution

Rapporteurs :
Alexander GOLDENSHLUGER
Pascal MASSART
Dominique PICARD

Date de soutenance : 8 Décembre 2015

Composition du jury :

Pascal MASSART	(Rapporteur)
Dominique PICARD	(Rapportrice)
Élisabeth GASSIAT	(Examinatrice)
Béatrice LAURENT	(Examinatrice)
Erwan LE PENNEC	(Examinateur)
Oleg LEPSKI	(Examinateur)
Fabienne COMTE	(Invitée)

Je souhaite remercier les membres du jury pour avoir accepté de prendre part à cette soutenance, en particulier Oleg Lepski et Béatrice Laurent qui ont fait le déplacement depuis Marseille et Toulouse. A ceci s'ajoute une gratitude particulière pour celles et ceux qui m'ont régulièrement aidée au cours de ces années, comme Elisabeth Gassiat avec qui j'ai une collaboration stimulante depuis plusieurs années, ou encore Erwan Le Pennec qui répondait déjà à mes questions à l'époque où j'étais en thèse – j'espère qu'il continuera encore longtemps. Je suis très heureuse de remercier à nouveau Fabienne Comte qui fut une directrice de thèse remarquable à tous points de vue.

Je suis très reconnaissante à Alexander Goldenshluger pour avoir bien voulu relire mon manuscrit. D'une façon générale, je remercie vivement mes rapporteurs/trice pour leur travail et pour les précieux conseils qu'ils et elle m'ont prodigués, à l'occasion de ce mémoire, mais aussi depuis de nombreuses années. Pascal Massart m'a été d'un grand soutien, et les nombreuses discussions que nous avons eues ont toujours été très agréables et enrichissantes ; quant à Dominique Picard c'est dès l'encadrement de mon mémoire de maîtrise qu'elle a su m'offrir son aide.

Au fil des ans j'ai eu l'occasion de nouer de fructueuses collaborations, dont certaines m'ont laissée d'excellents souvenirs tant personnels que professionnels. Ainsi, Yohann De Castro, qui en plus de savoir parfaitement approximer des matrices, est un parfait compagnon de bureau, ou encore Thanh Mai Pham Ngoc dont le dynamisme à tous les niveaux est vraiment appréciable. Plus éloignée géographiquement mais certainement pas dans mes pensées, je remercie également Gaëlle Chagny et Sandra Plancade, mes petites soeurs de thèse. Enfin, je ne saurais trop exprimer de gratitude envers Vincent Rivoirard, qui m'a toujours témoigné beaucoup de confiance et dont j'ai eu le plaisir de découvrir avec les années combien il était quelqu'un avec qui il est agréable de travailler et d'échanger.

Il serait fastidieux de dresser la liste de tous les collègues d'Orsay à qui je suis redevable : l'ensemble de l'équipe probabilités et statistiques du laboratoire pourrait y trouver sa place. Je tiens toutefois à mentionner particulièrement Marie-Anne Poursat, avec qui j'ai eu des moments vraiment agréables tout au long de ces années.

Merci à vous deux qui vous reconnaitrez.

Contents

Chapter 1

Preamble

This dissertation presents the work I have done between 2007 and 2014, that is to say at the end of my PhD at University Paris Descartes and mostly at University Paris-Sud as a "maître de conférences". The detail is in the corresponding papers. In addition to the content of the articles, here are some more general reflections.

1.1 Risk and approximation

1.1.1 Loss function

All of this work is placed in the context of nonparametric statistics. Essentially, we consider the following issue: we want to estimate a function f from observations $Z_1, \dots, Z_n$ identically distributed with law $\mathbb{P}_Z$ depending on f. The aim will be to provide a random function $\hat{f}$ depending only on the observations and approaching as close as possible to f. To calculate the performance of the estimator, the question of the distance between functions arises. The distances $\mathbb{L}^p$ are the most natural, with $p = 1, 2$ or ∞: then one is interested in $\|f - \hat{f}\|_p$. The distance $\mathbb{L}^1$ may seem more appropriate for densities, distance $\mathbb{L}^\infty$ being more sensitive to "bumps" and distance $\mathbb{L}^2$ a compromise between the two. In my work, I have always used the distance $\mathbb{L}^2$, this choice often being guided by technical considerations more than anything else (especially when working with Fourier transforms). Nevertheless, it remains a quite reasonable choice and probably the most common in nonparametric statistics. I have also sometimes considered the pointwise distance $|f(x_0) - \hat{f}(x_0)|$, which is closely linked to the $\mathbb{L}^\infty$ norm. When f is a density or a distribution function, and we then look for estimating a probability distribution, the question arises of working with distances between the intrinsic probability measures (i.e. not depending on the dominating measure) as the Kullback divergence or the Hellinger distance. Here, in all considered applications, the dominating measure will always be the Lebesgue measure naturally. So we only consider distance between functions rather than distributions.

1.1.2 Approximation spaces

Here we present a lot of oracle inequalities, that is to say comparisons between estimators. But secondly, to assess the optimality of our results, we consider the minimax framework. The estimate always requires an approximation: projection of f on a finite dimensional vector space or smoothing kernel. To control this approximation, I have always assumed that my target function belongs either to a Besov space or to a Sobolev space. These spaces are now well known of statisticians in the univariate framework. The interest of Besov spaces $B^\alpha_{p,q}$ is that they can

describe both very regular functions, but also those with peaks and spatial inhomogeneity. We retrieve Sobolev spaces by taking $p = q = 2$ and Hölder spaces if $p = q = \infty$. Thus, the space $B^{\alpha}_{2,\infty}$ contains the Sobolev space W^{α}_2 (actually all Besov spaces $B^{\alpha}_{2,q}$ for $q \geqslant 1$). It also contains $B^{\alpha}_{p,\infty}$ for any $p \geqslant 2$ if the domain is a bounded open set of $\mathbb{R}^n$ (see Amann, 2000). These inclusions (or embeddings) are the reason why we we often consider this space $B^{\alpha}_{2,\infty}$. In the multivariate framework, we must be rigorous enough to define anisotropic spaces (see Section 3.1.2 for the generalization of Sobolev spaces). Besov spaces are well defined in this context and previous embeddings remain true even in the anisotropic case (see Triebel, 2006).

However issues of smoothness spaces and estimating distance cannot be separated. Indeed, let us consider a function $f \in B^{\alpha}_{p,\infty}$ to be estimated with a $\mathbb{L}^q$ loss ($1 \leqslant q \leqslant \infty$), then we distinguish two cases:

- If $p \geqslant q$, f has a classical smoothness, typically with wavelet coefficients not too large, but many are non-negligible. This case is often called homogeneous. Then f is linearly approximable with $\mathbb{L}^q$ norm, the minimax rate of convergence with $\mathbb{L}^q$ loss is $n^{-\alpha/(2\alpha+1)}$ and it is achieved by linear estimators.

- If $p < q$, it is the inhomogeneous case. We can not approximate f with regular bases. If α is large enough ($\alpha > (q/p-1)/2$), the minimax rate is still $n^{-\alpha/(2\alpha+1)}$, otherwise it becomes like $(\log n/n)^{\beta(\alpha,p,q)}$.

In the second part of this manuscript, we simply study the homogeneous case ($p = q = 2$), while the inhomogeneous case is detailed in the first part.

1.2 Adaptive methods

This manuscript is divided into two parts, according to the two areas of research that I have explored during my short life as a researcher. The first part deals with the conditional density estimation, and the second one with noisy models. Common to these two parts is the adaptive nonparametric estimation. I have used two main methods for adaptive estimation: Birgé-Massart model selection, and Goldenshluger-Lepski method. Both are non-asymptotic (at least in theory).

1.2.1 Birgé-Massart method

Let us recall briefly the principle of model selection, the reference on the subject is of course Massart (2007). Given a family of estimators $(\hat{f}_m)_{m\in\mathcal{M}}$, the issue is the choice of m. It is assumed that the target function f can be written as a minimizer of a contrast function $f = \arg\min_t \mathbb{E}(\gamma(t, Z)) = \arg\min_t P\gamma(t)$, which naturally provides estimators by minimizing the empirical risk:

$$\hat{f}_m = \underset{t\in S_m}{\operatorname{argmin}}\ \gamma_n(t) = \underset{t\in S_m}{\operatorname{argmin}}\ \frac{1}{n}\sum_{i=1}^{n}\gamma(t, Z_i)$$

where γ_n is the empirical contrast and spaces S_m are called models. In this framework, by considering the loss associated with the contrast (typically the $\mathbb{L}^2$ loss for least squares, or Kullback divergence for a maximum likelihood estimator) the optimal m (or oracle) is the one which minimizes $P\gamma(\hat{f}_m)$. Then one might be tempted to minimize the empirical equivalent of this quantity, but $\gamma_n(\hat{f}_m)$ is a biased estimator, that needs to be corrected. The idea is to introduce a penalty function pen and to set

$$\hat{m} = \underset{m\in\mathcal{M}}{\operatorname{argmin}}\ \left\{\gamma_n(\hat{f}_m) + \operatorname{pen}(m)\right\}.$$

Here the penalty may depend on the "dimension" of each model and the "complexity" of the collection. This method allows to prove non-asymptotic oracle inequalities. It has now been extensively studied, both practically and theoretically, in many models, e.g. in change detection (Lebarbier, 2005), spatial statistics (Verzelen, 2010), classification (Maugis and Michel, 2011), or geometrical inference (Caillerie and Michel, 2011). Due to its practical and theoretical performance, we have extensively used it here. However, it cannot really apply outside the scope of the estimation by contrast minimization, controlled by the associated loss function. Thus, it is not suitable to control the pointwise risk, at a given point. Laurent et al. (2008) proposed an adaptation of the method of Birgé-Massart for linear functional. This is actually very similar to the method of Goldenshluger-Lepski described below, at least in the univariate case. To estimate $T(f)$, a linear functional of f, it consists in replacing the term $\gamma_n(\hat{f}_m)$, which estimated (up to a constant) the bias in the model selection, with the term $\sup_{m' \geqslant m}(T(\hat{f}_m) - T(\hat{f}_{m'}) - H(m, m'))$ where H is a compensation term to be specified. Here we find the idea of comparing estimators two by two, already present in Lepski's method. A problem arises when addressing multivariate functions. To manage the anisotropy of f, we need to use indices $m = (m_1, m_2)$ in two dimensions. The notion of order between m and m' is not clear to define, and it is difficult to simultaneously put in order the bias and the variance. By the way, as noticed by Kerkyacharian et al. (2001), this lack of natural ordering is one of the reasons that make adaptive methods difficult to implement in an anisotropic framework. Although the model selection can work in this framework (in regular cases, inequality $\dim(S_m + S_{m'}) \leqslant \dim(S_m) + \dim(S_{m'})$ allows to easily manage the complexity of anisotropic models), it is nevertheless limited to some estimators and associated risks. The handling of the anisotropy is one of the motivations for the introduction of the Goldenshluger-Lepski method.

1.2.2 Goldenshluger-Lepski method

While the model selection is provided to select among contrast minimization estimators, the Goldenshluger-Lepski method is designed for kernel estimators. It is based on pairwise comparison of estimators, which is perhaps its main drawback (for its practical application). These authors first develop their methodology in white noise model (Goldenshluger and Lepski, 2008, 2009), next for density estimation (Goldenshluger and Lepski, 2011) and then for various models (Goldenshluger and Lepski, 2013). Their initial objective was to provide an adaptive procedure for multivariate and anisotropic estimation. They use it to give minimax rates of convergence in a very general framework (see Goldenshluger and Lepski, 2014). For this purpose, they have established oracle inequalities to ensure that the final estimator is almost as efficient as the best one in the collection. This methodology has next been applied in concrete examples: see Doumic et al. (2012) for transport-fragmentation equations, or [L16] for relative density in two-sample problems (this paper will not be mentioned in this dissertation).

This method proposes a data-driven choice of h to select an estimator among a collection $(\hat{f}_h)_{h \in \mathcal{H}}$. To sum up, the selected $\hat{h}$ is chosen as a minimizer of $A(h) + V(h)$ with

$$A(h) = \sup\{[\|\hat{f}_{h'} - \hat{f}_{h,h'}\|^2 - V(h')]_+,\ h' \in \mathcal{H}\}$$

where x_+ denotes the positive part $\max(x, 0)$ and where $\hat{f}_{h,h'}$ are oversmoothed auxiliary estimators and $V(h)$ is a penalty term to be suitably chosen. Heuristically, the term $A(h)$ has the same order as $\sup\{[\|\mathbb{E}(\hat{f}_{h,h'}) - \mathbb{E}(\hat{f}_{h'})\|^2,\ h' \in \mathcal{H}\}$ because the distance to the expectation is canceled by $V(h')$. And, if h' tends to 0, $\|\mathbb{E}(\hat{f}_{h,h'}) - \mathbb{E}(\hat{f}_{h'})\|$ tends to the bias $\|\mathbb{E}(\hat{f}_h) - f\|$. Then the final choice $\hat{h} = \arg\min_h A(h) + V(h)$ mimics a bias-variance trade-off. Other heuristics are

presented in [L17]. Although this method is initially intended for selecting kernel estimators, we have also extended it to projection estimators, see Section 2.4.3.

1.2.3 Concentration inequalities

Concentration inequalities are deeply involved in almost all the proofs of this dissertation. Sometimes, a simple Bernstein inequality is enough, but we have generally used Talagrand inequality. Thus the main probabilistic tool for the sequel is the following result:

Lemma 1. *[Talagrand's Inequality, adapted from Klein and Rio (2005)] Let $X_1, \dots, X_n$ be a sequence of i.i.d. variables and $\nu(t) = n^{-1}\sum_{i=1}^{n}[g_t(X_i) - \mathbb{E}(g_t(X_i))]$ for t belonging to a countable set of functions $\mathcal{F}$. Assume that for all $t \in \mathcal{F}$ $\|g_t\|_\infty \leqslant b$ and $\mathrm{Var}(g_t(X_1)) \leqslant v$. Denote $H = \mathbb{E}(\sup_{t\in\mathcal{F}} \nu(t))$. Then, for any $\varepsilon > 0$, for $H' \geqslant H$,*

$$\mathbb{P}(\sup_{t\in\mathcal{F}} \nu(t) \geqslant (1+\varepsilon)H') \leqslant \max\left(\exp\left(-\frac{\varepsilon^2}{6}\frac{nH'^2}{v}\right), \exp\left(-\frac{\min(\varepsilon,1)\varepsilon}{24}\frac{nH'}{b}\right)\right),$$
$$\mathbb{P}(\sup_{t\in\mathcal{F}} \nu(t) \leqslant H - \varepsilon H') \leqslant \max\left(\exp\left(-\frac{\varepsilon^2}{6}\frac{nH'^2}{v}\right), \exp\left(-\frac{\min(\varepsilon,1)\varepsilon}{24}\frac{nH'}{b}\right)\right).$$

For dealing with dependent variables we also use the works of Adamczak (2008) and Paulin (2014).

1.3 Abstract

This dissertation is divided into two parts. The first part deals with the estimation of a conditional density. We first present the motivation and the bibliographic context of our study. Then our main estimator is introduced: it is a minimizer of an original empirical contrast $\gamma_n(t) = n^{-1}\sum_{i=1}^{n}[\int t^2(X_i, y)dy - 2t(X_i, Y_i)]$. The minimization is done on piecewise polynomials approximation spaces, and we detail how the minimization is possible in Section 2.2. Next we present an adaptive Birgé-Massart model selection procedure which leads to oracle inequalities for $\mathbb{L}^2$-risk. The rates of convergence are studied for the problem of estimating the conditional density, first for homogeneous regularity, next for inhomogeneous one. The latter requires the use of a specific collection of models based on dyadic partitions. Then, we study local adaptation for the estimation of the conditional density at a given point. In this case the selection is done via Goldenshluger-Lepski method, and we also study a kernel estimator. The next section is devoted to some generalizations of the previous results: we first extend to dependent data and censored data, and then to the estimation of a conditional cumulative distribution. In this case, interesting rates of convergence appear which combine parametric and nonparametric rates. To complete this part, we give some considerations on the penalty calibration issue. In particular we provide a minimal penalty for the Goldenshluger-Lepski method in the case of density estimation. We conclude with some prospects.

In the second part, we present our works in the framework of deconvolution and hidden Markov chains. Deconvolution models deal with the density f of a signal X which is contaminated by a noise ε, so that only the variable $X+\varepsilon$ is observed. First, we study the case of a multivariate signal. In this case we give lower and upper bounds for the rates of convergence: these are very complex, depending on the smoothness of both the signal and the noise. We also introduce an adaptive estimator which enjoys good theoretical (oracle inequality) and practical properties. Then we come back to the univariate background to investigate the case of an unknown noise distribution. We assume that an additional sample of the pure noise is available: $\varepsilon_{-1}, \dots, \varepsilon_{-M}$,

and we give upper bounds on the rates of convergence, depending both on M and n the size of the initial sample of Y. Adaptation is performed via Birgé-Massart model selection with a specific random penalty. This section ends with a work on pointwise estimation for Lévy process, this topic being very close to the deconvolution topic. Section 3.2 details a goodness-of-fit test procedure for a deconvolution problem on the sphere. Motivated by astrophysical applications, we are interested in testing uniformity of noisy spherical data. Again, the separation rates depend on the noise smoothness, and we provide both lower and upper bounds for these rates. We implement our procedure and try it on both simulated and real data. Finally, we investigate the case of a Markovian signal X with finite state space when only indirect observations Y are available. Using the model selection technique, we give nonparametric estimators of the density of Y given $X = k$ for this hard problem (nothing of X is assumed to be known but the cardinal of the hidden state space). They are proved to be adaptive optimal and computationally efficient, using a preliminary spectral estimator to initialize our algorithm.

Chapter 2

Estimation of a conditional distribution

2.1 Introduction

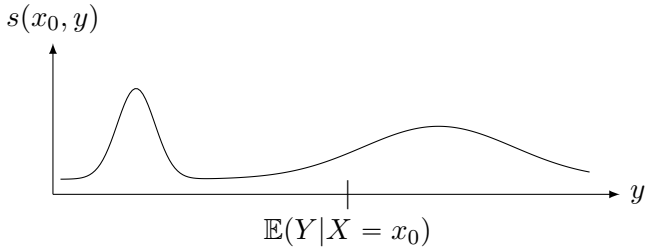
The first main theme of my research is the estimation of a conditional density. This problem can be stated in this way. Let (X, Y) be a pair of random variables, and we assume that the conditional distribution of Y given X admits a density, denoted by s in this document:

$$s(x, y)dy = \mathbb{P}(Y \in dy | X = x).$$

The goal is to find the conditional density from a sample with the same distribution as (X, Y): $(X_1, Y_1), \ldots, (X_n, Y_n)$.

2.1.1 Motivation

This question may arise as soon as we observe a (possibly multidimensional) response Y associated with a (possibly multidimensional) covariate X. Of course this issue is extremely large. We often study the regression function $\mathbb{E}(Y|X = x)$. But this information is restrictive, and the entire distribution is more informative than the mean. The most typical case is the case of a bimodal distribution:



This figure shows that the knowledge of $\mathbb{E}(Y|X = x_0)$ does not give enough information about the law of Y given $X = x_0$. Therefore the problem of estimating the conditional distribution is richer than the one of estimating a regression function, and is found in various application fields. Let us give here some examples:

- in meteorology: prediction of the electrical power produced by a wind turbine as a function of wind speed (Jeon and Taylor, 2012);
- insurance: conditional density of claim severity given claim count (Resti et al., 2012), or premium given credit score (Efromovich, 2010b);

- in medical studies: spinal bone mineral density given the age (Takeuchi et al., 2009), survival in days of cancer patients given months from diagnosis and age (Hall et al., 2004);
- in geology: waiting time between the starts of successive eruptions et the duration of the subsequent eruption for a geyser (Azzalini and Bowman, 1990)
- in astronomy: distribution of redshifts given spectroscopic measurements (Holmes et al., 2012).

Moreover, even within the statistical field, we must mention the ABC methods (Approximate Bayesian Computation): estimating the conditional distribution of θ given observations is really at the heart of the method, see Section 2.4. Moreover, the problem of estimating the conditional probability density is related to the estimation of the transition of a Markov chain with continuous state space: just consider $(X, Y) = (X_i, X_{i+1})$. More generally, making inference for a stochastic process $(X_t)_{t\geqslant 0}$ from discrete observations requires studying the law of $X_{t+\Delta}$ given $X_t = x$ (for example to maximize the likelihood $\prod s_\Delta(X_{i\Delta}|X_{(i-1)\Delta})$, see Aït-Sahalia (2001)).

All these considerations make the estimation of conditional density a subject of study which should not be left behind. Note that the problem of conditional density estimate is at the intersection of density estimation and regression. Since the conditional density $s(x, .)$ is the density of Y given $X = x$, we face a problem of density estimation in the y-direction, and the estimation of the regression function $\mathbb{E}(Y|X = x)$ in the x-direction. From a theoretical point of view, the conditional density estimation is thus a mixing of the two main models considered in nonparametric estimation.

2.1.2 State of the art

The most natural way to estimate a conditional density is to express it as a ratio of the density of the couple (X, Y) and the density of X. It is this method that was used by Rosenblatt (1969) and in the works of the 90s. Using a kernel K and a couple of bandwidths (h_1, h_2), we can estimate s by

$$\hat{s}(x, y) = \frac{\sum_{i=1}^n K_{h_1}(X_i - x)K_{h_2}(Y_i - y)}{\sum_{i=1}^n K_{h_1}(X_i - x)} \tag{1}$$

where $K_h(x) = K(x/h)/h$. We can obviously choose two different kernels rather than one. This estimator can also be seen as the Nadaraya-Watson estimator applied to data X_i and $Z_i(y) = K_{h_2}(Y_i - y)$. This can be understood by noticing that if $K = \mathbb{1}_{[-1,1]}/2$, the regression function verifies

$$\mathbb{E}(Z_i(y)|X_i = x) = \frac{1}{2h_2}\mathbb{E}(\mathbb{1}_{|Y_i-y|\leqslant h_2}|X_i = x) = \frac{F(y + h_2|x) - F(y - h_2|x)}{2h_2} \approx s(x, y).$$

This estimator is also proposed by Roussas (1969) in the framework of Markov chains and used in the context of stationary mixing processes (X_t), where one studies the conditional density of $(X_{i_{p+1}}, \ldots, X_{i_m})$ given $(X_{i_1}, \ldots, X_{i_p})$ (Masry, 1989; Cai, 1991). Starting from this standard estimator, several improvements have been suggested. Noting that there is a bias in the estimation of the conditional mean, that is to say the regression function $r = \int ys(., y)dy$, Hyndman et al. (1996) use a preliminary estimator $\hat{r}$. Then, in (1), Y_i is replaced by $Y_i^*(x) = \hat{r}(x) + (Y_i - \hat{r}(X_i))$. This is also what is used in Beaumont et al. (2002) or Blum (2010) with different estimators of r. Fan et al. (1996) suggest an approach with local polynomials that generalizes the classical method. If r is the degree of the polynomial, $s(x, y)$ is estimated by $\hat{\theta}_0$ where $\hat{\theta}$ is the vector

which minimizes

$$\sum_{i=1}^{n}\left(K_{h_2}(Y_i-y)-\sum_{k=0}^{r}\theta_k(X_i-x)^k\right)^2 K_{h_1}(X_i-x).$$

We can show that this estimator can be written $\sum_{i=1}^{n} w_i(x)K_{h_2}(Y_i-y)$. This estimator is further modified by Hyndman and Yao (2002) in order to get a positive function. From the same point of view, De Gooijer and Zerom (2003) introduce an estimator of the form $\sum_{i=1}^{n} w_i(x)K_{h_2}(Y_i-y)$ but with weights $w_i(x)$ different from the classical estimator or from the one of Fan et al. (1996) to keep the benefits of these (positivity and good bias respectively).

For all these estimators we may wonder about the choice of bandwidths. Different methods have been advocated: calculation of the optimal bandwidth assuming Gaussian data (Chen et al., 2001) or bootstrap approach (Bashtannyk and Hyndman, 2001). Fan and Yim (2004) make a numerical study that shows the superiority of the cross-validation. Hall et al. (2004), in addition to studying the case where X contains discrete components, are also interested in bandwidth selection by cross-validation and give a theoretical result (see also Efromovich (2010a) for the case where X includes discrete and continuous components). Holmes et al. (2012) propose a fast numerical method for cross-validation that minimizes the log-likelihood rather than the integrated squared risk. This kernel/cross-validation approach is also used by Bouaziz and Lopez (2010) who consider a semiparametric single-index model in the case of censored data.

All previous papers study kernel estimators, but there are some other methods. Stone (1994) introduce an estimator by maximizing the likelihood on a space of splines. Györfi and Kohler (2007) study a histogram type estimator. The approach of Faugeras (2009) is a kernel one but is original because the idea is to express the conditional density not as a quotient but as a product, using copula. In the Markov framework also, Clémençon (2000) introduces two wavelets thresholding estimators. The first as a quotient of an estimator of $f_{X,Y}$ and an estimator of f_X (see also [L2] for model selection approach), the second by the method of the boxes of Hoffmann, using an analogy with the framework regression. His work is the first where adaptation is really treated theoretically.

2.1.3 Notation and assumptions

Throughout this chapter, we assume that the conditional distribution of Y given X admits a density with respect to the Lebesgue measure. We denote s this conditional density. It is also assumed that the distribution of X admits a density f with respect to the Lebesgue measure. So the couple (X,Y) admits as density $f(x)s(x,y)$.

Here we consider the dimension 2, i.e. X and Y are one-dimensional. The extension to the multidimensional case ($X \in \mathbb{R}^{d_1}$ and $Y \in \mathbb{R}^{d_2}$) does not pose any particular problem, at least from a methodological point of view. The rates of convergence are amended in the usual way. Application problems posed by the "curse of dimensionality" will be discussed at the end of this chapter.

We will use four different norms, defined in this way, for $t : \mathbb{R}^2 \to \mathbb{R}$:

$$\begin{aligned}
\|t\|_2 &= \left(\iint t^2(x,y) dx dy \right)^{1/2}, \\
\|t\|_f &= \left(\iint t^2(x,y) f(x) dx dy \right)^{1/2}, \\
\|t\|_n &= \left(\frac{1}{n} \sum_{i=1}^{n} \int t^2(X_i, y) dy \right)^{1/2}, \\
\|t\|_{x,2} &= \left(\int t^2(x,y) dy \right)^{1/2}.
\end{aligned}$$

These norms will sometimes be used for univariate functions: for example, if the function t is univariate $t : \mathbb{R} \to \mathbb{R}$, we have $\|t\|_2^2 = \int t^2(x)dx$ and $\|t\|_f^2 = \int t^2(x)f(x)dx$. We also use the dot product $\langle ., . \rangle_f$ and the distance d_f associated to the norm $\|.\|_f$:

$$\langle t_1, t_2 \rangle_f = \iint t_1(x,y) t_2(x,y) f(x) dx dy, \quad d_f(t, S) = \min_{u \in S} \|t - u\|_f. \tag{2}$$

Furthermore it is assumed that X and Y are living in a compact set, assumed without loss of generality equal to $[0,1]$, that is to say that s has support $[0,1]^2$ and f has support $[0,1]$. In fact, one might assume that the distributions are not compactly supported, but our methods provide an estimation of s only on a compact set A, so that we would actually estimate s restricted to A. For the calculation of rates of convergence, assuming that s belongs to a Besov space over $\mathbb{R}^2$ implies that s restricted to A belongs to the Besov space (with the same smoothness) on A. So for the sake of simplicity we chose to always assume s with compact support $[0,1]^2$. Estimation on a non-compact set is another issue, and the rate can be degraded in some cases: see for instance Reynaud-Bouret et al. (2011).

The main assumption required in this document is the following:

Assumption (A) For all (x, y) in $[0,1]^2$,

$$s(x,y) \leqslant \|s\|_\infty < \infty, \qquad 0 < f_0 := \inf_{[0,1]} f \leqslant f(x) \leqslant \|f\|_\infty < \infty$$

It is not much restrictive to assume s and f bounded. The really strong assumption is the lower bound of f by f_0. However, the need for this assumption can be understood without difficulty. Estimating the distribution of X or the distribution of Y given X in the neighborhood of points x where $f(x)$ is equal to or close to 0 is of course very difficult, since there will be no observation X_i in this area. Thus this assumption is classical in a regression (or conditional density) framework. It is also required in most of the aforementioned works. It can be avoided if one uses only the norm $\|.\|_f$ but then coming back to the norm $\|.\|_2$ requires the assumption of lower bound.

We denote x_+ the positive part of x: $x_+ = \max(x, 0)$. For two sequences u, v, we denote $u_n \lesssim v_n$ if there exists a positive constant C not depending on n such that $u_n \leqslant C v_n$. For two functions φ, ψ, we denote

$$\varphi \otimes \psi : (x,y) \mapsto \varphi \otimes \psi(x,y) = \varphi(x)\psi(y)$$

Last, we define the harmonic mean $\bar{\alpha}$ of a couple of positive reals $\alpha = (\alpha_1, \alpha_2)$ by

$$\frac{1}{\bar{\alpha}} = \frac{1}{2} \left(\frac{1}{\alpha_1} + \frac{1}{\alpha_2} \right).$$

2.2 Least squares contrast and estimator

2.2.1 Definition

The motivation of my first works about conditional density was the following. Is it possible to find an alternative estimation method that avoids the disadvantages of the traditional method by kernel ratio? In particular, if we look at the rate of convergence of an estimator of the form $\hat{f}_{X,Y}/\hat{f}$, it will depend on the smoothness of f. But the smoothness of f can be much smaller than the smoothness of s and thus slow down considerably the estimate. Moreover, the goal was to develop an adaptive procedure, theoretically justified by oracle inequalities, taking into account the anisotropy of the function. Compared to Clémençon (2000), we also got a more implementable estimator with minimax rate of convergence without logarithmic loss.

We chose the method of contrast minimization. We want to build a random operator γ_n depending on the data, such that for any bivariate function t, $\mathbb{E}[\gamma_n(t)]$ is minimum for $t = s$. So it will be logical to take as an estimator a function t minimizing $\gamma_n(t)$. This method is used in density estimation with the contrast:

$$\frac{1}{n}\sum_{i=1}^{n}[\int t^2(y)dy - 2t(Y_i)],$$

whose expectation is minimum when t is equal to the density of Y_i. In regression, we have the least squares contrast:

$$\frac{1}{n}\sum_{i=1}^{n}[t^2(X_i) - 2t(X_i)Y_i].$$

Taking inspiration from these two well-known contrasts, we can define

Definition 2.

$$\gamma_n(t) = \frac{1}{n}\sum_{i=1}^{n}[\int t^2(X_i, y)dy - 2t(X_i, Y_i)].$$

The link with the usual least squares contrast for regression functions $\int s(x, y')\psi(y')dy' = \mathbb{E}(\psi(Y_1)|X_1 = x)$ is detailed in [L3], [L6]. This contrast verifies

$$\mathbb{E}\gamma_n(t) \quad = \quad \iint t^2(x, y)f(x)dxdy - 2\iint t(x, y)s(x, y)f(x)dxdy = \|t - s\|_f^2 - \|s\|_f^2$$

which is minimum when $t = s$.

2.2.2 Minimization

So we want to consider the following estimator

$$\hat{s} = \underset{t\in S}{\operatorname{argmin}}\ \gamma_n(t)$$

with S a set of functions to be specified, that we always consider of the form $S = \text{Vect}\{\varphi_j \otimes \psi_k, (j, k) \in \mathcal{L}\}$. But what is the meaning of this minimization? Assume for the sake of simplicity that $\mathcal{L}$ is a Cartesian product $\mathcal{L} = J \times K$. We can prove the following lemma.

Lemma 3. *If the function* $\hat{s}(x, y) = \sum_{j\in J}\sum_{k\in K}\hat{a}_{j,k}\varphi_j(x)\psi_k(y)$ *minimizes the empirical contrast function* γ_n *on* $S = \text{Vect}\{\varphi_j \otimes \psi_k, j \in J, , k \in K\}$, *then*

$$G\hat{A} = Z,$$

where $\hat{A}$ is the coefficients matrix $(\hat{a}_{j,k})_{j\in J,k\in K}$,

$$G = \left(\frac{1}{n}\sum_{i=1}^{n}\varphi_{j_1}(X_i)\varphi_{j_2}(X_i)\right)_{j_1,j_2\in J} \quad \textit{and} \quad Z = \left(\frac{1}{n}\sum_{i=1}^{n}\varphi_j(X_i)\psi_k(Y_i)\right)_{j\in J,k\in K}.$$

Moreover the minimum of the contrast is then $\gamma_n(\hat{s}) = \mathrm{tr}(-{}^tZ\hat{A})$.

We see on this formula that $\hat{s}$ might be not well defined at any point, if G is not invertible. This non-invertibility occurs when using localized bases, as the base of piecewise polynomials: if there exists j_0 in J such that there is no observation in the support of φ_{j_0}, then G has a null column. Actually, it is more a problem of uniqueness than existence: this corresponds to the case where the linear system has infinitely many solutions. We can show the following result.

Lemma 4. *Let $W = \{(t(X_i,y))_{1\leqslant i\leqslant n},\, t\in S\}$ and P_W be the orthogonal projection on W. If $\hat{s}$ minimizes the empirical contrast function on S, then $(\hat{s}(X_i,y))_{1\leqslant i\leqslant n}$ is uniquely defined as the projection*

$$(\hat{s}(X_i,y))_{1\leqslant i\leqslant n} = P_W\left(\left(\sum_k \psi_k(Y_i)\psi_k(y)\right)_{1\leqslant i\leqslant n}\right).$$

Then it is sufficient to interpolate to obtain values of $\hat{s}$ at any x, y. But we can say that this function is more a theoretical tool and the estimator constructed by our method is actually the vector $(\hat{s}(X_i,y))_{1\leqslant i\leqslant n}$. This explains that in all cases where we are interested in the global risk, we use the $\|.\|_n$ norm defined by

$$\|t\|_n \quad = \quad \left(\frac{1}{n}\sum_{i=1}^{n}\int t^2(X_i,y)dy\right)^{1/2}.$$

This empirical norm is the natural distance for our problem, and moreover we can notice that if t is a deterministic function, under assumption (**A**),

$$f_0\|t\|_2^2 \leqslant \mathbb{E}\|t\|_n^2 = \|t\|_f^2 \leqslant \|f\|_\infty\|t\|_2^2$$

and then the mean of the empirical norm is equivalent to the usual $\mathbb{L}^2$ norm. To prove the results that follow, we will always consider a space where the norm $\|.\|_n$ and its mean the norm $\|.\|_f$ are close. This space having a probability close to 1, it will be sufficient to do the study on this space and we will have the results with high probability.

In Section 2.4, we also study this estimator in the neighborhood of a given point x_0. In this case, the precise definition of $s(.,y)$, even outside of the vector of observations $X_1,\ldots,X_n$, will be required. Then we will set $\hat{s}(x,y) = \sum_j\sum_k \hat{a}_{jk}\varphi_j(x)\psi_k(y)$ with

$$(\hat{a}_{jk})_{j\in J,k\in K} = \begin{cases} G^{-1}Z & \text{if } \min(\mathrm{Spectrum}(G)) > \text{thresholding to be specified,} \\ 0 & \text{otherwise.} \end{cases}$$

2.2.3 Models

In the sequel of this document (Section 2), it is considered as space S a space of piecewise polynomials of degree smaller than a non negative integer r. To each partition m of the set $[0,1]^2$ into rectangles, one can associate S_m the space of all piecewise polynomial functions on $[0,1]\times[0,1]$ which are polynomial by coordinate with degree $\leqslant r$ on each rectangle $R = I_1\times I_2$

of m. Each approximation space, called model, is determined by a partition m. We denote $|m|$ the cardinality of the partition m and

$$D_m = \dim(S_m) = (r+1)^2|m|$$

the dimension of S_m. We denote by $\mathcal{M} = \mathcal{M}_n$ the set of all considered partitions, which will be detailed later. In practice we consider a basis of Legendre polynomials in each direction, with an affine transformation to come back to the given interval. Using again the previous notation, $S_m = \text{Vect}\{\varphi_j^m \otimes \varphi_k^m, (j,k) \in \mathcal{L}_m\}$ with

$$\mathcal{L}_m = \{((I_1,d_1),(I_2,d_2)), 0 \leqslant d_1, d_2 \leqslant r, R = I_1 \times I_2 \in m\}.$$

The φ_j^m are Legendre polynomials, the index j indexing both the degree d (between 0 and r) and the interval I:

$$\varphi_j^m(u) = \varphi_{I,d}^m(u) = \sqrt{\frac{2d+1}{|I|}} P_d(T(u)) \mathbb{1}_I(u)$$

where P_d is the Legendre polynomial of degree d on $[-1,1]$ and T is the affine mapping that maps I into $[-1,1]$ and $|I|$ the length of the interval I.

In Section 2.3.4 (estimation for a conditional density with inhomogeneous smoothness), we will use irregular partitions of the square. In this case $\mathcal{L}_m$ can not be written as a Cartesian product, as assumed in Lemma 3. However, for each rectangle R of the partition m, we will have a similar matrix equation and the uniqueness of the definition of the estimator will be true also by the same projection argument.

In the following, we assume that for all $m \in \mathcal{M}$, $S_m \subset S_{m^\star}$, that is to say, $S_{m^\star}$ is the maximum model. We also assume that the maximum model is a Cartesian product in the following sense:

$$S_{m^\star} = \{t : \mathbb{R}^2 \to \mathbb{R}, \quad t(x,y) = \sum_{j \in J_m^\star} \sum_{k \in K_m^\star} a_{jk} \varphi_j(x) \varphi_k(y),\ a_{jk} \in \mathbb{R}\}$$

with $D_{m^\star} = D_{m_1^\star} D_{m_2^\star} = |J_{m^\star}||K_{m^\star}|$.

2.3 Adaptation for the integrated risk

2.3.1 Study of the risk and model selection

For each model S_m, we denote by

$$\hat{s}_m = \underset{t \in S_m}{\text{argmin}}\ \gamma_n(t)$$

the associated estimator. We can prove

Proposition 5 ([L11]). *Under assumption* **(A)**, *there exists a constant $C(s,f)$ such that, for all $m \in \mathcal{M}$,*

$$\mathbb{E}\left[\|s - \hat{s}_m\|_n^2\right] \leqslant C(s,f) \left\{ d_f^2(s, S_m) + \frac{D_m}{n} \right\}.$$

We can recognize in the right hand side the usual bias-variance decomposition. The first term is a bias term, that will be small if S_m is a large approximation space, but then the second term will be big. If instead we try to minimize the stochastic error term, the bias will be large. So the goal is to choose the best estimator in the collection $\{\hat{s}_m\}_{m \in \mathcal{M}}$, the one that achieves the best bias-variance trade-off. If s has smoothness α, one can prove that the bias $d_f(s, S_m)$ is of

order $D_m^{-\alpha/2}$. If one knows the regularity of s, then it is not hard to choose S_m that minimizes $D_m^{-\alpha} + D_m/n$. Then we obtain that the estimation error $\mathbb{E}\|s - \hat{s}_m\|_n^2$ decreases as $n^{-2\alpha/(2\alpha+2)}$ which is the usual rate of convergence for a bivariate function. But this assumption of knowledge of the smoothness is obviously unrealistic. The goal here is to provide a procedure that adapts to the smoothness of s. To do this, we use the model selection method introduced by Birgé and Massart, which will allow us to select a data-driven estimator, by minimizing a penalized criterion. Thus, we consider the random selection procedure

$$\hat{m} = \underset{m \in \mathcal{M}}{\operatorname{argmin}} \{\gamma_n(\hat{s}_m) + \operatorname{pen}(m)\} \tag{3}$$

and the penalized estimator

$$\tilde{s} = \hat{s}_{\hat{m}}, \tag{4}$$

where pen : $\mathcal{M} \to \mathbb{R}^+$ is called a penalty, and it remains to choose it such that $\tilde{s}$ is a good estimator. Note that, with the notations of Lemma 3, $\gamma_n(\hat{s}_m) = \operatorname{tr}(-{}^t Z \hat{A})$: that makes this procedure easy to implement in practice.

2.3.2 Lower bound

In this section, we will show that the aforementioned rate of convergence $n^{-\alpha/(2\alpha+2)}$ is the best we can get. Since we study a bivariate function, it may have different smoothnesses in x and y axes: we call this anisotropic smoothness. We will see that it is then the harmonic mean of these two smoothnesses that plays the role of α. Considering anisotropy is particularly meaningful in our case of a conditional density. Indeed, the role of abscissa and ordinate are very different and there is no reason for the smoothnesses to be identical. For a bibliography on anisotropy in function estimation, see for example Autin et al. (2014) or Lepski (2014). So we will consider functions with smoothness $\alpha = (\alpha_1, \alpha_2)$, in the sense of Besov spaces. Let us recall the definition of anisotropic Besov spaces and the associated norm $\|.\|_{B^\alpha_{pp'}}$.

Let $A = [0,1]^2$ and e_1 and e_2 be the canonical basis vectors in $\mathbb{R}^2$ and for $i = 1, 2$, $A^r_{h,i} = \{x \in \mathbb{R}^2; x, x + he_i, \ldots, x + rhe_i \in A\}$. Next, for x in $A^r_{h,i}$, let

$$\Delta^r_{h,i} g(x) = \sum_{k=0}^{r} (-1)^{r-k} \binom{r}{k} g(x + khe_i)$$

the r-th difference operator with step h. For $t > 0$, the directional moduli of smoothness are given by

$$\omega_{r_i,i}(g,t) = \sup_{|h| \leqslant t} \left(\int_{A^{r_i}_{h,i}} |\Delta^{r_i}_{h,i} g(x)|^p dx \right)^{1/p}.$$

The norm of the Besov space $B^\alpha_{p,p'}(A)$ is defined by

$$\|g\|_{B^\alpha_{pp'}} = \begin{cases} \|g\|_p + \left(\int \left(\sum_{i=1}^{2} t^{-\alpha_i} \omega_{r_i,i}(g,t) \right)^{p'} \frac{dt}{t} \right)^{1/p'} & \text{if } p' < \infty \\ \|g\|_p + \sup_{t>0} \sum_{i=1}^{2} t^{-\alpha_i} \omega_{r_i,i}(g,t) & \text{if } p' = \infty \end{cases}$$

for r_i integers larger than α_i. We say that g belongs to the Besov space $B^\alpha_{p,p'}(A)$ if $\|g\|_{B^\alpha_{pp'}} < \infty$.

It is known that for $p < 1$, the spaces $\mathbb{L}^p$ are quite unusual, so we always assume that $p \geqslant 1$, and for the same reason that $p \geqslant 1$. The anistropic Besov balls are then

$$B(\alpha, p, p', R) = \{t : [0,1]^2 \to \mathbb{R} \text{ such that } \|t\|_{B^\alpha_{pp'}} \leqslant R\}.$$

We recall that the harmonic mean $\bar{\alpha}$ of α is defined by

$$\frac{1}{\bar{\alpha}} = \frac{1}{2}\left(\frac{1}{\alpha_1} + \frac{1}{\alpha_2}\right).$$

In the isotropic case $\alpha_1 = \alpha_2 = \alpha$, the harmonic mean $\bar{\alpha}$ is simply equal to α. We can now write the following result of lower bound, which generalizes that of Birgé (1983).

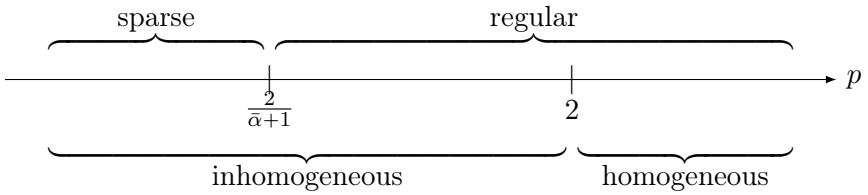
Theorem 6. *Let $1 \leqslant p < \infty$ and $1 \leqslant p' \leqslant \infty$. Assume that $\bar{\alpha}/2 > 1/p - 1/2$. Then, there exists $C > 0$ such that, for n large enough,*

$$\inf_{\hat{s}_n} \sup_{s \in B(\alpha,p,p',R)} \mathbb{E}_s \|\hat{s}_n - s\|_2^2 \geqslant C R^{\frac{2}{\bar{\alpha}+1}} n^{-\frac{2\bar{\alpha}}{2\bar{\alpha}+2}}$$

where the infimum is taken over all estimators $\hat{s}_n$ of s based on data $(X_1, Y_1), \ldots, (X_n, Y_n)$.

So we cannot expect a best rate than $n^{-\frac{2\bar{\alpha}}{2\bar{\alpha}+2}}$ for an estimator. In the following, we will show that our estimator achieves this rate of convergence. This proves that it is the minimax rate of convergence.

Actually there is an elbow phenomenon, which is usual in nonparametric estimation (see Härdle et al., 1998). The case stated in the theorem is the regular case, when $p > \frac{2}{\bar{\alpha}+1}$. It is opposed to the "sparse" case: very regular functions in general but with some sharply localized singularities (the name comes from the sparsity of the description in term of wavelet coefficients). In this case the minimax rate has a logarithmic factor, and is of the form $(\log n/n)^\delta$. We will not consider such functions in this manuscript. We will only study "regular" functions, which themselves can be divided into two cases. Functions with regular smoothness ($p \geqslant 2$) can be estimated by a linear estimator, while in the opposite case, the linear estimators are insufficient and we will use other approximation spaces.



Finally, the space of values for p is divided into three zones:

1. $p \leqslant 2/(\bar{\alpha}+1)$: "sparse" zone, the minimax rate has a logarithmic factor, and is of the form $(\log n/n)^\beta$. This case is not treated here.

2. $2/(\bar{\alpha}+1) < p < 2$: intermediate zone (regular but inhomogeneous). Since $p > 2/(\bar{\alpha}+1)$, the rate is polynomial, but since $p < 2$, the smoothness is inhomogeneous. In this case, linear estimators are not sufficient. Since, for $p \geqslant 1$, $B^\alpha_{p,p'} \subset B^\alpha_{1,p'} \subset B^\alpha_{1,\infty}$, it is sufficient to consider $s \in B^\alpha_{1,\infty}$.

3. $p \geqslant 2$: homogeneous regular zone. The rate is still polynomial, but moreover we manage to estimate s by linear estimators. In this case, we consider $p' = \infty$. Indeed we recall that $B^\alpha_{p\infty}$ contains all spaces $B^\alpha_{pp'}$, for $p' > 0$. In addition we have $B^\alpha_{p,\infty} \subset B^\alpha_{2,\infty}$ for any $p \geqslant 2$. In this case, it is sufficient to consider that $s \in B^\alpha_{2,\infty}$, which includes all others.

2.3.3 Homogeneous smoothness

First, we assume that $p \geqslant 2$ (homogeneous smoothness). In this case, the bias can be bounded by $\|s - s_m\| \leqslant C(D_{m_1}^{-\alpha_1} + D_{m_2}^{-\alpha_2})$ using simple linear spaces (this will not be true when $p < 2$). Nevertheless, to manage the anisotropy of s, we need anisotropic models.

We then consider partitions of the square into dyadic rectangles of the form $[\frac{j-1}{2^{m_1}}, \frac{j}{2^{m_1}}[\times[\frac{k-1}{2^{m_2}}, \frac{k}{2^{m_2}}[$ (by closing the intervals on the right when $j = 2^{m_1}$ or $k = 2^{m_2}$). The partition m is then fully determined by the number of cuts m_1 on the x-axis and the number of cuts m_2 on the y-axis. S_m can be written $S_m = E_{m_1} \otimes H_{m_2}$ with $E_{m_1} = \mathrm{Vect}(\varphi_j^m,\ j \in J_m)$, $H_{m_2} = \mathrm{Vect}(\varphi_k^m,\ k \in K_m)$ and $D_m = D_{m_1} D_{m_2} = 2^{m_1}(r+1)2^{m_2}(r+1)$. We denote by $\mathcal{M}^{reg}$ the set of partitions of this form*.

A useful property of these models is that they satisfy the norm connection between the $\mathbb{L}^2$ norm and the infinite norm: there exists $\phi_0 = 2r + 1 > 0$ such that

$$\forall t \in S_m \qquad \|t\|_\infty \leqslant \phi_0 \sqrt{D_m} \|t\|.$$

Moreover, we are always in the case where the directional sub-models E and H are nested ($D_{m_1} \leqslant D_{m_1'} \Rightarrow E_{m_1} \subset E_{m_1'}$ and $D_{m_2} \leqslant D_{m_2'} \Rightarrow H_{m_2} \subset H_{m_2'}$). We use dyadic intervals to ensure that property. Then we have that, for all m and m' in $\mathcal{M}$, $S_m + S_{m'}$ is included in a model ($S_m + S_{m'} \subset S_{m''}$ with $D_{m_1''} = \max(D_{m_1}, D_{m_1'})$ and $D_{m_2''} = \max(D_{m_2}, D_{m_2'})$) but this model has large dimension because of the anisotropy. For the same reason there is not an only model per dimension (one may have $D_{m_1} D_{m_2'} = D_{m_1'} D_{m_2}$), but there is an only directional sub-model per sub-dimension. Thus, the essential ingredient for the proof of the following theorem is

$$\sum_m e^{-K\sqrt{D_m}} = \sum_{m_1}\sum_{m_2} e^{-K\sqrt{D_{m_1} D_{m_2}}} \leqslant \sum_{m_1} e^{-(K/2)\sqrt{D_{m_1}}} \sum_{m_2} e^{-(K/2)\sqrt{D_{m_2}}} < \infty.$$

Then we can state the following result.

Theorem 7 ([L6], improved version)**.** *We assume that assumption* (**A**) *is verified, and* $\mathcal{M} \subset \mathcal{M}^{reg}$ *and*

$$\forall m \in \mathcal{M} \qquad D_{m_1} \lesssim n/\log^2(n) \quad \text{and} \quad (\log n)^3 \lesssim D_{m_1} D_{m_2} \lesssim n.$$

For $\gamma > 0$*, we define the estimator of* s *by* (3) *and* (4) *with penalty*

$$\mathrm{pen}(m) = (1+\gamma)^2 \|s\|_\infty \frac{D_{m_1} D_{m_2}}{n}.$$

Then, with probability larger than $1 - C_0 \exp\{-(\log n)^{5/4}\}$,

$$\|\tilde{s} - s\|_n^2 \leqslant \inf_{m \in \mathcal{M}} \left(C_1 d_f^2(s, S_m) + C_2 \mathrm{pen}(m) \right)$$

where $C_1 > (1 + 2\gamma^{-1})^2$*,* $C_2 > 2(1 + 2\gamma^{-1})$ *and* C_0 *depends on* $\|s\|_\infty, \|f\|_\infty, f_0, r,\ \gamma$*. Moreover*

$$\mathbb{E}\|\tilde{s} - s\|_n^2 \leqslant C_3 \inf_{m \in \mathcal{M}} \left(d_f^2(s, S_m) + \frac{D_{m_1} D_{m_2}}{n} \right) + \frac{C_4}{n}$$

where C_3 *depends on* $\|s\|_\infty$*,* γ*, and* C_4 *depends on* $\|s\|_\infty, \|f\|_\infty, f_0, r,\ \gamma$*.*

*In this section 2.3.3 could also be used as an approximation basis trigonometric polynomials or wavelets.

Before discussing the proof, let us make a few remarks about the penalty. The issue of calibration of γ will be mentioned later, Section 2.6. The presence of $\|s\|_\infty$ is obviously problematic. This term appears because of the anisotropy, it could be avoided if we only used isotropic spaces. In practice, we replace it by an upper bound, or by $\|\hat{s}\|_\infty$ where $\hat{s}$ is an estimator of s. Then we can prove the same result by adding some regularity assumption. We write this in more detail in the next section.

Note on the proof:
We observe that for all functions t_1, t_2

$$\gamma_n(t_1) - \gamma_n(t_2) = \|t_1 - s\|_n^2 - \|t_2 - s\|_n^2 - 2\nu(t_1 - t_2)$$

where

$$\nu(t) = \frac{1}{n}\sum_{i=1}^{n}\left\{t(X_i, Y_i) - \int_{\mathbb{R}} t(X_i, y)s(X_i, y)dy\right\} = \frac{1}{n}\sum_{i=1}^{n}\left\{t(X_i, Y_i) - \mathbb{E}[t(X_i, Y_i)|X_i]\right\}.$$

The heart of the proof lies in the study of the centered empirical process ν, in particular in the control of $\sup_{\substack{t\in S_m \\ \|t\|_f=1}} \nu^2(t)$. We are reduced to show that, with great probability,

$$\forall m, m' \in \mathcal{M}, \qquad \sup_{\substack{t\in S_m + S_{m'} \\ \|t\|_f=1}} \nu^2(t) \leqslant \frac{\text{pen}(m)}{1+\gamma} + \frac{\text{pen}(m')}{1+\gamma}.$$

To do this, it is sufficient to use Talagrand inequality. Finally we use $\dim(S_m + S_{m'}) \leqslant \dim(S_m) + \dim(S_{m'})$. This simple inequality is actually crucial and allows to easily manage the anisotropy. ■

From Theorem 7, we can deduce the minimax rate of convergence of the risk.

Corollary 8. *Assume that s belongs to the ball $B(\alpha, 2, \infty, R)$ with smoothness $\alpha = (\alpha_1, \alpha_2)$ with $\alpha_1 > 0$ and $\alpha_2 > 0$ †. We consider that the maximal degree r of the polynomials is larger than $\alpha_i - 1$. Then, under assumptions of the above theorem,*

$$\mathbb{E}\|s - \tilde{s}\|_n^2 \leqslant CR^{\frac{2}{\bar{\alpha}+1}} n^{-\frac{2\bar{\alpha}}{2\bar{\alpha}+2}}.$$

Thus we obtain the rate of convergence $n^{-\frac{2\bar{\alpha}}{2\bar{\alpha}+2}}$, which is optimal in the minimax sense (see Theorem 6). The estimation procedure allows an adaptation of the approximation space to each directional regularity. For example, if $\alpha_2 > \alpha_1$, then the procedure chooses a space of dimension $D_{m_2} = D_{m_1}^{\alpha_1/\alpha_2} < D_{m_1}$.

The empirical norm is the more natural in this problem, but if we were interested in a $\mathbb{L}^2$ control of the risk, we may modify the estimation procedure as follows:

$$\tilde{s}^* = \begin{cases} \tilde{s} & \text{if } \|\tilde{s}\|_2 \leqslant n^{2/3}, \\ 0 & \text{otherwise.} \end{cases}$$

We can prove a result similar to Theorem 7 but bounding $\mathbb{E}\|\tilde{s}^* - s\|^2$ instead of its empirical version:

† If we don't use a localized basis (for example the trigonometric basis), the condition on D_{m_1} in Theorem 7 is stronger, which entails that Corollary 8 is true only for $\alpha_1 > 1/2$ and $\alpha_2 > 1/2$.

Corollary 9. *Under assumptions of Theorem 7, then*

$$\mathbb{E}\|s - \tilde{s}^*\|_2^2 \leqslant C^* \inf_{m\in\mathcal{M}}\{d^2(s, S_m) + \text{pen}(m)\} + \frac{C'^*}{n}$$

where C^* depends on $\gamma, f_0, \|f\|_\infty$ and C'^* depends on $r, \|s\|_\infty, f_0, \|f\|_\infty, \gamma$. Moreover, under assumptions of Corollary 8,

$$\mathbb{E}\|s - \tilde{s}^*\|_2^2 \leqslant CR^{\frac{2}{\bar{\alpha}+1}} n^{-\frac{2\bar{\alpha}}{2\bar{\alpha}+2}}.$$

Simultaneously to my early work (2007), other papers have studied the conditional density estimation with a requirement of theoretical results for adaptation. Efromovich (2007, 2010b) presents an estimator by Fourier decomposition and preliminary estimate of f by an estimator $\hat{f}$ and then block thresholding (blockwise-shrinkage Efromovich Pinsker estimator). He gets an oracle inequality and finds the same minimax rate of convergence in the context of Sobolev anisotropic spaces or analytic functions. However, it must be assumed that f is differentiable with bounded derivative. He was also interested in the particular case where Y is in fact independent of X and therefore $s(x, y)$ depends only on y and not on x.

More recently, Chagny (2013) also searched for oracles inequality to estimate the conditional density. Her estimator uses a preliminary estimator $\hat{F}$ of the distribution function F of X and is written

$$\hat{s}(x, y) = \sum_{j,k} \hat{a}_{jk}\varphi_j(\hat{F}(x))\varphi_k(y)$$

with $\hat{a}_{jk} = \frac{1}{n}\sum_{i=1}^n \varphi_j(\hat{F}(X_i))\varphi_k(Y_i)$ and (φ_j) the Fourier basis. This projection method has the advantage of not requiring the implementation of matrix inversion. The chosen method for adaptation is that of Goldenshluger-Lepski.

Another very recent work is that of Cohen and Le Pennec (2013) with a maximum likelihood approach

$$\hat{s}_m = \underset{t\in S_m}{\text{argmin}}\ \{-\sum_{i=1}^n \log(t(X_i, Y_i))\}$$

followed by Birgé-Massart model selection. They give an oracle inequality for tensored and convexified Kullback divergence. The method is implemented for spaces of square root of polynomials on tree-structured partitions, or mixtures of Gaussian. It is applied to the segmentation of hyperspectral images.

Even more recently, Bayesian studies of the subject have been published: see Scricciolo (2015) and references therein. We can also cite a study that focuses on the case of noisy data: Wang and Ye (2015).

These recent works are interesting but (except Clémençon (2000)) do not address the case of functions with non-homogeneous smoothness. So I would like to handle now this case of inhomogeneous smoothness. It is then required to use irregular models.

2.3.4 Inhomogeneous smoothness

When $p < 2$ (inhomogeneous case), it is impossible to have a good approximation of s by a projection onto a linear subspace. To get the same kind of result $\|s - s_m\| \leqslant C(D_{m_1}^{-\alpha_1} + D_{m_2}^{-\alpha_2})$, it is necessary to have nonlinearity, that is to say, given a dimension, the possibility to choose several linear models of the same dimension. Figure 1 allows to get an idea of what is an inhomogeneous smoothness.

Here, we still consider as a model S_m a space of piecewise polynomial functions with degree smaller than or equal to r, but this time, the pieces will not be necessarily of the same size. For

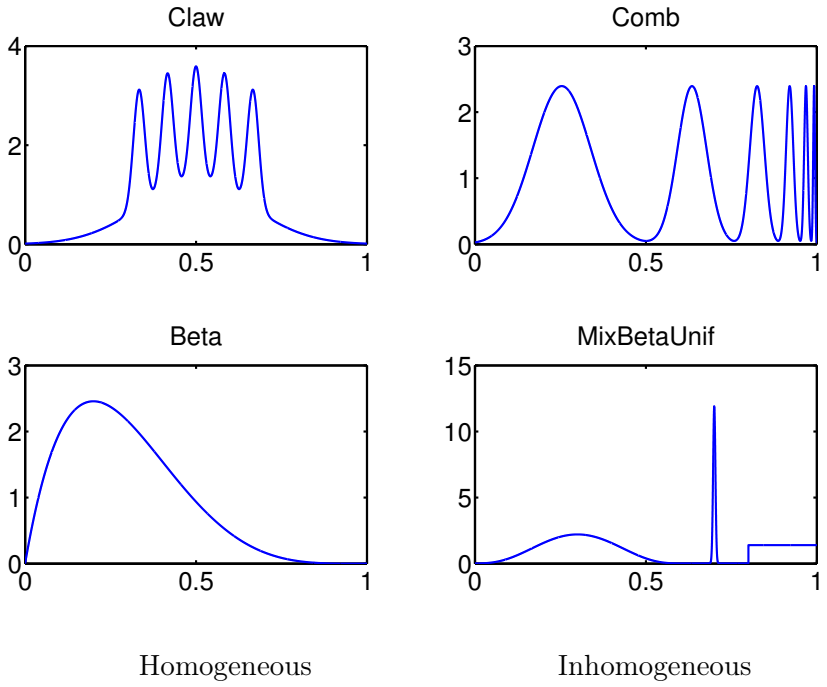


Figure 1: Example of homogeneous and inhomogeneous functions

the moment, we no longer assume anything on the set of partitions $\mathcal{M}$. Then we can prove the following result.

Theorem 10 ([L11]). *We assume that* (**A**) *is verified, and that the maximum model* $S_{m^\star}$ *is based on a regular partition in squares such that* $D_{m^\star} \lesssim \sqrt{n}$. *We also assume that there exists* $\{L_m\}_{m\in\mathcal{M}}$ *a family of reals greater than or equal to 1, that may depend on* n, *such that*

$$\sum_{m\in\mathcal{M}} \exp(-L_m D_m) \leqslant 1. \tag{5}$$

For some large enough positive absolute constant κ, *we choose*

$$\mathrm{pen}(m) = \kappa \left(\|s\|_\infty + \frac{(2r+1)^2}{f_0} \right) \frac{L_m^2 D_m}{n}$$

Then

$$\mathbb{E}\left[\|\tilde{s} - s\|_n^2\right] \leqslant C \left(\max_{m\in\mathcal{M}} L_m^2 \right) \min_{m\in\mathcal{M}} \left\{ d_f^2(s, S_m) + \frac{D_m}{n} \right\}.$$

where C *only depends on* κ, r, $\|s\|_\infty$, $f_0, \|f\|_\infty$.

This result generalizes the previous theorem, which corresponds to $L_m = cst/(\|s\|_\infty + (2r+1)^2 f_0^{-1})$. The price for this generality is the appearance of f_0 in the penalty.

Here we will detail the replacement in the penalty of $\|s\|_\infty$ and f_0 by $\|\hat{s}\|_\infty$ and $\hat{f}_0$. We choose $m^\bullet = m_1^\bullet \times m_2^\bullet$ a regular partition into cubes such that $|m^\bullet|^2 \leqslant n$, we denote by $\hat{f}_{m_1^\bullet}$ the classical projection estimator of f, and $\hat{f}_0 = \max(\inf_{[0,1]} \hat{f}_{m_1^\bullet}, 1/n)$.

We denote $\min(\alpha) = \min(\alpha_1, \alpha_2)$ and $\bar{\alpha}$ the harmonic mean of α_1, α_2. Then we can write

Corollary 11. *Assume that* $s \in B(\alpha, p, p', R)$ *and* $f \in B(\beta, p, p', R_1)$ [‡] *with* $p' = \infty$ *if* $p = 1$ *or* $p \geqslant 2$, $p' = p$ *if* $1 < p < 2$, *and*

$$\frac{\bar{\alpha}}{2}\left(1 - \frac{1}{\min(\alpha)}\right) > \frac{1}{p}, \qquad \beta > \left(\frac{1}{p} - \frac{1}{2}\right)_+ + 1.$$

[‡]Besov space in dimension 1 are defined as in dimension 2.

Assume that $|m_1^\bullet| \geqslant \ln n$ *and, for all* $m \in \mathcal{M}$,

$$\mathrm{pen}(m) = \bar{\kappa}\left(\|\hat{s}_{m^\bullet}\|_\infty + \frac{(2r+1)^2}{\hat{f}_0}\right)\frac{L_m^2 D_m}{n}$$

for some positive constant $\bar{\kappa}$ *large enough. Then, under assumptions of Theorem 10, for* n *large enough,*

$$\mathbb{E}_s\left[\|s-\tilde{s}\|_n^2\right] \leqslant C_1'\left(\max_{m\in\mathcal{M}} L_m^2\right)\min_{m\in\mathcal{M}}\left\{d_f^2(s,S_m)+\frac{D_m}{n}\right\}.$$

where C_1' *is a positive constant depending only on* $\bar{\kappa}$, r, $\|s\|_\infty$,f_0, $\|f\|_\infty$.

We have already explained that we need irregular partitions to estimate inhomogeneous functions. However, irregular partitions often form a too rich collection. If L_m only depends on D_m, Condition (5) means that L_m has to be large enough to balance the number of models of same dimension D_m. If the number of models for each dimension is high, the L_m's have to be high too. For instance, Birgé and Massart (1997) use weights $(L_m)_{m\in\mathcal{M}}$ of order $\log(n)$ to ensure Condition (5), which spoils the rates of convergence.

Here we use an especially interesting collection of partitions, for which the factor $\max_{m\in\mathcal{M}} L_m^2$ can be bounded by a constant, although the collection is rich enough to have good approximation qualities with respect to functions of inhomogeneous smoothness. Let us describe this collection. We call dyadic rectangle of $[0,1]^2$ any set of the form $I_1 \times I_2$ where, for $l=1$ or $l=2$,

$$I_l = [(k_l-1)2^{-j_l}, k_l 2^{-j_l}[$$

with $j_l \in \mathbb{N}$ and $k_l \in \{1,\ldots,2^{j_l}\}$ [§]. Otherwise said, a dyadic rectangle of $[0,1]^2$ is defined as a product of two dyadic intervals of $[0,1]$ that may have different lengths. We denote by $\mathcal{M}^{irreg}$ such a collection of partitions. Let us underline that a partition of $\mathcal{M}^{irreg}$ may be composed of rectangles with different Lebesgue measures, as illustrated by Figure 2. This was not the case for the models used in the previous section. The partitions were then composed of rectangles with all the same size $2^{-m_1}2^{-m_2}$.

We consider the collection of partitions of $[0,1]^2$ into dyadic rectangles with side $\geqslant 2^{-J^\star}$ where $2^{2J^\star} = |m^\star| \leqslant \sqrt{n}$, so that the assumptions of Theorem 10 are verified.

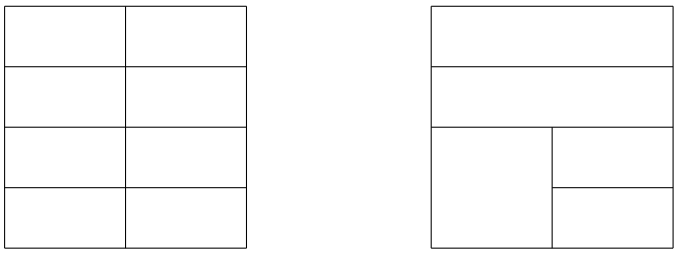


Figure 2: A partition in $\mathcal{M}^{reg}$ (left) and a partition in $\mathcal{M}^{irreg}$ (right)

Notice that choosing such a partition into D dyadic rectangles amounts to choosing cutting directions and a binary tree with D leaves. For instance, the tree corresponding to the above figure is represented in Figure 3.

[§]and the interval is closed on the right if $k_l = 2^{j_l}$.

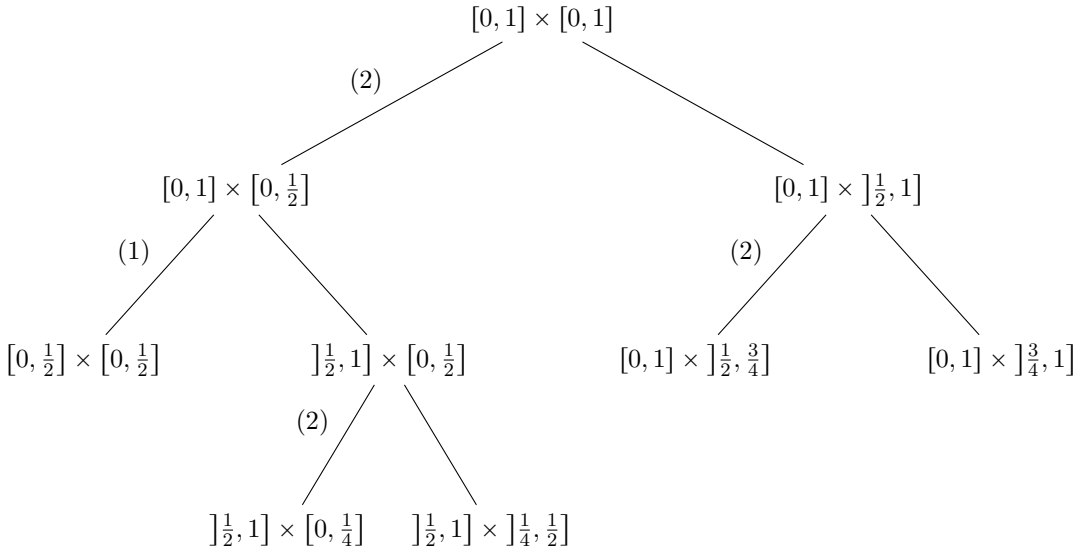


Figure 3: Binary tree labeled with the sequence of cutting directions $(2,1,2,2)$ corresponding with the dyadic partition represented on the right side of Figure 2.

Hence it can be deduced that for this collection of partitions, $L_m = \log(16)$ works. This gives the following result. If $\tilde{s}$ is built using $\mathcal{M} \subset \mathcal{M}^{irreg}$, then

$$\mathbb{E}_s\left[\|s-\tilde{s}\|_n^2\right] \leqslant C_2 \min_{m\in\mathcal{M}}\left\{d_f^2(s,S_m)+\frac{D_m}{n}\right\}$$

where C_2 is a positive real number only depending on κ, r, $\|s\|_\infty$,f_0, $\|f\|_\infty$.

We are now able to compute estimation rates, using approximation results of Akakpo (2012). Let

$$q(\alpha,p)=\frac{\min(\alpha)}{\bar{\alpha}}\frac{1+\bar{\alpha}}{\bar{\alpha}}\left(\frac{\bar{\alpha}}{2}-\left(\frac{1}{p}-\frac{1}{2}\right)_+\right).$$

Contrary to Klemelä (2009), we have chosen a parameter $J^\star$ that does not depend on the unknown smoothness of s, hence the factor $\lambda = \min(\alpha)/\bar{\alpha}$ in the above definition. That factor, which is inferior or equal to 1 with equality only in the isotropic case, may be interpreted as an index measuring the lack of isotropy. We assume that $q(\alpha,p)>1$, which is equivalent to

$$\bar{\alpha} > \begin{cases} \frac{2}{\lambda}-1 & \text{if } p=2 \\ \frac{1}{\lambda}+\sqrt{\frac{1}{\lambda^2}+1} & \text{if } p=1 \end{cases} \qquad \text{where } \lambda=\min(\alpha)/\bar{\alpha}$$

(for instance in the isotropic case and $p=2$, it means $\alpha>1$).

As explained above, if $p \geqslant 2$, $B^\alpha_{p,p'} \subset B^\alpha_{2,\infty}$, and if $1 \leqslant p < 2$, $B^\alpha_{p,p'} \subset B^\alpha_{1,\infty}$, then we only consider $p=2$ or $p=1$ and $p'=\infty$.

Theorem 12 ([L11])**.** *We assume that* (**A**) *is verified, and* $\mathcal{M} \subset \mathcal{M}^{irreg}$*, and the maximum model verifies*

$$D_{m^\star} \lesssim \sqrt{n}.$$

Assume that $s \in B(\alpha,p,\infty,R)$ *with* $p=1$ *or* $p=2$*, and* $q(\alpha,p)>1$*. If* $n^{-1} \leqslant R^2 \leqslant n^{q(\alpha,p)-1}$*, then there exists a postive real number* $C(\alpha,r,p)$ *such that*

$$\mathbb{E}_s\left[\|s-\tilde{s}\|_n^2\right] \leqslant C_2 C(\alpha,r,p)\|f\|_\infty\left(R^2 n^{-\bar{\alpha}}\right)^{2/(2\bar{\alpha}+2)}.$$

We can extend this result to higher dimensions (voir [L11]). The rate $\left(R^2 n^{-\bar{\alpha}}\right)^{1/(2\bar{\alpha}+2)}$ is the minimax one. Here we are able to reach that rate not only for functions with homogeneous smoothness, but also for functions with inhomogeneous smoothness, *i.e.* for $0 < p < 2$. It was impossible with the collection of regular models considered above. Besides, among the cited references, only Klemelä (2009) can deal simultaneously with anisotropy and inhomogeneous smoothness (but in the context of density estimation). Here we improve its result by allowing to approximately reach the minimax risk up to a factor that does not depend on n and considering smoothness parameters possibly larger than 1.

I also would like to mention some work published after ours. Birgé (2013) shows that his estimation method from a family of tests (T-estimators) can be applied to the case of the conditional density, and thus provides an oracle inequality. Unfortunately, it is more a theoretical than a practical method. In this line, the work of Sart (2014) is more implementable. It deals with the estimation of the transition of a Markov chain by a piecewise constant function on a random partition. The model selection is a mixture between a contrast minimization and a test procedure *à la* Birgé and Baraud. This method allows the author to reduce the number of assumptions (even if the lower bound condition for the stationary density is necessary when using $\mathbb{L}^2$ loss rather than Hellinger loss), and to avoid unknown quantities in the penalty, at the cost of a logarithmic loss in the rate. His simulations are fast (linear complexity in n) and promising.

2.3.5 Numerical illustrations

The implementation of our estimation procedure is quite simple. Indeed, when the model is written in the form of a Cartesian product (regular partitions), using again Lemma 3, a matrix inversion allows to find $\hat{s}_m$ and then $\hat{m} = \arg\min_{m\in\mathcal{M}}\{-\mathrm{tr}({}^t Z\hat{A}) + \kappa\|\hat{s}_{m\bullet}\|_\infty D_m/n\}$ where $\|\hat{s}_{m\bullet}\|_\infty$ is an estimator of $\|s\|_\infty$. In the case of irregular partitions, there is an equivalent of Lemma 3 for each rectangle of the partition. Thus

$$\hat{m} = \underset{m\in\mathcal{M}}{\mathrm{argmin}} \sum_{R\in m} \{-\mathrm{tr}({}^t Z_R \hat{A}_R) + \kappa\frac{\|\hat{s}_{m\bullet}\|_\infty}{n}\}$$

where Z_R and $\hat{A}_R$ are $(r+1)\times(r+1)$ matrices, restrictions of Z and $\hat{A}$ to the rectangle R. That characterization allows to determine $\hat{m}$ without having to compute all the estimators of the collection $\{\hat{s}_m\}_{m\in\mathcal{M}}$. Indeed, we can for instance adapt to our estimation framework the algorithm proposed by Donoho (1997), which gives a computational complexity at most linear in the number of observations. The implementation, however, requires the choice of κ in the penalty. In the proofs, an upper bound of this constant is obtained, but it is unfortunately very rough and useless in practice. The calibration of this constant is a sensitive topic that we will discuss in Section 2.6.

Figures 4 and 5 illustrate the simulation results (with a histogram basis on $r = 0$), with κ calibrated by hand on a preliminary set of examples.

2.4 Pointwise adaptation

2.4.1 Specific Motivation

Here we are interested in estimating the conditional density s of Y knowing $X = x$ at a given point x. The motivation comes from methods used in population genetics. The purpose of these statistical approaches in population genetics is to infer the evolutionary processes that generated the observed data, typically the gene pool of studied populations, and to possibly

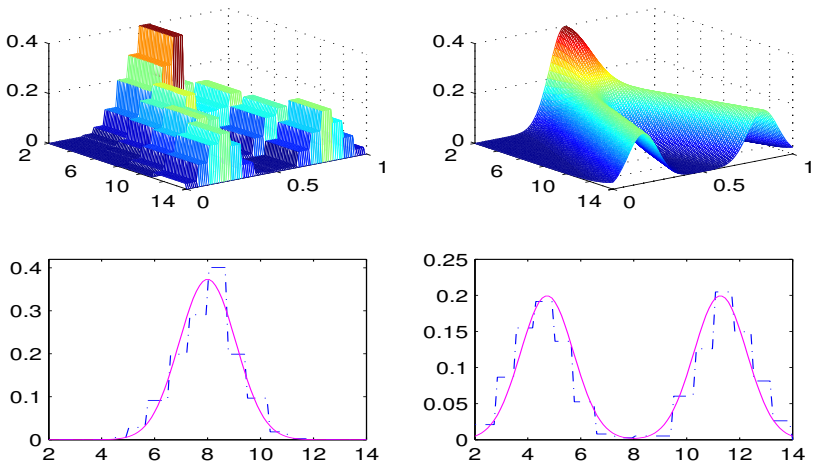


Figure 4: Estimation when the X_i's are i.i.d. $\mathcal{U}([0,1])$ and, given $X_i = x$, $Y_i \sim 0.5\mathcal{N}(8 - 4x, 1) + 0.5\mathcal{N}(8 + 4x, 1)$. Top left: Estimator using $\mathcal{M}^{reg}$ for $n = 2000$. Top right: True conditional density s. Bottom : two sections of s together with the corresponding sections of $\tilde{s}$ for $x = 0.1$ (bottom-left) and $x = 0.82$ (bottom-right).

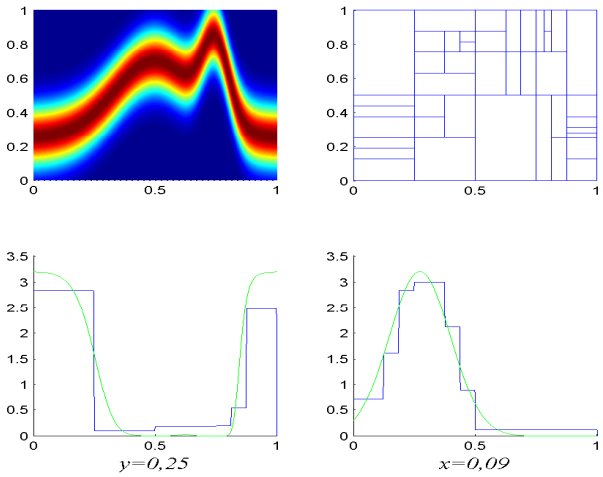


Figure 5: Estimation when the X_i's are i.i.d. $\mathcal{U}([0,1])$ and $Y_i = \frac{1}{4}(g(X_i) + 1) + \frac{1}{8}\epsilon_i, i = 1, \ldots, n$ where $(\epsilon_i)_{1 \leqslant i \leqslant n}$ are i.i.d. standard normal and g is the density of $\frac{3}{4}\mathcal{N}(1/2, (1/6)^2) + \frac{1}{4}\mathcal{N}(3/4, (1/18)^2)$. Top left: Level lines of the conditional density s. Top right: selected partition for $n = 1000$. Bottom : two sections of s together with the corresponding sections of $\tilde{s}$.

infer phylogenetic trees. Classically, used statistical methodologies are based on the likelihood but the complexity of population genetics data often makes the computation or even writing the likelihood impossible. ABC methods (for Approximate Bayesian Computation) have been developed to address this problem. Proposed in population genetics where the data live in very large dimension, they have spread to other areas such as social sciences, ecology and in other areas of biology (see the survey Marin et al. (2012)). The standard ABC procedure is very intuitive and consists in

- simulating a lot of parameters values using the prior distribution and, for each parameter value, a corresponding dataset,
- comparing this simulated dataset to the observed one;
- finally, keeping the parameter values for which distance between the simulated dataset and the observed one is smaller than a tolerance level.

That is a crude nonparametric approximation of the target posterior distribution (the conditional distribution of the parameters given the observation). Even if some nonparametric perspectives have been considered (see Blum (2010) or Biau et al. (2012)), we easily imagine that, using the simulated couples (parameters and datasets), a good nonparametric estimation of the posterior distribution can be a credible alternative to the ABC method. Such a procedure has to consider that the conditional density has to be estimated only for the observed value in the conditioning.

So the ABC algorithms show that it is necessary that the regularization parameter m depends on the point x. All the methods discussed above allow to globally select the "best" $\hat{m}$, then used for all x. Here, we would like that this $\hat{m}$ depends on the estimation point x. Indeed, according to the point x, the distribution of Y given $X = x$ can be easily approximated by a simple model or requires a more complex model. In this section, we would therefore introduce an estimator with the same qualities as before (easily implementable and fully data-driven) but that also satisfies non-asymptotic oracle inequalities depending on the estimation point x. We will only consider a function s with homogeneous smoothness and regular models. The assumption of lower bound of the marginal density f will be needed again, even if it is now limited to a small neighborhood of x. However, we try to study the presence of f_0 in detail, and show that it is unavoidable in some sense. To do this, in addition to studying previous contrast-minimization estimator, we also introduce a kernel estimator in Section 2.4.6. Since $y \to s(x, y)$ is a density, we will assess the quality of an estimator $\hat{s}$ at a given point $x \in \mathbb{R}$ and in $\mathbb{L}^2$ norm with respect to variable y. In other words, we will use the norm defined for all function t by

$$\|t\|_{x,2} = \left(\int_{\mathbb{R}} t^2(x, y) dy \right)^{1/2}.$$

We denote $V_n(x)$ the neighborhood of x on which we conducts the study. Given A a positive number and (k_n) a sequence of real numbers tending to infinity, we set

$$V_n(x) = \left[x - \frac{2A}{k_n}, x + \frac{2A}{k_n} \right].$$

Notice that the size of $V_n(x)$ tends to 0. The study will be made only on $V_n(x)$. In particular, Assumption (**A**) is now

Assumption (A) For all y in $[0, 1]$ and for all u in $V_n(x)$,

$$s(u, y) \leqslant \|s\|_\infty < \infty, \qquad 0 < f_0 := \inf_{V_n(x)} f \leqslant f(u) \leqslant \|f\|_\infty < \infty.$$

2.4.2 Estimators

We use the same estimator as before by minimizing the contrast γ_n on a model S_m, where $m \in \mathcal{M}^{reg}$. The only difference is that this time we make a partition of the rectangle $V_n(x) \times [0,1]$ instead of $[0,1]^2$ (for a partition m, $V_n(x)$ is divided into 2^{m_1} pieces and $[0,1]$ into 2^{m_2} pieces). Moreover, since x is given, for each m we can define the interval I^m among 2^{m_1} possible intervals, such that x belongs to the interval I^m. Hence we have

$$\hat{s}_m(x,y) = \sum_{K, I^m \times K \in m} \sum_{d_1,d_2=0}^{r} \hat{a}^m_{K,d_1,d_2} \varphi_{I^m,d_1}(x) \varphi_{K,d_2}(y)$$

with $G_m \hat{A}_m = Z_m$ where $\hat{A}_m = (\hat{a}^m_{K,d_1,d_2})_{0 \leqslant d_1,d_2 \leqslant r, I^m \times K \in m}$ and

$$G_m = \left(\frac{1}{n} \sum_{i=1}^{n} \varphi_{I^m,d_1}(X_i) \varphi_{I^m,d_2}(X_i) \right)_{0 \leqslant d_1,d_2 \leqslant r} \qquad Z_m = \left(\frac{1}{n} \sum_{i=1}^{n} \varphi_{I^m,d}(X_i) \varphi_{K,d_2}(Y_i) \right)_{0 \leqslant d_1,d_2 \leqslant r, I^m \times K \in m}$$

Then we define the estimator at point x by

$$\hat{A}_m := \begin{cases} (G_m)^{-1} Z_m & \text{if } \min(\mathrm{Sp}(G_m)) > (1+\gamma)^{-2/5} \widehat{f_0}, \\ 0 & \text{otherwise,} \end{cases}$$

where γ is a positive real number, and $\widehat{f_0}$ an estimate of f_0. Here, for a symmetric matrix M, $\mathrm{Sp}(M)$ denotes the spectrum of M, i.e. the set of its eigenvalues. This expression allows us to overcome problems if $\hat{G}_m$ is not invertible. Note that, when $r = 0$, where r is maximal degree of Legendre polynomials, this estimator can be written more simply:

$$\hat{s}_m(x,y) = \sum_{K, I^m \times K \in m} \frac{\mathrm{card}\{i, (X_i, Y_i) \in I^m \times K\}}{\mathrm{card}\{i, X_i \in I^m\}} \frac{\mathbb{1}_K(y)}{|K|}.$$

Then we have defined the collection of estimators from which to select the final estimator. To do this, we need a preliminary estimator of f denoted $\hat{f}$. We assume that $\hat{f}$ satisfies assumption

Assumption (B)

$$\widehat{f_0} := \inf_{t \in V_n(x)} |\hat{f}(t)| > 0.$$

and

$$\forall\, \lambda > 0, \quad \mathbb{P}\left(\sup_{t \in V_n(x)} \left| \frac{f(t) - \hat{f}(t)}{\hat{f}(t)} \right| > \lambda \right) \leqslant C \exp\{-(\log n)^{3/2}\},$$

where C is a constant only depending on λ and f.
It is shown in [L15] that we can find such an estimator if f belongs to a Hölder space. For kernel method (Section 2.4.6), we moreover need that $\hat{f}$ is independent of the data $(X_1, Y_1), \ldots, (X_n, Y_n)$. In this case it is assumed that we have available additional data $X_{n+1}, \ldots, X_{2n}$ needed to build this preliminary estimator $\hat{f}$.

2.4.3 Adaptation

We use the Goldenshluger-Lepski methodology mentioned in Section 1. Let us describe it in a more general context. Given a set of parameters $\mathcal{M}$, for any $m \in \mathcal{M}$, we assume we are given

a smoothing linear operator denoted $\mathcal{K}_m$ and an estimate $\hat{s}_m$. For any $m \in \mathcal{M}$, $\hat{s}_m$ is related to $\mathcal{K}_m(s)$ via its expectation and we assume that $\mathbb{E}[\hat{s}_m]$ is close to (or equal to) $\mathcal{K}_m(s)$. The main assumptions needed for applying the methodology are

$$\mathcal{K}_m \circ \mathcal{K}_{m'} = \mathcal{K}_{m'} \circ \mathcal{K}_m \tag{6}$$

and

$$\mathcal{K}_m(\hat{s}_{m'}) = \mathcal{K}_{m'}(\hat{s}_m) \tag{7}$$

for any $m, m' \in \mathcal{M}$. This method is a convenient way to select an estimate among $(\hat{s}_m)_{m\in\mathcal{M}}$ which amounts to select $m \in \mathcal{M}$ and can be described as follows: For $\|\cdot\|$ a given norm and σ a function to be chosen later ($\sigma^2(m)$ is the equivalent to pen(m)), we set for any m in $\mathcal{M}$,

$$A(m) := \sup_{m'\in\mathcal{M}} \left\{ \|\hat{s}_{m'} - \mathcal{K}_{m'}(\hat{s}_m)\| - \sigma(m') \right\}_+ .$$

Then we estimate s by using $\tilde{s} := \hat{s}_{\hat{m}}$, where $\hat{m}$ is selected as follows:

$$\hat{m} := \operatorname*{argmin}_{m\in\mathcal{M}} \left\{ A(m) + \sigma(m) \right\} .$$

This choice can be seen as a bias-variance trade-off, with $\sigma(m)$ an estimator of the standard deviation of $\hat{s}_m$ and $A(m)$ an estimator of the bias. This method was originally used for kernel estimators, and in this case $\mathcal{K}_m$ is the convolution operator with a kernel dilated of $h = 1/m$. This is what we will use in Section 2.4.6. Here, we adapt the procedure somewhat, taking for $\mathcal{K}_m$ the projection on $(S_m, \langle ., . \rangle_f)$. Of course (6) is satisfied, but not (7). Therefore, we modify this approach to overcome this problem. The idea is the following. Let us denote $S_{m\wedge m'} = S_m \cap S_{m'}$. Taking inspiration from the fact that $\mathcal{K}_m \circ \mathcal{K}_{m'}(s) = \mathcal{K}_{m\wedge m'}(s)$, let us set for any $(m, m') \in \mathcal{M}^2$,

$$\tilde{\mathcal{K}}_m(\hat{s}_{m'}) = \hat{s}_{m\wedge m'}.$$

This operator is only defined on the set of the estimators $\hat{s}_m$ but verifies (7). Now the previous reasoning can be reproduced and the above method can be applied by replacing $\mathcal{K}_{m'}$ by $\tilde{\mathcal{K}}_{m'}$ in $A(m)$.

More precisely, we denote

$$m \wedge m' = (m_1 \wedge m'_1, m_2 \wedge m'_2) = (\min(m_1, m'_1), \min(m_2, m'_2))$$

and we estimate s by $\tilde{s} = \hat{s}_{\hat{m}}$ where

$$\hat{m} = \hat{m}(x) := \operatorname*{argmin}_{m\in\mathcal{M}} \left\{ A(m) + \sigma(m) \right\}$$

and

$$A(m) := \sup_{m'\in\mathcal{M}} \left[\|\hat{s}_{m'} - \hat{s}_{m'\wedge m}\|_{x,2} - \sigma(m') \right]_+ .$$

2.4.4 Oracle inequality

We recall that $\mathcal{K}_m$ is the orthogonal projection on $(S_m, \langle , \rangle_f)$ where $\langle , \rangle_f$ is the dot product defined in (2).

Theorem 13 ([L15]). *We assume that* (**A**), (**B**) *are verified, and for all* $m \in \mathcal{M} \subset \mathcal{M}^{reg}$

$$k_n(r+1) \leqslant D_{m_1} \leqslant \frac{\widehat{f}_0 n}{(\log n)^3} \quad \text{and} \quad \log^2(n) \leqslant D_{m_2} \leqslant n.$$

For $\gamma > 0$, *we choose*

$$\sigma(m) = \hat{\chi}\sqrt{\frac{D_{m_1} D_{m_2}}{\widehat{f}_0 n}} \quad \text{with} \quad \hat{\chi}^2 = (1+\gamma)^2 \frac{4(r+1)^3}{|V_n(x)|} \frac{\widehat{\|f\|_\infty}}{\widehat{f}_0}.$$

Then, with probability larger than $1 - C_0 \exp\{-(\log n)^{5/4}\}$,

$$\|\tilde{s} - s\|_{x,2} \leqslant \inf_{m \in \mathcal{M}} \left(C_1 \sup_{t \in V_n(x)} \|\mathcal{K}_m(s) - s\|_{t,2} + \frac{5}{2}\hat{\chi}\sqrt{\frac{D_{m_1} D_{m_2}}{\widehat{f}_0 n}} \right)$$

where $C_1 = 1 + 2(r+1) f_0^{-1} \|f\|_\infty$ *and* C_0 *depends on* $|V_n(x)|, r, \gamma, \|s\|_\infty$ *and* f. *Moreover*

$$\mathbb{E}^{1/2}\|\tilde{s} - s\|_{x,2}^2 \leqslant \tilde{C}_1 \inf_{m \in \mathcal{M}} \left(\sup_{t \in V_n(x)} \|\mathcal{K}_m(s) - s\|_{t,2} + \sqrt{\frac{D_{m_1} D_{m_2}}{f_0 n}} \right) + \frac{\tilde{C}_2}{\sqrt{n}}$$

where $\tilde{C}_1$ *depends on* $|V_n(x)|, r, \gamma, \|f\|_\infty, f_0$ *and* $\tilde{C}_2$ *depends on* $|V_n(x)|, r, \gamma, \|s\|_\infty, f$.

The right hand side corresponds to the best trade-off between a bias term and a variance term. Here the constant $\hat{\chi}$ in the penalty depends on $\widehat{f}_0$ and $\|\hat{f}\|_\infty$ but however, in the case where $r = 0$ (histogram basis), it is possible to use the simpler penalty term $\hat{\chi} = (1+\gamma)2/\sqrt{|V_n(x)|}$ and the previous result still holds.

Note on the proof:
Let us fix $m \in \mathcal{M}$. By definition of $A(m)$ and $\hat{m}$

$$\begin{aligned} \|\hat{s}_{\hat{m}} - s\|_{x,2} &\leqslant \|\hat{s}_{\hat{m}} - \hat{s}_{m\wedge\hat{m}}\|_{x,2} + \|\hat{s}_{m\wedge\hat{m}} - \hat{s}_m\|_{x,2} + \|\hat{s}_m - s\|_{x,2} \\ &\leqslant A(m) + \sigma(\hat{m}) + A(\hat{m}) + \sigma(m) + \|\hat{s}_m - s\|_{x,2} \\ &\leqslant 2A(m) + 2\sigma(m) + \|\hat{s}_m - \mathcal{K}_m(s)\| + \|\mathcal{K}_m(s) - s\|. \end{aligned}$$

It remains essentially to upper bound $A(m)$. By splitting $\|\hat{s}_{m'} - \hat{s}_{m'\wedge m}\|_{x,2}$ into bias+variance, one can show that

$$A(m) \leqslant (r+1)\|f\|_\infty f_0^{-1} \sup_{t \in V_n(x)} \|s - \mathcal{K}_m(s)\|_{t,2} + \sup_{m' \in \mathcal{M}_n} \left\{ 2\|\hat{s}_{m'} - \mathcal{K}_{m'}(s)\| - \sigma(m') \right\}_+ .$$

Using a Talagrand inequality, we prove that the last term vanishes on a space of great probability.
■

2.4.5 Rate of convergence

Since we are interested in the smoothness at a given point, it is assumed here that s belongs to a Hölder space. This is as considering $B(\alpha, \infty, \infty, R)$. Let

$$H_2(\alpha, R) = \Big\{ s : \mathbb{R}^2 \to \mathbb{R} \text{ such that for all } x, y, x', y' \in \mathbb{R},$$
$$\left| \frac{\partial^{\lfloor \alpha_1 \rfloor} s}{\partial x^{\lfloor \alpha_1 \rfloor}}(x', y) - \frac{\partial^{\lfloor \alpha_1 \rfloor} s}{\partial x^{\lfloor \alpha_1 \rfloor}}(x, y) \right| \leqslant R|x' - x|^{\alpha_1 - \lfloor \alpha_1 \rfloor} \text{ and } \left| \frac{\partial^{\lfloor \alpha_2 \rfloor} s}{\partial y^{\lfloor \alpha_2 \rfloor}}(x, y') - \frac{\partial^{\lfloor \alpha_2 \rfloor} s}{\partial y^{\lfloor \alpha_2 \rfloor}}(x, y) \right| \leqslant R|y' - y|^{\alpha_2 - \lfloor \alpha_2 \rfloor} \Big\}$$

where $\lfloor \beta \rfloor = \max\{l \in \mathbb{N} : l < \beta\}$. Let us first state a lower bound result for the used norm.

Theorem 14 ([L15]). *There exists a positive constant C depending neither on R nor n such that, for n large enough,*

$$\inf_{\hat{s}_n} \sup_{(s,f)\in\tilde{H}(\alpha,R)} \left\{ (f(x))^{\frac{2\bar{\alpha}}{2\bar{\alpha}+2}} \mathbb{E}_s \|s-\hat{s}_n\|_{x,2}^2 \right\} \geqslant C R^{\frac{2}{\bar{\alpha}+1}} n^{-\frac{2\bar{\alpha}}{2\bar{\alpha}+2}}$$

where the infimum is taken on all estimators $\hat{s}_n$ of s based on observations $(X_1,Y_1),\dots,(X_n,Y_n)$ and $\tilde{H}(\alpha,R)$ is the set such that the conditional density s belongs to $H_2(\alpha,R)$ and the marginal density f is continuous.

Here we will show that our estimator achieves this rate.

Theorem 15 ([L15]). *Let $s\in H_2(\alpha,R)$. We assume that the models satisfy $r>\alpha_1$ and $r>\alpha_2$. Under assumptions (**A**), (**B**), the estimator $\tilde{s}$ defined is Section 2.4.3 enjoys*

$$\mathbb{E}\|s-\tilde{s}\|_{x,2}^2 \leqslant C R^{\frac{2}{\bar{\alpha}+1}} n^{\frac{-2\bar{\alpha}}{2\bar{\alpha}+2}}$$

where C depends on r,α_1,α_2,f_0 and $\|f\|_\infty$.

Thus our estimator is adaptive minimax. This procedure can be extended to higher dimensions, at least theoretically.

2.4.6 Kernel procedure

Here we present another approach to estimate the conditional density s. Our motivation is to study more finely the presence of f_0. Moreover kernel estimators are natural and computationally efficient.

Let us assume for a while that f is known and positive. We introduce a kernel K, namely a bounded integrable function K such that $\iint K(u,v)dudv=1$ and $\|K\|_2<\infty$. Then, given a regularization parameter, namely a bandwidth $h=(h_1,h_2)$ belonging to a set $\mathcal{H}$ to be specified later, we set

$$K_h(u,v)=\frac{1}{h_1h_2}K\left(\frac{u}{h_1},\frac{v}{h_2}\right).$$

It is simpler to consider K of the form $K(u,v)=K^{(1)}(u)K^{(2)}(v)$, and we will assume that the function $K^{(1)}$ is supported by $[-A,A]$. For all $h=(h_1,h_2)\in\mathcal{H}$, we set

$$\hat{s}_h(x,y):=\frac{1}{n}\sum_{i=1}^n \frac{1}{f(X_i)}K_h(x-X_i,y-Y_i). \tag{8}$$

For any given $h\in\mathcal{H}$, we provide a lower bound of the risk of $\hat{s}_h$ by using the following bias-variance decomposition

$$\mathbb{E}\left[\|\hat{s}_h-s\|_{x,2}^2\right]=\|K_h\star s-s\|_{x,2}^2+\int \mathrm{Var}(\hat{s}_h(x,y))dy,$$

where $K_h\star s(x,y):=\int K_h(x-u,y-v)s(u,v)dudv$.

Proposition 16 ([L15]). *We assume that s is bounded, and that f is positive and continuous in the neighborhood of x, and that $\max\mathcal{H}\to 0$ when $n\to+\infty$, then*

$$\mathbb{E}\left[\|\hat{s}_h-s\|_{x,2}^2\right]\geqslant\|K_h\star s-s\|_{x,2}^2+\frac{\|K\|_2^2}{f(x)nh_1h_2}\times(1+o(1))+O\left(\frac{1}{n}\right). \tag{9}$$

This lower bound can be viewed as a benchmark for our procedure. In particular, our challenge is to build a data-driven kernel procedure whose risk achieves the lower bound given in (9). This is the goal of the next paragraph where we modify $\hat{f}_h$ by estimating f when f is unknown.

We use the aforementioned estimator $\hat{f}$, but this time more fully. As already explained, it is assumed in this section that additional data $X_{n+1}, \dots, X_{2n}$ are available and used in the construction of $\hat{f}$ so that this estimator is independent of the data $(X_1, Y_1), \dots, (X_n, Y_n)$. Naturally, we replace $\hat{s}_h$ defined in (8) by

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

Then, we use the Goldenshluger-Lepski method, as previously, except that the regularization parameter is denoted h, instead of m to match with usual notation of the literature. Similarly, the set of bandwidths is denoted by $\mathcal{H}$, instead of $\mathcal{M}$. We estimate s by $\tilde{s} = \hat{s}_{\hat{h}}$ where

$$\hat{h} = \hat{h}(x) := \operatorname*{argmin}_{h \in \mathcal{H}} \left\{ A(h) + \sigma(h) \right\},$$

$$A(h) := \sup_{h' \in \mathcal{H}} \left\{ \left\| \hat{s}_{h'} - \hat{s}_{h,h'} \right\|_{x,2} - \sigma(h') \right\}_+,$$

and

$$\hat{s}_{h,h'}(x,y) = \frac{1}{n}\sum_{i=1}^{n} \left[\hat{f}(X_i) \right]^{-1} (K_h \star K_{h'})(x - X_i, y - Y_i) = (K_{h'} \star \hat{s}_h)(x,y).$$

Then we can prove an oracle inequality for this estimator.

Theorem 17 ([L15])**.** *We assume* (**A**), (**B**) *and the bandwidths* (h_1, h_2) *are such that* $h_1 = \frac{1}{k}, h_2 = \frac{1}{l}$, *with* k *and* l *integers and*

$$k_n \leqslant \frac{1}{h_1} \leqslant \frac{\widehat{f}_0 n}{(\log n)^3} \quad et \quad \log^2(n) \leqslant \frac{1}{h_2} \leqslant n.$$

For a given $\gamma > 0$, *we choose*

$$\sigma(h) = \frac{\chi}{\sqrt{\widehat{f}_0 n h_1 h_2}} \quad with \quad \chi = (1+\gamma)(1 + \|K\|_1)\|K\|_2,$$

We have with probability larger than $1 - C\exp\{-(\log n)^{5/4}\}$,

$$\|\tilde{s} - s\|_{x,2} \leqslant \inf_{h \in \mathcal{H}} \left\{ C_1 \sup_{t \in V_n(x)} \|K_h \star s - s\|_{t,2} + \frac{C_2}{\sqrt{\widehat{f}_0 n h_1 h_2}} \right\} + \frac{C_3}{f_0} \sup_{t \in V_n(x)} |\hat{f}(t) - f(t)|,$$

where $C_1 = 1 + 2\|K\|_1$, $C_2 = (1+\gamma)\|K\|_2(3 + 2\|K\|_1)$, C_3 *depends on* K, γ *et* $\|s\|_\infty$ *and* C *depends on* K, γ, f *et* $\|s\|_\infty$. *Moreover*

$$\mathbb{E}\left[\|\tilde{s} - s\|_{x,2}^2 \right]^{1/2} \leqslant \tilde{C}_1 \inf_{h \in \mathcal{H}} \left\{ \sup_{t \in V_n(x)} \|K_h \star s - s\|_{t,2} + \frac{1}{\sqrt{f_0 n h_1 h_2}} \right\} + \frac{\tilde{C}_2}{f_0} \mathbb{E}^{1/2} \left(\sup_{t \in V_n(x)} |\hat{f}(t) - f(t)|^2 \right) + \frac{\tilde{C}_3}{\sqrt{n}},$$

where $\tilde{C}_1$ *depends on* K, γ, $\tilde{C}_2$ *depends on* K, γ *et* $\|s\|_\infty$ *and* $\tilde{C}_3$ *depends on* K, γ, f, $\|s\|_\infty$. *Moreover, in the case where* f *is known* ($\hat{f} = f$, $\widehat{f}_0 = f_0$), C *and* $\tilde{C}_3$ *do not depend on* f.

This result almost matches up to the aforementioned lower bound, since the size of $V_n(x)$ goes to 0 when $n \to +\infty$. So we have optimality since the term f_0 (or at least $f(x)$) in the variance is unavoidable. This leads to the following result of rate, for $s \in H_2(\alpha, R)$ (assuming that the kernel K is of sufficiently large order)

$$\mathbb{E}\left[\|\tilde{s} - s\|_{x,2}^2\right] \leqslant C\left(R^{\frac{2}{\bar{\alpha}+1}}(nf_0)^{-\frac{2\bar{\alpha}}{2\bar{\alpha}+2}} + \frac{1}{f_0^2}\mathbb{E}\left(\sup_{t\in V_n(x)}|\hat{f}(t) - f(t)|^2\right)\right).$$

The term $\sup_{t\in V_n(x)}|\hat{f}(t) - f(t)|$ is obviously cumbersome, but it does not degrade the rate if f is smooth enough. (Note that methods using a quotient of an estimator of $f_{X,Y}$ and an estimator of f have an even slower rate, because $f_{X,Y}(x,y) = f(x)s(x,y)$ is less regular than s).

In the numerical study lead in [L15] on classic examples (so regular enough), it turned out that the kernel estimator was more efficient in terms of risk, although slower to implement. The calibration of γ in the penalty was also easier to implement for this kernel procedure. Perhaps it would be better, for the projection estimator, to replace $\sigma(m')$ by a penalty $\sigma(m, m')$ depending both on m and m' to be closer to the true variance.

2.5 Extensions

2.5.1 Extension to dependent data

The above results can be extended to the case of dependent variables. The observations framework is the following. Assume now that $\{(X_i, Y_i)\}_{i\in\mathbb{Z}}$ is a strictly stationary process and the variables X_i have a density f with respect to the Lebesgue measure. The goal is still to estimate the conditional density s of Y_i given X_i from the data $(X_1, Y_1), \ldots, (X_n, Y_n)$. The case of Markov chains is particularly important, especially in view of applications: if $(X_i)_{i\in\mathbb{Z}}$ is a homogeneous Markov chain of order 1, and if we denote $Y_i = X_{i+1}$ for all $i \in \mathbb{Z}$, then the conditional density s of Y_i given X_i is the transition density of the chain $(X_i)_{i\in\mathbb{Z}}$, and our aim is to estimate it from a trajectory $X_1, \ldots, X_{n+1}$.

To extend our results and manage dependence, we will assume that the process $\{Z_i\}_{i\in\mathbb{Z}} = \{(X_i, Y_i)\}_{i\in\mathbb{Z}}$ satisfies mixing conditions: we are particularly interested in β-mixing and ρ-mixing. So we will recall the definition of these concepts. For two sub σ-fields $\mathcal{A}$ and $\mathcal{B}$ of $\mathcal{F}$, the β-mixing coefficient, or absolute regularity, is defined by

$$\beta(\mathcal{A}, \mathcal{B}) = \mathbb{E}\left[\sup_{B\in\mathcal{B}}|\mathbb{P}(B|\mathcal{A}) - \mathbb{P}(B)|\right],$$

and the ρ-mixing coefficient, or maximal correlation, is defined by

$$\rho(\mathcal{A}, \mathcal{B}) = \sup_{X,Y}\frac{|\mathrm{Cov}(X, Y)|}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}},$$

where the supremum is taken over all real random variables X and Y respectively $\mathcal{A}$- et $\mathcal{B}$-measurable and square integrable. These quantities quantify the dependence, they are equal to 0 if $\mathcal{A}$ and $\mathcal{B}$ are independent and are all the larger than the dependence between $\mathcal{A}$ and $\mathcal{B}$ is strong. To precisely define our conditions, we denote $Z_i = (X_i, Y_i)$ and for all integer j,

$$\begin{aligned}\beta_j^{\boldsymbol{Z}} &= \beta\left(\sigma(Z_i, i \leqslant 0), \sigma(Z_i, i \geqslant j)\right),\\ \rho_j^{\boldsymbol{Z}} &= \rho\left(\sigma(Z_i, i \leqslant 0), \sigma(Z_i, i \geqslant j)\right).\end{aligned}$$

Thus we measure the dependence between the first variables and the next variables spaced out j ranks. The process is called β-mixing if $\lim_{j\to+\infty}\beta_j^{\boldsymbol{Z}}=0$ (idem for ρ).

We are now in position to state our dependence assumptions. In each case, we also define real numbers ϑ and δ that will appear in the next result.

Assumption $(\boldsymbol{D\beta})$ The process $(Z_i)_{i\in\mathbb{Z}}$ is geometrically β-mixing, with $a\geqslant 0$ and $b>0$ such that, for all positive integer j, $\beta_j^{\boldsymbol{Z}}\leqslant a\exp(-bj)$. Then we denote $\vartheta=1$ and $\delta=1$.

Assumption $(\boldsymbol{D\beta\rho})$ Assumption $(\boldsymbol{D\beta})$ is satisfied et, in addition, the series $S_\rho:=\sum_{j\in\mathbb{N}}\rho_{2^j}^{\boldsymbol{Z}}$ converges. Then we denote $\vartheta=250\prod_{j=0}^{\infty}\left(1+\rho_{\lfloor 2^{j/3}\rfloor+1}^{\boldsymbol{Z}}\right)$ and $\delta=0$.

Assumption $(\boldsymbol{D\beta 2\text{-}\rho})$ Assumption $(\boldsymbol{D\beta})$ is satisfied et, in addition, the series $S_{2\text{-}\rho}:=\sum_{j\geqslant 1}\rho(\sigma(Z_0),\sigma(Z_j))$ converges. Then we denote $\vartheta=(1+2S_{2\text{-}\rho})$ and $\delta=0$.

Assumption $(\boldsymbol{D\beta_{cond}})$ Assumption $(\boldsymbol{D\beta})$ is satisfied et, in addition, for all $j\geqslant 2$, Z_j is independent of Z_1 conditionally to X_j. Then we denote $\vartheta=1$ and $\delta=0$.

Compared to other conventional mixing notions, we know that the β and ρ-mixing are implied by the ϕ-mixing (uniform mixing) and entail the α-mixing, hence our assumptions are quite mild. If we try to compare our four assumptions together, we can say that the first $(\boldsymbol{D\beta})$ is the weakest, that $(\boldsymbol{D\beta 2\text{-}\rho})$ implies $(\boldsymbol{D\beta\rho})$ in the case of Markov chains and the last $(\boldsymbol{D\beta_{cond}})$ does not involve ρ-mixing condition. This last assumption is verified when estimating the transition density of a Markov chain (X_i) and $Y_i=X_{i+1}$. Many processes verify these assumptions, especially among Markov chains, ARMA or ARCH models: see [L11].

In the literature, in general, the assumptions required in the old papers are rather strong (ρ-mixing, Doeblin condition). However, several authors are able to assume only a a Harris recurrence assumption (Athreya and Atuncar, 1998) or α-mixing (Masry (1989), Cai (1991), Chen et al. (2001)). Then one can wonder if our assumptions could be weakened and if we can prove a result under α-mixing assumption. This is indeed the case as long as we do not try to make adaptation and to prove oracle inequalities. Akakpo (2009) proved a version of Proposition 5, which bounds the risk $\mathbb{E}\left[\|s-\hat{s}_m\|_n^2\right]$ for a given model m, under assumption of geometrical α-mixing. Actually a sufficient condition to ensure that $\mathbb{E}_s\left[\|\hat{s}_m-s_m\|_n^2\right]$ is of the same order as in the independent case is that for some constant C and for any $t\in S_m$

$$\mathrm{Var}\left(\sum_{i=1}^{n}t(Z_i)\right)\leqslant Cn\mathrm{Var}\left(t(Z_1)\right). \tag{10}$$

Still, Assumptions $(\boldsymbol{D\beta\rho})$ and $(\boldsymbol{D\beta 2\text{-}\rho})$ are almost optimal for obtaining such an inequality, in the following sense: a Harris ergodic and reversible Markov chain $(Z_i)_{i\in\mathbb{Z}}$ satisfies (10) if and only if it is ρ-mixing.

We now state a result of oracle inequality. We are interested in integrated risk and we consider the same conditions (same model, same notation) as in Section 2.3.4. The aim is to prove again Theorem 10 in the dependent framework. We will see that a logarithmic factor appears in the penalty (and therefore in the rate of convergence) under the only condition of β-mixing, but this term vanishes under the stronger assumptions $(\boldsymbol{D\beta\rho})$, $(\boldsymbol{D\beta 2\text{-}\rho})$ or $(\boldsymbol{D\beta_{cond}})$. This is what means the factor $\log^{\delta}(n)$ with $\delta\in\{0,1\}$.

Theorem 18 ([L11]). *We assume that $(Z_i)_{i\in\mathbb{Z}}$ satisfies assumption $(\boldsymbol{D\beta})$ and s, f satisfy Assumption $(\boldsymbol{A})$. We assume that the maximum model $S_{m^\star}$ is of Cartesian type and is made of a regular partition in cubes such that $D_{m^\star} \lesssim \sqrt{n}/\log n$. We also assume that there exists a collection $\{L_m\}_{m\in\mathcal{M}}$ of real numbers such that $\sum_{m\in\mathcal{M}} \exp(-L_m D_m) \leqslant 1$. For a large enough numerical constant $\kappa > 0$, we choose*

$$\mathrm{pen}(m) = \kappa \left(\vartheta (b^{-1}\log(n))^{\delta} \|s\|_\infty + \frac{(2r+1)^2}{b^2 f_0} \right) \frac{L_m^2 D_m}{n}$$

(where b, δ and ϑ are defined in the dependence assumptions). Then

$$\mathbb{E}\left[\|\tilde{s} - s\|_n^2\right] \leqslant C \left(\max_{m\in\mathcal{M}} L_m^2 \right) \min_{m\in\mathcal{M}} \left\{ d_f^2(s, S_m) + \frac{D_m}{n} \right\}.$$

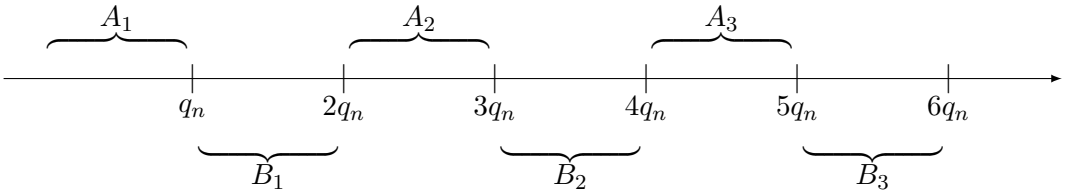
where C depends on κ, ϑ, δ, a, b, r, $\|s\|_\infty$, $f_0, \|f\|_\infty$.

Thus, under the weakest assumption $(\boldsymbol{D\beta})$ there is a term $\log(n)$ in the penalty. Under the stronger assumptions $(\boldsymbol{D\beta\rho})$, $(\boldsymbol{D\beta 2\text{-}\rho})$, we can avoid this but the price to pay is the presence of the term ϑ, which is rather troublesome. For practical purposes, it is necessary to include this term in the tuning parameter κ.

The best case is when Assumption $(\boldsymbol{D\beta_{cond}})$ is verified, which is the case when we estimate the transition density of a Markov chain. Indeed, in this case, $\delta = 0$ and $\vartheta = 1$, so that the penalty is almost as simple as in the independent case. In this case, it is even possible to only assume an arithmetical β-mixing, if we slightly strengthen the condition on the maximum model; and we can even avoid the term b^2 in the penalty if the mixing is strong enough. In the simple case of a transition with homogeneous smoothness and using regular models, a penalty $\kappa\|s\|_\infty D_{m_1} D_{m_2}/n$ is suitable (see [L3]).

Note on the proof:
As in the independent case, the result is deduced from the control of the deviation of the variable $\sup_t \nu(t)$. The goal here is to use the assumption of β-mixing to be reduced to the independent case. Specifically, we set $q_n = \lceil 3b^{-1}\log(n) \rceil$ (where b is defined in Assumption $(\boldsymbol{D\beta})$) and we perform the Euclidean division of n by q_n: $n = d_n q_n + r_n$. For the sake of simplicity, we can assume that $r_n = 0$ et $d_n = 2p_n > 0$ (the other cases being similar). We will group the data by spaced blocks. For $l = 0, \ldots, p_n - 1$, set

$$A_l = \{Z_i\}_{2lq_n+1 \leqslant i \leqslant (2l+1)q_n} \quad \text{and} \quad B_l = \{Z_i\}_{(2l+1)q_n+1 \leqslant i \leqslant (2l+2)q_n}.$$



As recalled for instance in Viennet (1997), we can build, for $l = 0, \ldots, p_n - 1$,

$$A_l^\bullet = \{Z_i^\bullet\}_{2lq_n+1 \leqslant i \leqslant (2l+1)q_n} \quad \text{and} \quad B_l^\bullet = \{Z_i^\bullet\}_{(2l+1)q_n+1 \leqslant i \leqslant (2l+2)q_n}$$

such that, for all $l = 0, \ldots, p_n - 1$,

- A_l, $A_l^\bullet$, B_l et $B_l^\bullet$ have the same law;
- $\mathbb{P}_s(A_l \neq A_l^\bullet) \leqslant \beta_{q_n}^{\boldsymbol{Z}}$ and $\mathbb{P}_s(B_l \neq B_l^\bullet) \leqslant \beta_{q_n}^{\boldsymbol{Z}}$;
- $(A_l^\bullet)_{0\leqslant l\leqslant p_n-1}$ are independent, as well as $(B_l^\bullet)_{0\leqslant l\leqslant p_n-1}$.

Thus, it is sufficient to consider $\Omega_\bullet = \bigcap_{i=1}^n \{Z_i^\bullet = Z_i\}$ to be reduced to blocks of independent variables. Moreover this set has a great probability since

$$\mathbb{P}(\Omega_\bullet^c) \leqslant 2p_n\beta_{q_n}^{\boldsymbol{Z}} \leqslant \frac{n}{q_n} a e^{-bq_n} \leqslant \frac{ab}{3}\frac{n^{-2}}{\log n}.$$

■

Finally, using the dyadic partitions described in Section 2.3.4, we can state the following result.

Corollary 19. *The notation is that of Theorem 18, Assumption* (**A**) *is supposed to be fulfilled. We assume that* $\mathcal{M} \subset \mathcal{M}^{irreg}$ *is defined as previously with the condition*

$$D_{m^\star} \lesssim \sqrt{n}/\log(n).$$

Assume that $s \in B(\alpha, p, \infty, R)$ *with* $p = 1$ *or* $p = 2$*, and* $\alpha \in (0, r+1)^2$ *such that* $q(\alpha, p) > 1$*. If* $\log^\delta(n)/n \leqslant R^2 \leqslant n^{q(\alpha,p)-1}\log(n)^{\delta-2q(\alpha,p)}$*, then there exists some positive real* $C(\alpha, r, p)$ *that only depends on* α, r, p *such that*

$$\sup_{s\in B(\alpha,p,R)} \mathbb{E}_s\left[\|s-\tilde{s}\|_n^2\right] \leqslant C(\alpha,r,p)\|f\|_\infty R^{\frac{2}{\bar{\alpha}+1}} \left(\frac{n}{\log^\delta(n)}\right)^{\frac{-2\bar{\alpha}}{2\bar{\alpha}+2}}.$$

Thus we recover the same rate of estimation as with independent data up to a logarithmic factor that disappears under Assumptions (***Dβρ***), (***Dβ2-ρ***) or (***Dβ$_{cond}$***).

2.5.2 Extension to censored data

In this section, we will see that the above results extend to the case of right censoring. This is a model frequently used in reliability or survival analysis where one studies the life (or time to failure) of individuals. There is right censorship when some individuals in the study are not observed until the end (death, remission, healing). In this case, we observe only a lower bound of the life time, the so-called survival time. Finally, the observations consist of the minimum between life time and censoring time, and the knowledge that it has been censored or not. For more details on censored models, see Andersen et al. (1993, chap. 3).

We consider the following censoring framework: the observations are $(X_i, T_i, \delta_i)_{1\leqslant i\leqslant n}$ where

$$T_i = \min(Y_i, C_i), \qquad \delta_i = \mathbb{1}_{\{Y_i\leqslant C_i\}}.$$

The explanatory variable X is unchanged, and the response Y is censored. The variable C is the censoring variable, and is assumed to be positive. If $Y > C$, we say that the variable Y is censored, we only observe $T = C$ and the censoring indicator δ is equal to 0. Otherwise, Y is not censored and we directly observe $T = Y$, the indicator is then $\delta = 1$. We will work under the (little strong) assumption that C is independent of (X, Y), which means that the censorship happens for external reasons unrelated to X and Y. A second assumption, which is classical in the censoring framework, is related to the cumulative distribution function G of C and its

survival function $\bar{G} = 1 - G$.

Assumption (C) The censoring variable C is independent of (X, Y), and there exists a positive constant c_G such that for all $y \in [0,1]$, $\bar{G}(y) = 1 - G(y) \geqslant c_G$.

Now we adapt our estimation procedure to this new observations context. To do this, we modify the contrast. We use a standard transformation of the data and introduce an empirical version of the weights

$$w_i = \frac{\delta_i}{\bar{G}(T_i)}$$

where $\bar{G}$ is the survival function associated with the censoring variables. Indeed

$$\begin{aligned}\mathbb{E}\left(w_i t(X_i,T_i)|X_i,Y_i\right) &= \mathbb{E}\left(\frac{\delta_i}{\bar{G}(T_i)}t(X_i,T_i)|X_i,Y_i\right) = \mathbb{E}\left(\frac{\mathbb{1}_{\{T_i=Y_i\}}}{\bar{G}(T_i)}t(X_i,T_i)|X_i,Y_i\right)\\ &= \frac{t(X_i,Y_i)}{\bar{G}(Y_i)}\mathbb{E}(\mathbb{1}_{\{Y_i\leqslant C_i\}}|X_i,Y_i) = \frac{t(X_i,Y_i)}{\bar{G}(Y_i)}\bar{G}(Y_i) = t(X_i,Y_i).\end{aligned}$$

We would like to use these weights in our procedure by replacing $t(X_i, Y_i)$ with $w_i t(X_i, T_i)$ in the contrast, but unfortunately $\bar{G}$ is unknown. Then we have to estimate it. We denote by $\hat{\bar{G}}$ the Kaplan-Meier estimator of the c.d.f G, modified in the way suggested by Lo et al. (1989), and defined by

$$\hat{\bar{G}}(y) = \prod_{T_{(i)}\leqslant y}\left(\frac{n-i+1}{n-i+2}\right)^{1-\delta_{(i)}}.$$

The new contrast is then

$$\gamma_n(t) = \frac{1}{n}\sum_{i=1}^{n}\left(\int_{\mathbb{R}} t^2(X_i,y)dy - 2\hat{w}_i t(X_i,T_i)\right),\ \hat{w}_i = \frac{\delta_i}{\hat{\bar{G}}(T_i)}.$$

It is the same contrast, replacing $t(X_i, T_i)$ with $\hat{w}_i t(X_i, T_i)$. Note that this contrast coincides with the previous one if no censoring occurs ($T_i = Y_i$), by defining the weights $\hat{w}_i = 1$. The estimator is the same *mutatis mutandis* and here we only study the case of global risk and homogeneous smoothness.

Theorem 20 ([L6])**.** *We assume that* (**A**), (**C**) *are verified. We assume that* $\mathcal{M} \subset \mathcal{M}^{reg}$ *and*

$$\forall m \in \mathcal{M} \qquad D_{m_1} \leqslant \sqrt{n}/\log(n) \quad \text{and} \quad D_{m_1}D_{m_2} \leqslant \sqrt{n}.$$

We consider two cases:

- *the models are anisotropic and*

$$\mathrm{pen}(m) = (1+\gamma)^2\frac{\|s\|_\infty}{c_G}\frac{D_{m_1}D_{m_2}}{n},$$

- *or the models are isotropic:* $m_1 = m_2$, $H_{m_1} = E_{m_1}$ *(i.e.* $S_m = E_{m_1} \otimes E_{m_1}$*) and*

$$\mathrm{pen}(m) = (1+\gamma)^2\frac{(2r+1)^2}{f_0}\mathbb{E}\left(\frac{\delta_1}{\bar{G}^2(T_1)}\right)\frac{D_{m_1}^2}{n}$$

Then

$$\mathbb{E}\|\tilde{s}-s\|_n^2 \leqslant C_3 \inf_{m\in\mathcal{M}} \left(d_f^2(s,S_m)+\mathrm{pen}(m)\right)+\frac{C_4}{n}$$

where C_3 depends on γ, and C_4 depends on $c_G, \|s\|_\infty, \|f\|_\infty, f_0, r, \gamma$.

The first case is the natural extension of Theorem 7. However, the term c_G appears in the penalty. This term is not easily estimable. That is why we present the second case. Again, there are additional quantities in the penalty: f_0 and $\mathbb{E}(\delta_1/\bar{G}^2(T_1))$ but these quantities are estimable. The estimation of f_0 has already been addressed. The term $\mathbb{E}(\delta_1/\bar{G}^2(T_1))$ can obviously be estimated by its empirical counterpart. The price to pay for this more realistic penalty is the loss of the anisotropic aspect of the procedure: we obtain the rate $n^{-\min(\alpha)/(2\min(\alpha)+2)}$ where $\min(\alpha)$ is the worst directional smoothness of s.

2.5.3 Extension to the conditional cumulative distribution function

If we are interested in the conditional distribution, it may be preferred to directly estimate the conditional cumulative distribution function $F(x,y)=\mathbb{P}(Y\leqslant y|X=x)$ rather than the density. This is particularly the case in reliability or survival analysis, among others, because it appears in the hazard rate. It is also used to compute the conditional quantiles. This estimation problem was first studied by Stute (1986) which shows the consistency of a conventional Nadaraya-Watson estimator. The method is refined by Hall et al. (1999) who also introduce a ratio estimator with well chosen weights and a bootstrap selection method, numerically studied. Hall and Yao (2005) proposed to estimate the distribution of Y given X by the one of Y given ${}^t\theta X$ with an optimized selection of θ which allows to handle the case of rather large covariate. The case of functional covariates is treated in Ferraty et al. (2006) and Chagny and Roche (2014). Finally, let us mention the work of Plancade (2013) that deals with the case of current status data.

We will see that our previous results can be adapted to this estimation issue. But this time, the adaptive aspect of the procedure is required only in the first direction. It must be well understood that there is a deep asymmetry between the two directions. In the direction x, one has to face a nonparametric regression problem. In the direction y, it is about estimation of a distribution function, which is much easier and is done with the parametric rate $1/n$.

The procedure is as above, but we change the contrast. This time, we set

$$\Gamma_n(t)=\frac{1}{n}\sum_{i=1}^n\int (t^2(X_i,y)-2t(X_i,y)\mathbb{1}_{\{Y_i\leqslant y\}})dy.$$

The expectation can be easily computed:

$$\begin{aligned}\mathbb{E}(\Gamma_n(t)) &= \iint t^2(x,y)f(x)dxdy-2\iiint t(x,y)1_{\{z\leqslant y\}}s(x,z)f(x)dxdydz\\ &= \iint (t^2(x,y)-2t(x,y)F(x,y))f(x)dxdy=\|t-F\|_f^2-\|F\|_f^2.\end{aligned}$$

This quantity being minimum when $t=F$, it is natural to minimize $\Gamma_n(t)$ to estimate F. To better understand this contrast, we can observe its value when applied to a function t which only depends on y, and not on x (no covariate). We obtain

$$\Gamma_n(t)=\frac{1}{n}\sum_{i=1}^n\int (t^2(y)-2t(y)\mathbb{1}_{\{Y_i\leqslant y\}})dy=\int t^2(y)dy-2\langle t,\hat{F}_n\rangle=\|t-\hat{F}_n\|_2^2-\|\hat{F}_n\|_2^2.$$

where $\hat{F}_n$ is the empirical cumulative distribution function of the sample $(Y_1, \dots, Y_n)$. The minimizer of this contrast on a univariate functions space H is then the projection of $\hat{F}_n$ on H, that we denote $P_H(\hat{F}_n)$. We can briefly study the performance of this estimator. Its expectation is $P_H(F)$. There is therefore a bias variance decomposition:

$$\mathbb{E}\|F - P_H(\hat{F}_n)\|_2^2 = \|F - P_H(F)\|_2^2 + \mathbb{E}\|P_H(F) - P_H(\hat{F}_n)\|_2^2.$$

The variance term is bounded by

$$\mathbb{E}\|P_H(F) - P_H(\hat{F}_n)\|_2^2 \leqslant \mathbb{E}\|F - \hat{F}_n\|_2^2 = \int_0^1 \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{Y_i \leqslant y}\right) dy \leqslant \frac{1}{n}$$

and thus does not depend on the dimension of H. The bias term is all the smaller as the approximation space H is large. We just have to choose H as large as possible to reduce the bias, and we preserve the (parametric) rate of the empirical c.d.f.. We have just made a smoothing of $\hat{F}_n$ without degrading its rate. In the presence of a covariate X, the approximation space is of the form $S = E \otimes H$ where E and H have different roles. As explained, it is sufficient to choose H large enough, while the model E must be selected from a collection (E_{m_1}) in order to balance the bias and variance in x.

For the sake of simplicity, we assume that the regularity of F is homogeneous. Therefore we take the simple models described in section 2.3.3. The estimator of the conditional distribution function is then

$$\hat{F}_{m_1} = \underset{t \in E_{m_1} \otimes H_{m_2^\star}}{\operatorname{argmin}} \Gamma_n(t),$$

Next, to perform model selection, we set

$$\hat{m} = \underset{m_1 \in \mathcal{M}}{\operatorname{argmin}} \left\{\Gamma_n(\hat{F}_{m_1}) + \mathrm{pen}(m_1)\right\}$$

and we denote $\tilde{F} = \hat{F}_{\hat{m}}$. Two transformations are added in order to obtain a c.d.f. type estimator:

- firstly, in order to provide a non-decreasing estimate with respect to the y variable, we apply the rearrangement proposed in Chernozhukov et al. (2009)

$$\tilde{F}^*(X_i, y) = \inf\left\{z \in \mathbb{R}, \int \mathbb{1}_{\{\tilde{F}(X_i,u) \leqslant z\}} du \geqslant y\right\}.$$

- secondly, to obtain a function taking its value between 0 and 1, a truncation is sufficient:

$$\breve{F}(x,y) = \begin{cases} \tilde{F}^*(x,y) & \text{if} \quad 0 \leqslant \tilde{F}^*(x,y) \leqslant 1, \\ 0 & \text{if} \quad \tilde{F}^*(x,y) < 0, \\ 1 & \text{if} \quad \tilde{F}^*(x,y) > 1. \end{cases}$$

Theorem 21 ([L8], improved version)**.** *We assume that assumption* (**A**) *is verified (only the part on f, no assumption on s), $\mathcal{M} \subset \mathcal{M}^{reg}$, and*

$$\forall m_1 \in \mathcal{M} \qquad D_{m_1} \lesssim n/\log(n).$$

For a given $\gamma > 0$, we choose

$$\mathrm{pen}(m_1) = (1+\gamma)^2 \frac{D_{m_1}}{n}$$

Then, with probability larger than $1 - C_0 \exp\{-(\log n)^{5/4}\}$,

$$\|\breve{F} - F\|_n^2 \leqslant \inf_{m_1 \in \mathcal{M}} \left(C_1 d_f^2(F, E_{m_1} \otimes H) + C_2 \mathrm{pen}(m_1) \right)$$

where $C_1 > (1 + 2\gamma^{-1})^2$, $C_2 > 2(1 + 2\gamma^{-1})$ *and* C_0 *depends on* $\|f\|_\infty, f_0, r$ *et* γ. *Moreover*

$$\mathbb{E}\|\breve{F} - F\|_n^2 \leqslant C_3 \inf_{m \in \mathcal{M}} \left(d_f^2(F, E_{m_1} \otimes H) + \frac{D_{m_1}}{n} \right) + \frac{C_4}{n}$$

where C_3 *depends on* γ, *and* C_4 *depends on* $\|f\|_\infty, f_0, r$ *et* γ.

We can deduce from this theorem the rate of convergence of the risk.

Corollary 22. *We assume the same assumptions as in Theorem 21 with* $\mathrm{pen}(m_1) = \kappa D_{m_1}/n$, *and moreover we suppose that* F *belongs to the anisotropic Besov ball* $B(\alpha, 2, \infty, R)$ *with smoothness* $\alpha = (\alpha_1, \alpha_2)$ *such that* $\alpha > 0$ *and* $\alpha_2 \geqslant 1$. *We assume that the models are such that the maximal degree* r *of the polynomials is larger than* $\alpha_i - 1$. *Then, if* $\dim(H) \geqslant \sqrt{n}$,

$$\mathbb{E}\|F - \breve{F}\|_n^2 \leqslant C R^{\frac{2}{2\alpha_1+1}} n^{-\frac{2\alpha_1}{2\alpha_1+1}}.$$

The assumption $\alpha_2 \geqslant 1$ can in fact be weakened to $\alpha_2 \geqslant \alpha_{\min}$, if we merely use a space H of dimension $\dim(H) \geqslant n^{1/(2\alpha_{\min})}$. The theorem is stated here with this assumption because we have always assumed the existence of the conditional density s, which corresponds to the differentiability of F in the second direction.

Thus we obtain the rate $n^{-\frac{2\alpha_1}{2\alpha_1+1}}$ which is the usual rate of convergence for estimating a univariate function of smoothness α_1. Actually this is the optimal rate for our estimation problem of the conditional distribution function. Indeed, we have the following lower bound result.

Theorem 23 ([L8])**.** *We assume that* $\alpha_2 > 1$. *Then for all bounded* f, *there exists a constant* $C > 0$ *such that, for* n *large enough,*

$$\inf_{\hat{F}_n} \sup_{F \in B(\alpha,2,\infty,R)} \mathbb{E}_{F,f}\|\hat{F}_n - F\|_2^2 \geqslant C R^{\frac{2}{2\alpha_1+1}} n^{-\frac{2\alpha_1}{2\alpha_1+1}}$$

where the infimum is taken over all estimators $\hat{F}_n$ *of* F *based on data* $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Note that in this case the estimation of the conditional distribution function, we can also work with censored data and get the same convergence rate by an *ad hoc* modification of the contrast and the penalty.

The simulations given in [L8] show that we obtain, on various examples, an oracle constant $C_{\mathrm{oracle}} = \mathbb{E}\|F - \breve{F}\|^2 / \inf_{m_1 \in \mathcal{M}_n} \mathbb{E}\|F - \hat{F}_{m_1}\|^2$ of order 1 or 2. Moreover, we have applied our conditional distribution estimator to experimental data about the strength of concrete. The data consisted in couples (x_i, y_i) for $1 \leqslant i \leqslant 635$ where x_i is the water-to-cement ratio ($\mathrm{kg/m^3}$) and y_i is the concrete compressive strength (MPa).

2.6 Penalty calibration

2.6.1 Calibration for Birgé-Massart model selection

In the previous sections, we have used model selection methodology with penalty of the form $\kappa L_m D_m / n$. In a first time the constant κ in the penalty term can be calibrated using simulations.

But it is obviously desirable to find a data-driven procedure. In the case of Birgé-Massart model selection, this procedure exists and is called slope heuristic: it is precisely described in Baudry et al. (2012). Can this method be applied to conditional density case? We can answer this question by using the results of Saumard (2012). Let K be the operator defined by

$$K(t) : (x, y) \mapsto \int t^2(x, u)du - 2t(x, y)$$

and $\gamma(t) = \mathbb{E}[K(t)(X, Y)]$. Thus $\gamma_n(t) = n^{-1} \sum_{i=1}^n K(t)(X_i, Y_i)$ is the empirical version of $\gamma(t)$. We can prove that this contrast is regular in the sense of Saumard. That is to say that it is C^3 in the sense of Fréchet derivative and associated to an Hilbertian loss function. This entails that, with probability larger than $1 - Cn^{-2}$,

$$\gamma_n(s_m) - \gamma_n(\hat{s}_m) \approx \frac{1}{4n} \sum_{j=1}^{D_{m_1}} \sum_{k=1}^{D_{m_2}} \mathrm{Var}(K'(s_m)(\varphi_j \otimes \varphi_k)) \approx \gamma(\hat{s}_m) - \gamma(s_m).$$

This result allows us to validate the practical use of slope heuristic. Indeed, if we denote the best theoretical model

$$m^* = \underset{m \in \mathcal{M}}{\mathrm{argmin}} \, \|s - \hat{s}_m\|_f^2 = \underset{m \in \mathcal{M}}{\mathrm{argmin}} \, \{\gamma_n(\hat{s}_m) + \mathrm{pen}^*(m)\},$$

easy computations lead to an optimal penalty $\mathrm{pen}^*(m) = (\gamma_n(s_m) - \gamma_n(\hat{s}_m)) + (\gamma(\hat{s}_m) - \gamma(s_m)) + \epsilon_n$, where ϵ_n is a negligible term. Then the previous result gives an optimal penalty equal to $2(\gamma_n(\hat{s}_m) - \gamma_n(s_m))$. Hence we can write

$$\kappa_{\mathrm{opt}} L_m D_m / n = 2(\gamma_n(\hat{s}_m) - \gamma_n(s_m)).$$

Thus the procedure is as follows: detect a linear relationship between $\gamma_n(\hat{s}_m)$ and $L_m D_m / n$, record the slope, and take κ_{opt} equal to two times this slope. This methodology has been implemented in other frameworks by Baudry et al. (2012) in the toolbox CAPUSHE (Matlab and R).

2.6.2 Calibration for Goldenshluger-Lespki method

In this section, I would like to address more precisely the issue of penalty calibration for the Goldenshluger-Lepski methodology. We focus on the calibration of the penalty term V or $\sigma = \sqrt{V}$. It is known that the method achieves good results for V large enough. But what is the minimal (and the optimal) value for V to keep this good behavior? In this section we deal with the case of simple density estimation (instead of conditional density). We consider this issue from a theoretical point of view but actually it is decisive for a practical implementation of the method. Our contribution is to evidence an explosion phenomenon: if the penalty term V is chosen smaller than some critical V_0, the risk is proven to dramatically increase, though for $V > V_0$ this risk is quasi-optimal. Proofs are extensively based on concentration inequalities. In particular, left tail concentration inequalities are used to prove the explosion result. We also implement numerical simulations which corroborate this behavior.

Kernel density estimation framework In this section, we consider the simpler framework of independent and identically distributed real variables $X_1, \ldots, X_n$ with unknown density f. For h a bandwidth we can define the classical kernel estimator

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

where K is a kernel and $K_h = K(./h)/h$. We assume here that the function to estimate is univariate and we study the Goldenshluger-Lepski methodology without oversmoothing. That is to say that we do not use auxiliary estimators. Indeed, this is not the heart of the method, and only induces slight changes in the bias term in our context. From $\{\hat{f}_h, h \in \mathcal{H}\}$ the collection of estimators, the procedure is the following. The bias is estimated by

$$A(h) = \sup_{h' \leqslant h} \left[\|\hat{f}_{h'} - \hat{f}_h\|_2^2 - V(h') \right]_+ \qquad \text{with } V(h') = a\frac{\|K_{h'}\|_2^2}{n} \tag{11}$$

and the selected bandwidth is

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

We introduce the following notation:

$$f_h := \mathbb{E}(\hat{f}_h), \quad h_{\min} := \min \mathcal{H}, \quad h_{\max} := \max \mathcal{H}$$
$$D(h) := \max(\sup_{h' \leqslant h} \|f_{h'} - f_h\|_2, \|f - f_h\|_2) \leqslant 2 \sup_{h' \leqslant h} \|f_{h'} - f\|_2.$$

We assume that the kernel verifies assumption

$$\int |K| = 1, \|K\|_2 < \infty \text{ and } \forall\ 0 \leqslant x \leqslant 1 \qquad \frac{\langle K, K(x.)\rangle}{\|K\|_2^2} \geqslant 1.$$

This is verified for classical kernels (Gaussian kernel, rectangular kernel, Epanechnikov kernel, biweight kernel; see [L17]). This entails that for all $h' \leqslant h$, $\|K_{h'} - K_h\|_2^2 \leqslant \|K_{h'}\|_2^2 - \|K_h\|_2^2$ which is a key property for our results.

Let us now recall what can be obtained if a is well chosen. Assume that f is bounded and $h_{\max}^{-1} \leqslant \sqrt{n}$. Then, if $a > 1$, there exists a positive $C = C(K, f)$ such that

$$\mathbb{E}\|\hat{f}_{\hat{h}} - f\|_2^2 \leqslant 2\left(\frac{3a-1}{a-1}\right)^2 \inf_{h \in \mathcal{H}} \left\{ D^2(h) + a\frac{\|K_h\|_2^2}{n} \right\} + C\frac{|\mathcal{H}|^2}{h_{\min}} e^{-\frac{(1-a^{-1})^2}{Ch_{\max}}}.$$

For $\mathcal{H} = \{e^{-k}, \lceil 2\log\log n \rceil \leqslant k \leqslant \lfloor \log n \rfloor\}$, the remaining term is bounded by $e^{-(1-a^{-1})^2(\log n)^2/C'}$ (see [L17]). We recognize in the right members the classical bias variance trade-off. This oracle inequality shows that the Goldenshluger-Lepski methodology works when $a > 1$. Here a is the constant in the penalty that we need to calibrate.

Minimal penalty Now we are interested in finding a minimal penalty $V(h)$, beyond which the procedure fails. Indeed, if a and then $V(h)$ is too small, the minimization of the criterion amounts to minimize the bias, and then to choose the smallest possible bandwidth. This leads to the worst estimator and the risk explodes.

In the following result $h_{\min}$ denotes the smallest bandwidth in $\mathcal{H}$ and is of order $1/n$.

Theorem 24 ([L17])**.** *Assume that f is bounded. Choose $\mathcal{H} = \{e^{-k}, \lceil 2\log\log n \rceil \leqslant k \leqslant \lfloor \log n \rfloor\}$ as a set of bandwidths. Consider for K the Gaussian kernel, the rectangular kernel, the Epanechnikov kernel or the biweight kernel. If $a < 1$ where a is defined in* (11)*, then, for n large enough (depending on f and K), the selected bandwidth $\hat{h}$ satisfies*

$$\exists C > 0 \qquad \mathbb{P}(\hat{h} \geqslant 3h_{\min}) \leqslant C(\log n)^2 \exp(-(\log n)^2/C)$$

i.e. $\hat{h} < 3h_{\min}$ with high probability. Moreover

$$\liminf_{n \to \infty} \mathbb{E}\|f - \hat{f}_{\hat{h}}\|_2^2 > 0$$

Note on the proof:
The heart of the proof is to control the deviations of $S(h,h') = n^{-1}\sum_{i=1}^n (K_{h'} - K_h)(x - X_i) - \mathbb{E}((K_{h'} - K_h)(x - X_i))$, and we prove that with high probability

$$(1-\varepsilon)\frac{\|K_{h'} - K_h\|_2}{\sqrt{n}} \leqslant \|S(h,h')\|_2 \leqslant (1+\varepsilon)\frac{\|K_{h'} - K_h\|_2}{\sqrt{n}}.$$

To do this we use here deviation in both sides, as cited in Lemma 1. Using this control, we can evaluate the precise behavior of $A(h)$ and then $\hat{h}$. ■

This theorem is proved in [L17] for more general kernels and bandwidth sets. It ensures that the critical value for the parameter a is 1. Beyond this value, the selected bandwidth $\hat{h}$ is of order $1/n$, which is very small (remember that for minimax study of a density with regularity α, the optimal bandwidth is $n^{-1/(2\alpha+1)}$), then the risk cannot tend to 0.

Simulations and discussion Let us now illustrate the role of tuning parameter a, the constant in the penalty term V. The aim is to observe the evolution of the risk for various values of a. Is the critical value $a = 1$ observable in practice? To do this, we simulate data $X_1, \ldots, X_n$ for several densities f. Next, for a grid of values for a, we compute the selected bandwidth $\hat{h}$, the estimator $\hat{f}_{\hat{h}}$ and the integrated risk $\|\hat{f}_{\hat{h}} - f\|_2^2$.

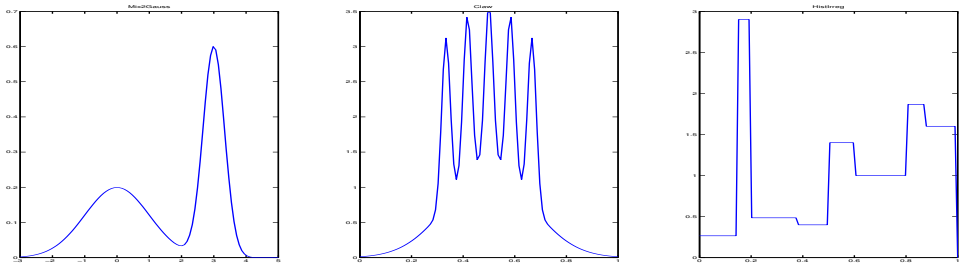


Figure 6: Plots of true density f for Examples 4–6

We consider the following examples, see Figure 6:

Example 1: f is the Cauchy density

Example 2: f is the uniform density $\mathcal{U}(0,1)$

Example 3: f is the exponential density $\mathcal{E}(1)$

Example 4: f is a mixture of two normal densities $\frac{1}{2}\mathcal{N}(0,1) + \frac{1}{2}\mathcal{N}(3,9)$

Example 5: f is a mixture of normal densities sometimes called Claw

Example 6: f is a mixture of eight uniform densities

We implement the method for various kernels, but we only present results for Gaussian kernel, since the choice of kernel does not modify the results. On the other hand, the method is sensitive to the choice of bandwidths set $\mathcal{H}$: here we use

$$\mathcal{H} = \{e^{-k}, 3 \leqslant k \leqslant 10\} \cup \{0.002 + k \times 0.02, 0 \leqslant k \leqslant 24\}.$$

For $n = 5000$ and $n = 50000$, and several values of a, the Figure 7 plots

$$C_0 = \tilde{\mathbb{E}} \frac{\|\hat{f}_{\hat{h}} - f\|_2^2}{\min_{h \in \mathcal{H}} \|\hat{f}_h - f\|_2^2}$$

where $\tilde{\mathbb{E}}$ means the empirical mean on $N = 50$ experiments. Thus smaller C_0 better the estimation. Moreover, we also plot on Figure 8 the selected bandwidth compared to the optimal bandwidth in the selection (for $N = 1$ experiment), i.e.

$$\hat{h} - h_0 \qquad \text{where} \qquad \|\hat{f}_{h_0} - f\|_2^2 = \min_{h \in \mathcal{H}} \|\hat{f}_h - f\|_2^2.$$

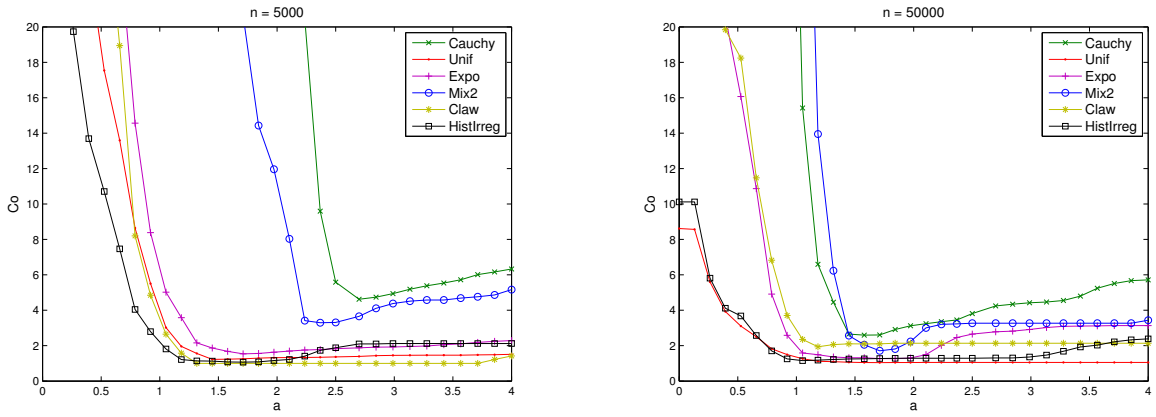


Figure 7: Oracle constant C_0 as a function of a, for Examples 1–6

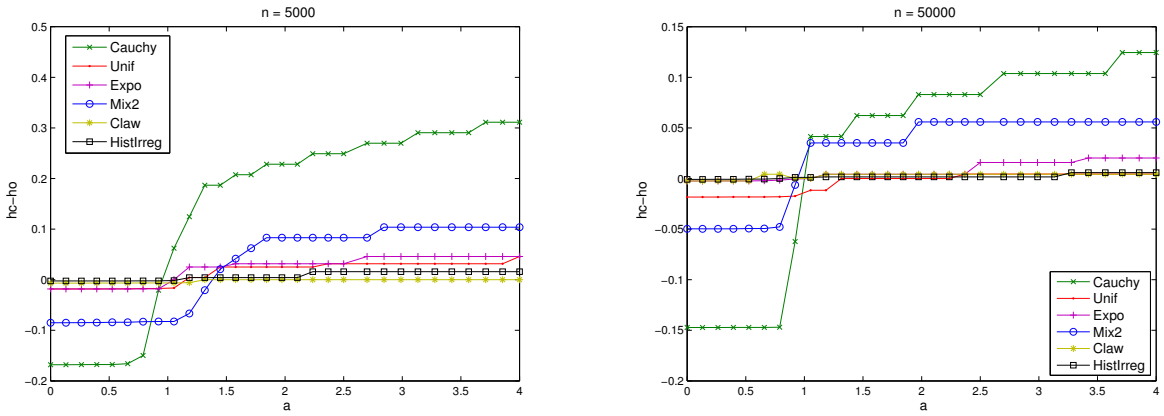


Figure 8: $\hat{h} - h_0$ as a function of a, for Examples 1–6

We can observe that the risk (and then the oracle constant C_0) is very high for small values of a, as expected. Then it jumps to a small value, that indicates the method begins to work well. For too large values of a the risk finally goes back up. Thus we observe in practice what was announced by the theory. Notice that the theory is asymptotic. That is why in practice, the jump may be not exactly at $a = 1$, especially for small values of n. For irregular densities

(examples 2, 5, 6), the optimal bandwidth is very low, then it is consistent to observe a smaller jump for the bandwidth choice. However the jump does exist and this is the interesting point.

To precisely calibrate the penalty V, we face a practical problem: just before $a = 1$, the risk explodes, and just after the result is optimal. Then we could consider another procedure:

$$A(h) = \sup_{h' \leqslant h} \left[\|\hat{f}_{h'} - \hat{f}_h\|_2^2 - a\frac{\|K_{h'}\|_2^2}{n} \right]_+ ,$$
$$\hat{h} = \arg\min_{h \in \mathcal{H}} \left\{ A(h) + b\frac{\|K_h\|_2^2}{n} \right\}.$$

with $b \neq a$ (here we just study the case $a = b$). The distinction between a and b could enable a best calibration. Preliminary computations indicate that $a = 1$ and $b = 2$ may be optimal. A good track for practical purpose seems to use the procedure of Section 2.6.2 to find a_0 where there is a jump in the risk: $a_0 = 1$ in the theory but could be slightly different in practice (simulations show that this jump is very perceptible), and then to choose $b = 2a_0$.

2.7 Some prospects

Several prospects are natural in the area of conditional density estimation. Inference on conditional distribution could be completed by the study of confidence bands, in the spirit of Giné and Nickl (2010). Even more recently, models of privacy have been introduced, in which data remain private even from the statistician, see for example Duchi et al. (2014). Nonparametric estimation of the conditional density in this framework could be a challenging task. In a more classic way, data are generally spoiled by a noise and are not directly workable: estimation for conditional law could be more realistic in this context.

However the most promising research direction is the one of the great dimension. That is why I have proposed with Vincent Rivoirard a PhD project starting in September 2015.

2.7.1 Conditional distribution in large dimension: PhD project

In Section 2.4, we have seen the interest of conditional density estimation for ABC methods. However, our procedure suffers from the curse of dimensionality and is not very adapted for dimensions greater than 2 or 3. The aim of this thesis project is then to propose an alternative to classical ABC methods, combining the sharpness of nonparametric kernel methods and the speed of greedy algorithms. Several goals are pursued: an automatic calibration of the procedure leading to an easy use for practitioners, a reasonable running time when the dimension of the conditional distributions is of order a few tens, and the theoretical validation of the implemented procedures via oracle or minimax approaches.

We consider again the issue of estimating a conditional density. To do this, we assume that we are given a n-sample $(X_i, Y_i)_{1 \leqslant i \leqslant n}$ of couples of random vectors with respective dimensions d_1 and d_2. We denote by $s(\cdot, x)$ the conditional density of Y_i given $X_i = x$, to estimate. The fundamental assumption that we propose (in order to avoid the curse of dimensionality) is that the function s depends only on a small number of variables $k < d := d_1 + d_2$. Formally, this means that there exists an unknown set $\mathcal{I}$ with cardinality k such that, for all vector $u = (x, y) \in \mathbb{R}^d$ with $u = (x_1, \ldots, x_{d_1}, y_1 \ldots, y_{d_2})$

$$s(u) = s(u_{\mathcal{I}}),$$

where $u_{\mathcal{I}} = (u_j : \ j \in \mathcal{I})$. This assumption of sparsity is less restrictive than the structural constraints proposed by Raskutti et al. (2012) or Bouaziz and Lopez (2010) and we expect a benefit in term of estimation rate, if we achieve a good inference for the set $\mathcal{I}$.

The first goal of this thesis is to develop a greedy algorithm to infer $(\mathcal{I}, f)$. Greedy algorithms are studied since several decades in signal processing and approximation theory (Davis et al., 1997; Tropp, 2004). Their use in statistics is more recent (Barron et al., 2008; Lafferty and Wasserman, 2008, 2007). In this area, they are used for variable selection issues, where each step of the algorithm proposes a decision rule for the inclusion or not of a new variable. In a second step, this new estimator could be applied to real data, interacting with population geneticists, with a comparison with current ABC methods. From the mathematical point of view, it would be interesting to reconsider the theoretical contributions of other variable selection methods (Bertin and Lecué, 2008; Comminges and Dalalyan, 2012) in the framework of conditional density estimation. Finally, other related statistical issues can be considered. For instance, we may want to estimate the set $\mathcal{I}$ in a very large dimension (which can possibly depend on n, polynomially or exponentially). In this case, we shall try to provide a multiple testing procedure, based on test statistics taken from the previous estimation work.

2.7.2 Calibration

In the more general context of adaptive estimation, there remain a lot of issues: see the open problems listed in Lepski (2014). In particular, it remains difficult to find optimally-adaptive estimator whose construction is computationally reasonable. In the line of Section 2.6.2, I would like to keep exploring the Goldenshluger-Lepski method in the framework of density estimation. A collaboration with Pascal Massart and Vincent Rivoirard has already started to study this topic. We consider the asymmetric procedure:

$$A(h) = \sup_{h' \leqslant h} \left[\|\hat{f}_{h'} - \hat{f}_h\|_2^2 - a\frac{\|K_{h'}\|_2^2}{n} \right]_+ ,$$
$$\hat{h} = \arg\min_{h \in \mathcal{H}} \left\{ A(h) + b\frac{\|K_h\|_2^2}{n} \right\} .$$

with $b \neq a$. This distinction between a and b may seem slight, but it makes appear very different behavior of the estimator, both theoretically and practically. In this framework, finding minimal and optimal penalty is a challenge. One can also consider a more refined penalty, which should depend on h and h' instead of only h. Other important investigations include the multivariate case and the use of diverse loss functions.

Chapter 3

Indirect observations models

3.1 Deconvolution on $\mathbb{R}^d$

3.1.1 Context

Model The so-called convolution model is the following one:

$$Y_i = X_i + \varepsilon_i, \qquad i = 1, \dots, n,$$

where (ε_i) and (X_i) are two independent sequences of i.i.d. variables. The aim is to estimate the density f of the signal X_i when only the noised variables Y_i are observed. Since in this case

$$f_Y = f \star f_\varepsilon = \int f(. - u) f_\varepsilon(u) du,$$

this issue is named deconvolution. Motivations and applications for this problem are numerous. On can cite:

- analysis of DNA content obtained by microfluorometry (Mendelsohn and Rice, 1982),
- intensity of a probe in genome-wide microarrays (Plancade et al., 2012),
- in astrophysics, density of metallicities of F and G dwarfs (Bissantz et al., 2007),
- in medical statistics: measures of peak expiratory flow rate, or the ventricule-brain ratio (Delaigle et al., 2008); health effects of radiation exposure (Stefanski and Carroll, 1990); estimation of onset of pregnancy (Comte et al., 2014),
- other applications in econometrics, medicine or astronomy can be found in Meister (2009) and Stefanski and Carroll (1990).

This model is closely linked with the regression with errors-in-variables, that we shall not evoke here. Another related model is the so-called Inverse Problem. Indeed, denoting $A : f \mapsto f \star F_\varepsilon$ and $U_n = \sqrt{n}(\widehat{(F_Y)}_n - F_Y)$, we can write $\widehat{(F_Y)}_n = Af + \frac{1}{\sqrt{n}} U_n$, where U_n is an empirical process which converges to a Brownian bridge $B_{F_Y(t)}$. However this link is rather artificial and actually inverse problems are about regression when deconvolution is about density estimation.

Note that without additional assumption, this problem is not identifiable. The classical way to make the problem identifiable is to assume that the distribution of the noise ε is known.

Main references There have been a lot of studies dedicated to the problem of recovering the distribution f of a signal when it is measured with an additive noise with known density. See Carroll and Hall (1988), Devroye (1989), Liu and Taylor (1989), Masry (1991), Stefanski and Carroll (1990), Zhang (1990), Hesse (1999), Cator (2001), Delaigle and Gijbels (2004) for mainly kernel methods, Koo (1999) for a spline method, Pensky and Vidakovic (1999) for wavelet strategies, and Comte et al. (2006) for adaptive projection strategies. Efromovich (1997) and Goldenshluger (2002) study the case of circular data. The book of Meister (2009) reviews the issue of deconvolution. The question of the optimality of the rates revealed real difficulties, after the somehow classical cases studied by Fan (1991) and the case of logarithmic rates studied by Goldenshluger (1999). The case of supersmooth noise (i.e. with exponential decay of its characteristic function) in presence of possibly also supersmooth density implies non standard bias variance compromises that require new methods for proving lower bounds. These problems have been studied by Butucea (2004), Butucea and Tsybakov (2007, 2008), [L1] and by Butucea and Comte (2009). See also Li and Liu (2014) for $\mathbb{L}^p$-risk. Recent advances concern uniform risk and confidence bands: see Lounici and Nickl (2011) and Bissantz et al. (2007).

Other estimation problems in the convolution model The issue of change-point estimation in this context of noisy observations has been addressed by Neumann (1997b) and Goldenshluger et al. (2006), when Schmidt-Hieber et al. (2013) are interested in detecting qualitative features of the unknown density, for example testing for local monotonicity. In Chesneau (2011), f is assumed to be a mixture with unknown components. Chesneau et al. (2015) deal with the estimation of the l-fold convolution $f^{\star(l)}$ in the convolution model.

More important is the issue of estimating the cumulative distribution function. This problem has been studied by Hall and Lahiri (2008), Dattner et al. (2011), Söhl and Trabs (2012), Dattner and Reiser (2013). See also Dattner et al. (2013) for quantile estimation. The distribution measure of X is also estimated with a Wasserstein control in Dedecker et al. (2015) and references therein.

Particular cases of noise modelization In the main references, it is always assumed that the characteristic function of the noise never vanishes. But new directions lead researchers to release this assumption: see Hall and Meister (2007); Meister (2008). Deconvolution when the noise is uniform is studied in van Es (2011) and Feuerverger et al. (2008). See also Abbaszadeh et al. (2013) for uniform multiplicative noise. Deconvolution when the noise is Laplace or Gamma is studied in van Es and Kok (1998), and Gaussian deconvolution is addressed in Masry and Rice (1992). In Mabon (2015), both random variables X and ε are assumed nonnegative.

Partly known noise It turns out that the knowledge of the error distribution is often an unrealistic assumption. Then it is important to relax this assumption, at least partially. A popular deconvolution procedure, the so-called SIMEX estimator only requires the knowledge of order-two-moments of the noise, see Stefanski and Cook (1995). See also Meister (2004), who studied the effect of misspecifying the error density.

Several authors consider a semiparametric approach for the case of normal error distribution with unknown variance : Matias (2002), Meister (2006), Schwarz and Van Bellegem (2010). In the papers of Butucea and Matias (2005), Butucea et al. (2008) and Meister (2006) this model is considered under the assumption that the Fourier transform of f has a specific known positive lower bound, so that f is finally identifiable. Meister (2007) establishes consistency in a model where f is compactly supported, and the noise characteristic function has to be known on some bounded interval.

Other models with supplementary observations Sometimes, the problem of unknown noise density can be circumvented by repeated observations of the same variable of interest, each time with an independent error. This is the model of panel data, see for example Li and Vuong (1998), Delaigle et al. (2008), Neumann (2007) or Comte and Samson (2012). Another model of double measurement is proposed in Meister et al. (2010).

In physical contexts, it is often possible to obtain samples of noise alone. This is this point of view which is considered in our Section 3.1.3 (see references therein).

Rates of convergence for classical deconvolution Now let us present the classical results for univariate deconvolution. In classical deconvolution, Fourier analysis is used to write:

$$f_Y^*(t) = \mathbb{E}[e^{itY}] = \mathbb{E}[e^{itX}]\mathbb{E}[e^{it\varepsilon}] = f^*(t)f_\varepsilon^*(t)$$

and $f_Y^*(t) = \mathbb{E}[e^{itY}]$ can be estimated by its empirical version. Then the inverse of the characteristic function of the noise plays a crucial role. When this characteristic function decreases with an exponential rate, as in the normal case (or Cauchy distribution), the deconvolution is particularly hard, and the rate of convergence suffer these effects. By the way it is a classical way to smooth a function: to convolve it with a normal density. The initial function is then difficult to retrieve after this smoothing. When the noise is less regular, with a polynomial decrease of the Fourier transform (e.g. Laplace, Gamma distributions) the reconstruction is easier. Then the rates of convergence for the problem of estimating f are different according to whether the noise is "supersmooth" (exponential decrease of the characteristic function, denoted SS) or "ordinary smooth" (polynomial decrease of the characteristic function, denoted OS). In the same way, the target f can be assumed to be supersmooth or ordinary smooth. Then it is now well-known that the rates of convergence for the problem of deconvolution in $\mathbb{R}$ are the following:

		noise OS	noise SS
f	OS	$n^{-\Diamond}$	$(\log n)^{-\triangle}$
f	SS	$\dfrac{(\log n)^{\triangledown}}{n}$	$\dfrac{(\log n)^{\triangledown}}{n} < . < (\log n)^{\triangle}$

We do not give rates in the case where both functions can be supersmooth, because it is very intricate. General formula in dimension 1 are given in [L1], see also Butucea and Tsybakov (2007, 2008). For example, if the signal is $\mathcal{N}(0,\sigma^2)$ and the noise $\mathcal{N}(0,\sigma_\varepsilon^2)$, the rate is $n^{-1/(1+\theta^2)}[\log(n)]^{-(1+1/(1+\theta^2))/2}$ with $\theta^2 = \sigma_\varepsilon^2/\sigma^2$. We can just emphasize that in such case the rates can be considerably improved, compared to the logarithmic issue above.

Notation In the sequel, we denote by g^* the Fourier transform of an integrable function g, $g^*(t) = \int e^{i\langle t,x\rangle} g(x)dx$ where $\langle t,x\rangle = \sum_{j=1}^d t_j x_j$ is the standard scalar product in $\mathbb{R}^d$. Moreover the convolution product of two functions g_1 and g_2 is denoted by $g_1 \star g_2(x) = \int g_1(x-u)g_2(u)du$. We recall that $(g_1 \star g_2)^* = g_1^* g_2^*$. As usual, we define

$$\|g\|_1 = \int |g(x)|dx \quad \text{and} \quad \|g\| = \|g\|_2 = \left(\int |g(x)|^2 dx\right)^{1/2}.$$

The notation x_+ means $\max(x, 0)$, and $a \leqslant b$ for $a, b \in \mathbb{R}^d$ means $a_1 \leqslant b_1, \ldots, a_d \leqslant b_d$. For two functions u, v, we denote $u(x) \lesssim v(x)$ if there exists a positive constant C not depending on x such that $u(x) \leqslant Cv(x)$ and $u(x) \approx v(x)$ if $u(x) \lesssim v(x)$ and $v(x) \lesssim u(x)$.

3.1.2 Multivariate deconvolution

We consider the following d-dimensional convolution model

$$Y_i = \begin{pmatrix} Y_{i,1} \\ \vdots \\ Y_{i,d} \end{pmatrix} = X_i + \varepsilon_i = \begin{pmatrix} X_{i,1} \\ \vdots \\ X_{i,d} \end{pmatrix} + \begin{pmatrix} \varepsilon_{i,1} \\ \vdots \\ \varepsilon_{i,d} \end{pmatrix}, \ i = 1, \ldots, n. \tag{12}$$

We assume that the ε_i's and the X_i's are i.i.d. and the two sequences are independent. Only the Y_i's are observed and our aim is to estimate the density f of X_1 when the density f_ε of ε is known.

Almost all previous works are in one dimensional setting. Our aim here is to study the multidimensional setting, and to propose adaptive strategies that would take into account possible anisotropy for both the function to estimate and the noise structure. Few papers study the multidimensional deconvolution problem; we can only mention Masry (1991, 2003) who considers mainly the problem of dependency between the variables without anisotropy nor adaptation, and Youndjé and Wells (2008) who consider a cross-validation method for bandwidth selection in an isotropic and ordinary smooth setting. Our work considerably generalizes their results with a different method, and provides new results and new rates in both pointwise and global setting. We have to mention also Dedecker and Michel (2013) who estimate geometric features of the distribution of X_i which is assumed to be supported on an unknown compact subset G of $\mathbb{R}^d$. As already explained in Kerkyacharian et al. (2001), adaptive procedures are delicate in a multidimensional setting because of the lack of natural ordering. For instance, the model selection method is difficult to apply here since it requires to bound terms on sums of anisotropic models. Here we use the Goldenshluger-Lepski methodology to face anisotropy problems, with the use of Talagrand inequality as the key of the deviation.

The estimator Let us now define our collection of estimators. It easily follows from Model (12) and independence assumptions that, if f_Y denotes the common density of the Y_j's, then $f_Y = f \star f_\varepsilon$ and thus $f_Y^* = f^* f_\varepsilon^*$. Note that this basic equality can be obtained for a noise with discrete distribution, and the whole method can be generalized to that case. Therefore, under the classical assumption:

Assumption (E) $\forall x \in \mathbb{R},\ f_\varepsilon^*(x) \neq 0$

the equality $f^* = f_Y^*/f_\varepsilon^*$ yields an estimator of f^* by considering the following estimate of f_Y^*:

$$\widehat{f_Y^*}(t) = \frac{1}{n} \sum_{k=1}^{n} e^{i\langle t, Y_k \rangle}.$$

Then, we should use inverse Fourier transform to get an estimate of f. As $1/f_\varepsilon^*$ is in general not integrable (consider a Gaussian density for instance), this inverse Fourier transform does not exist, and a smoother is used. Let K be a kernel in $\mathbb{L}^2(\mathbb{R}^d)$ such that K^* exists. Then we define, for $h \in (\mathbb{R}_+^*)^d$,

$$K_h(x) = \frac{1}{h_1 \ldots h_d} K\left(\frac{x_1}{h_1}, \ldots, \frac{x_d}{h_d}\right) \text{ and } L_{(h)}^*(t) = \frac{K_h^*(t)}{f_\varepsilon^*(t)}.$$

The kernel K is such that Fourier inversion can be applied:

$$L_{(h)}(x) = (2\pi)^{-d} \int e^{-i\langle t,x\rangle} \frac{K_h^*(t)}{f_\varepsilon^*(t)} dt,$$

A natural estimator of f is such that

$$\hat{f}_h^*(t) = K_h^*(t) \frac{\widehat{f_Y^*}(t)}{f_\varepsilon^*(t)} = \widehat{f_Y^*}(t) L_{(h)}^*(t),$$

since, under **(E)**, f_ε^* does not vanish, and thus, by Fourier inversion,

$$\hat{f}_h(x) = \frac{1}{n} \sum_{k=1}^{n} L_{(h)}(x - Y_k).$$

For the estimators to be correctly defined, the kernel must be chosen sufficiently regular to recover integrability in spite of the noise density. In this dissertation, we simply use the sinus cardinal kernel denoted by $K = \text{sinc}$ and defined by $K(t) = K_1(t_1) \dots K_d(t_d)$ with

$$K_j^*(t) = \mathbb{1}_{[-1,1]}(t) \Leftrightarrow K_j(x) = \frac{\sin(x)}{\pi x} \quad \left(K_j(0) = \frac{1}{\pi} \right).$$

But many other kernels are possible: see [L12].

Study of the integrated risk A usual, the integrated risk can be decomposed in a bias term plus a variance term: $\mathbb{E}\|\hat{f}_h - f\|^2 = B^2(h) + V(h)$. Let us first study the variance. A straightforward computation gives

$$V(h) := \mathbb{E}\|\hat{f}_h - \mathbb{E}(\hat{f}_h)\|^2 \overset{\text{Parseval}}{=} \frac{1}{(2\pi)^d} \mathbb{E} \int |\hat{f}_h^* - \mathbb{E}(\hat{f}_h^*)|^2 = \frac{1}{(2\pi)^d} \int \text{Var}(\hat{f}_h^*)$$

$$= \frac{1}{(2\pi)^d} \int \text{Var}\left(\frac{K_h^*}{f_\varepsilon^*} \widehat{f_Y^*} \right) = \frac{1}{(2\pi)^d n} \int \text{Var}\left(\frac{K_h^*}{f_\varepsilon^*} e^{i\langle u, Y_1\rangle} \right) \leqslant \frac{1}{(2\pi)^d n} \left\| \frac{K_h^*}{f_\varepsilon^*} \right\|^2.$$

To understand the behavior of this quantity, it is necessary to make assumptions on the noise. We assume that the characteristic function of the noise has a polynomial or exponential decrease. We denote by OS (for ordinary smooth) the set of directions j with ordinary smooth regularity and by SS (for supersmooth) the set of directions j with supersmooth regularity. Thus we suppose that

$$|f_\varepsilon^*(t)| \approx \prod_{j \in OS} (t_j^2 + 1)^{-\beta_j/2} \prod_{k \in SS} (t_k^2 + 1)^{-\beta_k/2} \exp(-\alpha_k |t_k|^{\rho_k}).$$

With this hypothesis $V(h) \lesssim \frac{1}{n} \prod_{j=1}^d h_j^{-1-2\beta_j+\rho_j} \exp(2\alpha_j h_j^{-\rho_j})$. It means that, when $h \to 0$, the variance grows polynomially for OS noise components and exponentially for SS noise components.

Let us now study the bias term. The estimator verifies

$$\mathbb{E}(\hat{f}_h^*(t)) = K_h^*(t) \frac{f_Y^*(t)}{f_\varepsilon^*(t)} = K_h^*(t) f^*(t) = (K_h \star f)^*(t)$$

so that $B(h) := \|f - \mathbb{E}(\hat{f}_h)\| = \|f - K_h \star f\|$. To express the regularity of the target f, we shall consider general anisotropic Sobolev spaces $\mathcal{S}(b, a, r, L)$ defined as the class of integrable functions $f : \mathbb{R}^d \to \mathbb{R}$ satisfying

$$\sum_{j=1}^{d} \int |f^*(t_1, \dots, t_d)|^2 (1 + t_j^2)^{b_j} \exp(2a_j |t_j|^{r_j}) dt_1 \dots dt_d \leqslant L^2,$$

for $a_j \geqslant 0, r_j \geqslant 0, b_j \in \mathbb{R}$ ($b_j > 1/2$ if $r_j = 0$), when $j = 1, \ldots, d$. If some a_j are nonzero, the corresponding directions are associated with so-called supersmooth regularities. The spaces of ordinary smooth functions correspond to classic Sobolev classes, while supersmooth functions are infinitely differentiable. It includes for example normal ($r_j = 2$) and Cauchy ($r_j = 1$) densities. Notice that in [L12], we also consider other function spaces (Hölder and Nikolski). We obtain for such a function $B(h) \lesssim L \sum_{j=1}^{d} h_j^{b_j} \exp(-a_j h_j^{-r_j})$ and then

$$\mathbb{E}\|\hat{f}_h - f\|^2 \lesssim L^2 \sum_{j=1}^{d} h_j^{2b_j} \exp(-2a_j h_j^{-r_j}) + \frac{1}{n} \prod_{j=1}^{d} h_j^{-1-2\beta_j+\rho_j} \exp(2\alpha_j h_j^{-\rho_j}).$$

It appears that the bias-variance balance is very complex and depends on the behavior of each coordinate of the noise and the signal.

Rates of convergence To understand the diversity of possible rates, let us first present some examples.

Example 1 - Cauchy distribution: $f(x,y) = (\pi^2(1+x^2)(1+y^2))^{-1}$ with a Laplace/Laplace noise, i.e.

$$f_\varepsilon(x,y) = \frac{\lambda^2}{4} e^{-\lambda|x|} e^{-\lambda|y|}; \quad f_\varepsilon^*(x,y) = \frac{\lambda^2}{\lambda^2+x^2} \frac{\lambda^2}{\lambda^2+y^2}.$$

The smoothness parameters are $b_1 = b_2 = 0$, $r_1 = r_2 = 1$, $\beta_1 = \beta_2 = 2$ and $\rho_1 = \rho_2 = 0$. For this example, we can compute that the rate is upperbounded by $(\log(n))^{10}/n$.

Example 2 - Mixed Gaussian distribution: $X_{i,1} = W/\sqrt{7}$ with $W \sim 0.4\mathcal{N}(0,1) + 0.6\mathcal{N}(5,1)$, and $X_{i,2}$ independent with distribution $\mathcal{N}(0,1)$. We consider that the noise follows a Laplace/Gaussian distribution, i.e.

$$f_\varepsilon(x,y) = \frac{\lambda}{2} e^{-\lambda|x|} \frac{1}{\mu\sqrt{2\pi}} e^{-y^2/(2\mu^2)}; \quad f_\varepsilon^*(x,y) = \frac{\lambda^2}{\lambda^2+x^2} e^{-\mu^2 y^2/2}.$$

The smoothness parameters are $b_1 = b_2 = 0$, $r_1 = r_2 = 2$, $\beta_1 = 2$, $\beta_2 = 0$ and $\rho_1 = 0$, $\alpha_2 = \mu^2/2$, $\rho_2 = 2$. Here the rate of convergence is $n^{-16/17}[\log(n)]^{63/34}$ for the bandwidths $h_1^{-1} = \sqrt{7\log(n)}$ and $h_2^{-1} = \sqrt{16\log(n) - 40\log(\log(n))}/\sqrt{17}$.

Example 3 - Gamma distribution: $X_{i,1} \sim \Gamma(5, 1/\sqrt{5})$ and $X_{i,2} \sim \Gamma(5, 1/\sqrt{5})$. We estimate the density on $[0,8]^2$. The noise follows a Gaussian/Gaussian distribution, i.e.

$$f_\varepsilon(x,y) = \frac{1}{2\pi\mu^2} e^{-(x^2+y^2)/(2\mu^2)}; \quad f_\varepsilon^*(x,y) = e^{-\mu^2(x^2+y^2)/2}.$$

So $b_1 = b_2 = 5$, $r_1 = r_2 = 0$, $\beta_1 = \beta_2 = 0$, $\alpha_1 = \alpha_2 = \mu^2/2$ and $\rho_1 = \rho_2 = 2$. This is an example with integrated rate $(\log(n))^{-9/2}$ (which is not so large for practical values of n, for instance, for $n = 1000$, this term is smaller than $1/n$).

General detailed rates of convergences are given in [L12]. Since they are very intricate, we just give here the major conclusions. We call "homogeneous" the cases where all components have the same behavior, and "mixed cases" otherwise.

- For homogeneous cases, we obtain natural extensions of the univariate rates, and in particular the important fact that the rates can be logarithmic if the noise is SS (for instance in the Gaussian case) but are much improved if the signal is also SS: for instance, if the signal is also Gaussian, then polynomial rates are recovered.

- We obtain surprising results in the mixed cases: if one component only of the noise is SS (all the others being OS), in presence of an OS signal, then the rate of convergence of the estimator is logarithmic. In this case, no bandwidth selection is required. Indeed, we just have to take $h_j = (\log(n)/2\alpha_j)^{-1/\rho_j}$ for the supersmooth components and $h_j = n^{-1/(2d(2\beta_j+1))}$ for ordinary smooth components, and the rate has a logarithmic order determined by the bias term.
- On the contrary, if the signal has k out of d components SS in presence of an OS noise, then the rate of the estimator is almost as good as if the dimension of the problem was $d-k$ instead of d.

To get a validation of our method, we have proved lower bounds for the rates computed above, at least in part of the cases. In particular, we can extend the results of Fan (1991) and of Butucea and Tsybakov (2008) to the multivariate setting. Our assumption is the following:

Assumption (Hε) We assume that the noise has its components independent. We also assume that, for $j = 1, \dots, d$, and for almost all u_j in $\mathbb{R}$, $f^*_{\varepsilon_{1,j}}(u_j)$ admits a derivative and $|u_j|^{\beta'_j} \exp(\alpha_j |u_j|^{\rho_j}) |(f^*_{\varepsilon_{1,j}})'(u_j)|$ is bounded, for a constant β'_j such that $\beta'_j > \beta_j$ if $\varepsilon_{1,j}$ is OS. If the signal f verifies $1 \leqslant r_j < 2$, we assume that $f^*_{\varepsilon_{1,j}}(u_j)$ admits in addition a second order derivative for almost all u_j in $\mathbb{R}$ such that $|u_j|^{\beta''_j} \exp(\alpha_j |u_j|^{\rho_j}) |(f^*_{\varepsilon_{1,j}})''(u_j)|$ is bounded, with β''_j a positive constant.

Theorem 25 ([L12])**.** *Under assumption* **(E)** *and* **(Hε)**, *consider the following cases:*

Case A *All the the components of ε are ordinary smooth and, for the signal: $r_j < 2$, or*

Case B *There exists at least one component of ε which is supersmooth and the signal is ordinary smooth (all $r_j = 0$)*

Then for any estimator $\hat{f}_n$, and for n large enough,

$$\sup_f \mathbb{E}_f \left[\|\hat{f}_n - f\|^2 \right] \gtrsim \psi_n$$

where ψ_n are the previously described rates.

Note on the proof:
This result is not straightforward and requires specific constructions, since it captures mixed cases which could not be encountered in univariate setting. We need to define a collection of alternatives $(f_\theta)_\theta$ in case A, and a single alternative in case B. If H denotes a specific kernel function and g_s the symmetric stable law with characteristic function $g^*_s(u) = \exp(-|u|^s)$, we introduce $f_0(x) = \prod_{j=1}^d \frac{1}{c_j} g_{s_j}\left(\frac{x_j}{c_j}\right)$ with c_j positive constants large enough and $s_j > r_j$. Then we build alternative functions far from f_0 in $\mathbb{L}^2$ distance but with close corresponding likelihoods. We take in case A densities of the form

$$f_\theta(x) = f_0(x) + c\sqrt{V(h_n)} \sum_{k \in \mathcal{K}} \theta_k \prod_{j=1}^d H\left(\frac{x_j - x_{nkj}}{2h_{n,j}}\right)$$

and in case B

$$f_1(x) = f_0(x) + c \sum_{j=1}^d h_{n,j}^{b_j - 1/2} H\left(\frac{x_j}{h_{n,j}}\right) \prod_{1 \leqslant i \leqslant d,\ i \neq j} H(x_i).$$

Note that our condition on the noise improves Fan (1991)'s conditions: in the OS case, Fan requires a second order derivative of f_ε^* and in the SS case, he gives a technical condition which is difficult to link with the functions at hand. The improvement took inspiration in the book of Meister (2009) who also had first order type conditions. ■

We therefore conclude that the rates reached by our estimators for estimating an ordinary smooth function are optimal. We also have optimality in the case of a supersmooth function f with $r_j < 2$ and ordinary smooth noise.

Adaptation As the bandwidth choice is very difficult to describe in the general case, this enhances the interest of automatic adaptation which is proposed below. In spite of the difficulty of the problem, in particular because of the large number of parameters required to formalize the regularity indexes of the functions, we exhibit very synthetic penalties than can be used in all cases.

We use the Goldenshluger-Lepski method. We introduce auxiliary estimators $\hat{f}_{h,h'} = K_{h'} \star \hat{f}_h$ and a penalty $\tilde{V}(h) = cC(h)V(h)$ where $C(h)$ a corrective term to be defined later. The estimator $\tilde{f} = \hat{f}_{\hat{h}}$ is defined by

$$A(h) = \sup_{h' \in \mathcal{H}} \left[\|\hat{f}_{h'} - \hat{f}_{h,h'}\| - \sqrt{\tilde{V}(h')} \right]_+$$
$$\hat{h} = \arg\min_{h \in \mathcal{H}} \left\{ A(h) + \sqrt{\tilde{V}(h)} \right\}$$

where $\mathcal{H}$ is the set of bandwidth to be defined later. We obtain the result

Theorem 26 ([L12])**.** *Under assumption* **(E)***, if we choose*

$$\begin{aligned} \mathcal{H} &= \{h^{(k)}, C(h) \max(1, \|K_h^*/f_\varepsilon^*\|_2^2 / \|K_h^*/f_\varepsilon^*\|_\infty^2) \geqslant (\log n)^2, \\ &\quad V(h^{(k)}) \leqslant 1 \quad \textit{for } k = 1, \dots, \lfloor n^\epsilon \rfloor \}. \end{aligned}$$

then, with probability larger than $1 - n^\epsilon e^{-K(\log n)^2}$

$$\|\tilde{f} - f\| \leqslant 3 \inf_{h \in \mathcal{H}} \left\{ B(h) + \sqrt{\tilde{V}(h)} \right\}.$$

Moreover

$$\mathbb{E}(\|\tilde{f} - f\|) \leqslant 3 \inf_{h \in \mathcal{H}} \left\{ B(h) + \sqrt{\tilde{V}(h)} \right\} + \frac{C}{\sqrt{n}}.$$

The bias-variance trade-off is then achieved as soon as $\tilde{V}(h) \sim V(h)$, i.e. when the corrective $C(h)$ is close to 1. To choose $C(h)$, we have to consider that the optimal rate of convergence will be obtained only if the optimal bandwidth $h_{opt} \in \mathcal{H}$. The question is then: is it possible to choose $C(h)$ close to 1 and to have $h_{opt} \in \mathcal{H}$? To give the answer, we have to distinguish two cases:

f **ordinary smooth** In we know that we are in this case, and if there exists a supersmooth component in the noise (case of very low rates), then $h_{opt} \in \mathcal{H}$ does not depend on f, then an adaptive procedure is useless. If all the components are ordinary smooth, then one can take $C(h) = 1$ and $h_{opt} \in \mathcal{H}$. So the optimal rates are achieved.

f **supersmooth** In all cases, a corrective $C(h) = (\log n)^2$ will always work. It implies a rate spoiled with a $(\log n)^2$ (which is negligible). For the mean oracle inequality, we can even take a $C(h)$ smaller which leads to an optimal rate if the noise is ordinary smooth or weakly supersmooth, see [L12].

Remark on the pointwise risk The pointwise risk of the same estimator can be precisely studied. Both bias and variance are slightly modified, that leads to different but similar rates of convergence, see [L12]. To build an adaptive estimator we define in this case $\tilde{V}_0(h) = c_0 \log(n) V_0(h)$ where $V_0(h)$ is the variance of estimator $\hat{f}_h$ at given point x_0. We also set

$$A_0(h, x_0) = \sup_{h' \in \mathcal{H}_0} \left[|\hat{f}_{h'}(x_0) - \hat{f}_{h,h'}(x_0)| - \sqrt{\tilde{V}_0(h')} \right]_+, \quad \hat{h}(x_0) = \arg\min_{h \in \mathcal{H}_0} \left\{ A_0(h, x_0) + \sqrt{\tilde{V}_0(h)} \right\}$$

and the final estimator is $\tilde{f}(x_0) = \hat{f}_{\hat{h}(x_0)}(x_0)$. With this procedure, we obtain both trajectorial and mean oracle inequalities, similar to Theorem 26, with a corrective term always equal to $\log(n)$. This always leads to the optimal rates with respect to a sample size $n/\log(n)$.
This logarithmic loss, due to adaptation, is known to be nevertheless adaptive optimal for $d = 1$, see Butucea and Tsybakov (2007, 2008) and Butucea and Comte (2009), and we can conjecture that it is also the case for larger dimension.

Numerical illustrations In this section, we consider the case $d = 2$. The kernel sinc has good properties for practical purposes. Denoting $\varphi_{h,j}(x) = \frac{\pi}{\sqrt{h_1 h_2}} K(\frac{x_1}{h_1} - \pi j_1, \frac{x_2}{h_2} - \pi j_2)$ we can prove that $\hat{f}_h = \sum_j \hat{a}_j^h \varphi_{h,j}$ and the coefficients $\hat{a}_j^h$ can be computed with Fast Fourier Transform algorithm, remarking that

$$\hat{a}_j^h = \frac{1}{4\pi^2} \int \hat{f}_h^* \varphi_{h,j}^* = \frac{\sqrt{h_1 h_2}}{4\pi} \int_{-1/h_1}^{1/h_1} \int_{-1/h_2}^{1/h_2} \frac{\widehat{f_Y^*}}{f_\varepsilon^*}(u_1, u_2) e^{i\pi(u_1 h_1 j_1 + u_2 h_2 j_2)} du_1 du_2.$$

Moreover $\hat{f}_{h,h'} = \hat{f}_{h \vee h'}$ where $h \vee h' = (\max(h_1, h_1'), \max(h_2, h_2'))$, and $\|\hat{f}_{h'} - \hat{f}_{h \vee h'}\|^2 = \|\hat{f}_{h'}\|^2 - \|\hat{f}_{h \vee h'}\|^2$. Thus our estimator can be computed very fast. We take $\mathcal{H}$ and $\mathcal{H}_0$ included in $\{4/m, 1 \leqslant m \leqslant 3n^{1/4}\}$, $\tilde{V}_0(h) = 0.01 \log(n) V_0(h)$ and $\tilde{V}(h) = 0.05 \log^2(n) V(h)$ (the calibration of constants has been done on a preliminary training set of various examples). We compute estimators for Examples 1–3 with $\lambda = 6$, $\mu = 1/4$.

For each path, we compare the MISE for the global procedure with the minimum risk for all bandwidths of the collection. Let us define

$$C_{\text{oracle}} = \mathbb{E}\left(\frac{\|\tilde{f} - f\|^2}{\inf_{h \in \mathcal{H}} \|\hat{f}_h - f\|^2} \right).$$

Then empirical version of C_{oracle} averaged over 100 samples is given below.

	$n = 100$	$n = 300$	$n = 500$	$n = 750$	$n = 1000$
Ex 1	1.48	2.04	2.01	1.96	1.97
Ex 2	1.08	1.03	1.05	1.07	1.25
Ex 3	1.36	1.53	1.57	1.57	1.62

It shows that the adaptation is performing, since the risk for the chosen $\hat{h}$ is very close to the best possible in the collection (the nearest of one C_{oracle}, the better the algorithm).

We also illustrate the results with some figures. Figure 9 shows the surface $z = f(x, y)$ for Example 2 and the estimated surface $z = \tilde{f}(x, y)$ obtained by global bandwidth selection. For more visibility, sections of the previous surface are drawn. We can see the curves $z = f(x, -0.3)$ versus $z = \tilde{f}(x, -0.3)$ and the curves $z = f(-0.3, y)$ versus $z = \tilde{f}(-0.3, y)$. For this figure, the selected bandwidth is $\hat{h} = (0.29, 0.57)$. Thus, the bandwidth in the first direction is

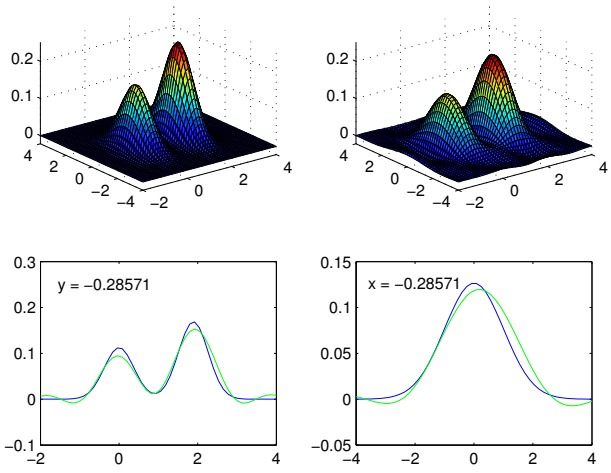


Figure 9: Example 2, global bandwidth selection, with $n = 500$. Top right: true density f, top left: estimator $\check{f}$, bottom: sections, dark line for f and light line for the estimator.

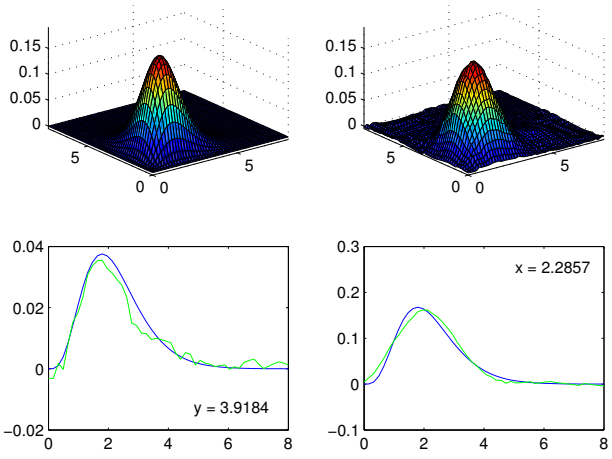


Figure 10: Example 3, pointwise bandwidth selection,with $n = 500$. Top right: true density f, top left: estimator $\tilde{f}$, bottom: sections, dark line for f and light line for the estimator.

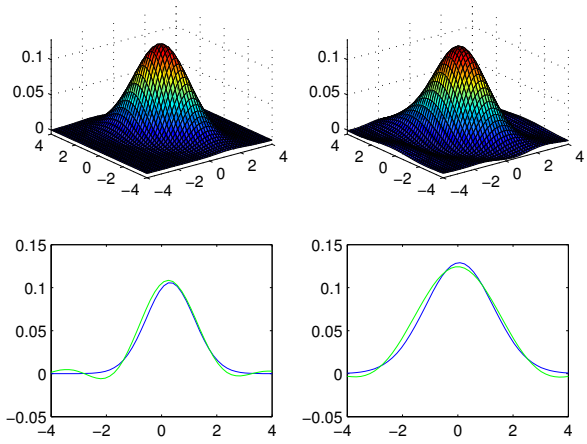


Figure 11: Dependent case, global bandwidth selection,with $n = 500$. Top right: true density f, top left: estimator $\tilde{f}$, bottom: sections, dark line for f and light line for the estimator

twice smaller, to recover the two modes: this shows that our procedure takes really anisotropy into account. Figure 10 shows an analogous illustration for Example 3, but with a pointwise bandwidth selection. We obtain a slightly more angular figure.

To conclude this section, we would like to mention that we can keep good results even in case of dependent components of both the noise and the signal. More precisely, we can take $X \sim \mathcal{N}(0, \Sigma)$ and $\varepsilon \sim \mathcal{N}(0, \Sigma_\varepsilon)$ with $\Sigma = \begin{pmatrix} 1 & -0.7 \\ -0.7 & 2 \end{pmatrix}$ and $\Sigma_\varepsilon = 10^{-2} \begin{pmatrix} 1 & 0.25 \\ 0.25 & 1.0625 \end{pmatrix}$, with X and ε independent. We present in Figure 11 an illustration of this example.

3.1.3 Case of an unknown noise distribution

We still consider the model of sample

$$Y_j = X_j + \varepsilon_j \qquad j = 1, \dots, n$$

where the aim is to estimate the density f of X when only $Y_1, \dots, Y_n$ are observed. Until now, we have always assumed that the error distribution was known. Unfortunately, it is clear that this assumption is often unrealistic. Sometimes, this problem can be circumvented by repeated observations of the same variable of interest, each time with an independent error. However, there are also many application fields where it is not possible to do repeated measurements of the *same* variable. In that case, information about the error distribution can be drawn from an additional experiment: a training set is used by experimenters to estimate the noise distribution. Think of ε as a measurement error due to the measuring device; then preliminary calibration measures can be obtained in the absence of any signal X (this is often called the instrument line shape of the measuring device). Mathematically, this means that the knowledge of f_ε can be replaced by observations $\varepsilon_{-1}, \dots, \varepsilon_{-M}$, a noise sample with distribution f_ε, independent of $(Y_1, \dots, Y_n)$. It has the advantage that only one measuring device is needed, instead of two or more for repeated measurement strategies. Note that the availability of two distinct samples makes the problem identifiable.

Actually, this is a natural method used by practitioners. One of the most typical domains where a preliminary estimation of the measurement error is done is spectrometry, or spectrofluorimetry, and we detail in [L10] an example in microscopy from Odiachi and Prieve (2004).

It is worth mentioning that the problem of adaptation which is studied here is of non linear type and thus difficult to solve, in spite of the apparent simplicity of the estimator of the Fourier Transform of f_ε. See the study of similar questions in the context of inverse problem with error in the operator in Hoffmann and Reiß (2008) or Cavalier and Raimondo (2007). Similar methods also appear in Lévy model: see for instance Kappus (2012) or Gugushvili (2012).

Although there exists a huge literature concerning the estimation of f when f_ε is known, this problem without the knowledge of f_ε has been less studied. One can cite Efromovich (1997) and Johannes and Schwarz (2013) in a context of circular data and Diggle and Hall (1993) who examine the case $M \geqslant n$. Meister (2004) studies the effect of noise misspecification. A few authors have studied the exact problem which is considered in this paper, but only for particular type of smoothness for f_ε or f or other type of risks. Neumann (1997a) gives an upper bound and a lower bound for the integrated risk in the case where both f and f_ε are ordinary smooth, and Johannes (2009) gives upper bounds for the integrated risk in a larger context of regularities.

Estimation procedure We start with the same estimator as previously, with $1/h = \pi m$, defined by

$$\hat{f}^*(t) = \mathbb{1}_{[-\pi m,\pi m]}(t)\frac{\widehat{f_Y^*}(t)}{f_\varepsilon^*(t)}.$$

Here f_ε^* is also unknown and need to be estimate. Therefore, we use the preliminary noise sample and we define the natural estimator of f_ε^*: $\hat{f}_\varepsilon^*(x) = \frac{1}{M}\sum_{j=1}^M e^{-ix\varepsilon_{-j}}$. To avoid the problem of a vanishing denominator, we introduce the truncated estimator:

$$\frac{1}{\tilde{f}_\varepsilon^*(x)} = \frac{\mathbb{1}_{\{|\hat{f}_\varepsilon^*(x)|\geqslant M^{-1/2}\}}}{\hat{f}_\varepsilon^*(x)} = \begin{cases} \dfrac{1}{\hat{f}_\varepsilon^*(x)} & \text{if } |\hat{f}_\varepsilon^*(x)| \geqslant M^{-1/2} \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

Then we can consider

$$\hat{f}_m(x) = \frac{1}{2\pi}\int_{-\pi m}^{\pi m} e^{ixu}\frac{\hat{f}_Y^*(u)}{\tilde{f}_\varepsilon^*(u)}du. \tag{14}$$

We can prove (see [L10]) that this estimator is also a minimum contrast estimator, and it has also another expression given in (17).

Study of the integrated risk Let f_m such that $f_m^* = f^*\mathbb{1}_{[-\pi m,\pi m]} = \mathbb{E}(\hat{f}_Y^*/f_\varepsilon^*\mathbb{1}_{[-\pi m,\pi m]})$. We introduce the notations

$$\Delta(m) = \frac{1}{2\pi}\int_{-\pi m}^{\pi m}\frac{1}{|f_\varepsilon^*(u)|^2}du \text{ and } \Delta_f(m) = \frac{1}{2\pi}\int_{-\pi m}^{\pi m}\frac{|f^*(u)|^2}{|f_\varepsilon^*(u)|^2}du.$$

Then we can prove the existence of a numerical constant C such that:

$$\mathbb{E}(\|\hat{f}_m - f\|^2) \leqslant \|f_m - f\|^2 + C\frac{\Delta(m)}{n} + (C+2)\frac{\Delta_f(m)}{M}. \tag{15}$$

The first term is the classical bias ($= B(h)$ with previous notations). The second term is the variance term for deconvolution problems (with previous notations $V(h) = \Delta(m)/n$). The third term $\Delta_f(m)/M$ is due to the estimation of f_ε^*. As $|f^*(x)| \leqslant 1$, we have $\Delta_f(m) \leqslant \Delta(m)$. It

follows that for any $M \geqslant n$, then $\mathbb{E}\|\hat{f}_m - f\|^2 \leqslant \|f_m - f\|^2 + C\Delta(m)/n$ and we recover the usual risk bound for deconvolution estimation when f_ε^* is known. Therefore, in all cases, the condition $M \geqslant n$ ensures that the rate of the estimator is the same as when f_ε^* was known. Notice that inequality (15) is also true for multivariate deconvolution (see [L12]). However, for the rates of convergence and the following of this section, we restrict to the case of real variables.

As previously we assume a parametric description of the rate of decrease of f_ε^* written as follows:

$$|f_\varepsilon^*(t)| \approx (t^2+1)^{-\beta/2} \exp(-\alpha|t|^\rho).$$

Moreover, the distribution function f to estimate is assumed to belong to the extended Sobolev space $\mathcal{S}(b, a, r, L)$ defined as the class of integrable functions satisfying

$$\int |f^*(t)|^2 (1+t^2)^b \exp(2a|t|^r) dt \leqslant L^2,$$

We recall that when $r > 0$, the function f is called supersmooth, and ordinary smooth otherwise. In the same way, the noise distribution is called ordinary smooth if $\rho = 0$ and supersmooth otherwise.

Under these assumptions the bias and the variance terms can be evaluated. This allows us to obtain the upper bounds for the rate of convergence of the Mean Integrated Squared Risk (for a good choice of m depending on each case):

	noise OS	noise SS
f OS	$n^{-\frac{2b}{2b+2\beta+1}} + M^{-[1\wedge(\frac{b}{\beta})]}$	$(\log n)^{-\frac{2b}{\rho}} + (\log M)^{\frac{-2b}{\rho}}$
f SS	$\dfrac{(\log n)^{\frac{2\beta+1}{r}}}{n} + \dfrac{1}{M}$	see the discussion below.

The last case, when both functions are supersmooth ($r > 0$ and $\rho > 0$), is much more tedious, in particular if one wants to evaluate the rates. Three cases have to be distinguished. When $r = \rho$, these are polynomial rates in n and M. Roughly speaking, when f is more regular than the noise $r > \rho$, the rate is of the form $w_n + \frac{1}{M}$, and is the opposite case $u_n + v_M$, where u, v, w are sequences decreasing faster than logarithm: $\frac{(\log n)^\nabla}{n} < u, v, w < (\log n)^\triangle$.

A similar table of rates for the pointwise risk can be found in [L9].

Concerning lower bounds for the integrated risk, we obtain a partial result. The following proposition establishes the optimality of our estimator with respect of both risks in the cases where f is smoother than f_ε and $r \leqslant 1$.

Proposition 27 ([L10])**.** *Under assumption* **(E)**, *if* $r = \rho = 0$ *and* $\beta < b - 1/2$, *or if* $0 \leqslant \rho < r \leqslant 1$ *then*

$$\inf_{\hat{f}} \sup_{f, f_\varepsilon} \mathbb{E}\|\hat{f} - f\|_2^2 \geqslant CM^{-1}$$

where the infimum is taken over all estimators $\hat{f}$ *of* f *based on the observations* $Y_1, \ldots, Y_n, \varepsilon_{-1} \ldots, \varepsilon_{-M}$.

Adaptation The above study shows that the choice of m is both crucial et difficult. Thus, we provide a data driven strategy to perform automatically this choice. We assume that we are able to manage with M much larger than n: this means that we need a careful calibration step, but this step is done once for all. More precisely, we assume in the following that $M = n^2$. This preliminary ε-sample will enable us to provide a density estimator for any new n-sample of the Y_i's. Consequently, our aim is to preserve here the rate corresponding to the case of an n-sample of observations Y_i when f_ε^* is known.

The estimation procedure is completed as follows. We choose the best estimator among the collection $(\hat{f}_m)_{m\in\mathcal{M}_n}$ where $\mathcal{M}_n \subset \{1,\dots,n\}$ is the set of all considered indexes. To do this, we consider

$$\hat{m} = \arg\min_{m\in\mathcal{M}_n}\{-\|\hat{f}_m\|^2 + \mathrm{pen}(m)\}$$

where pen is a penalty term. A traditional choice for the penalty is of order of the variance, here $V(m) = \Delta(m)/n$. But $\Delta(m)$ needs to be estimated. Then we define

$$\mathrm{pen}(m) = \kappa\left(\frac{\log(\hat{\Delta}(m))}{\log(m+1)}\right)^2 \frac{\hat{\Delta}(m)}{n}, \quad \hat{\Delta}(m) = \int_{-\pi m}^{\pi m} |\tilde{f}_\varepsilon^*(x)|^{-2}dx. \tag{16}$$

Let us also define $\hat{m}_n = \arg\max\left\{m \in \{1,\dots,n\}, 1/4 \leqslant \hat{\Delta}(m)/n \leqslant 1/2\right\}$.

Then the following theorem shows that we have completely solved our problem with a data-driven procedure. Notice that here, the penalty is not only data driven, but the collection of models is also randomly selected on the basis of the observations.

Theorem 28 ([L10])**.** *Assume that assumption* **(E)** *is fulfilled. Consider the estimator* $\tilde{f} = \hat{f}_{\hat{m}}$ *defined by* (14) *and*

$$\hat{m} = \arg\min_{m\in\{1,\dots,\hat{m}_n\}}\{\gamma_n(\hat{f}_m) + \mathrm{pen}(m)\}$$

with $\mathrm{pen}(m)$ *defined by (16),* κ *being a pure numerical constant (*$\kappa = 128$ *would work). Then, for* n *large enough, we have*

$$\mathbb{E}\|\tilde{f} - f\|^2 \leqslant C_1 \inf_{m\in\mathcal{M}_n}\{\|f_m - f\|^2 + \mathrm{pen}(m)\} + \frac{C_2 \log^{\mathbb{1}_{\rho>1}}(n)}{n},$$

where C_1 *is a pure numerical constant (*$C_1 = 40$ *would suit), and* C_2 *is a constant depending on* f *and* f_ε*.*

The result of Theorem 28 is an oracle inequality which states that the estimator $\hat{f}_{\hat{m}}$ makes the compromise between the squared bias term $\|f - f_m\|^2$ and the penalty, except for a possible $\log(n)$ factor (which appears only if $\rho > 1$, in which case it is negligible). If the penalty has exactly the order of the variance $\Delta(m)/m$, then the optimal rate is reached. Otherwise, a loss may occur, since we have a multiplicative factor $m^{2\rho}/\log^2(m+1)$. In the cases $\rho = 0$ (ordinary smooth error) or $r = 0, \rho > 0$ or $0 < r < \rho$, we can easily prove that the rate of convergence of the estimate is not affected. If $r \geqslant \rho > 0$, the loss in the rate concerns only the logarithmic terms, which are negligible with respect to the rate. Therefore, if a loss in the rate occurs, as price of the adaptive property of the procedure, we know that it is negligible with respect to the rate of convergence of the estimator (it follows from results in Butucea and Tsybakov (2007) that a loss may be unavoidable in the adaptive procedure; in that case, the rate is called adaptive optimal).

Let us emphasize here that the interest of the penalty (16) is that the terms required in the supersmooth case are added without requiring the information: are the errors ordinary smooth or supersmooth, and what is the value of ρ. Nevertheless, this procedure has been improved by the very recent work (posterior to this one) of Kappus and Mabon (2014). Their considerations on a very different penalty term allow to handle the case of small noise samples $M < n$ (with nevertheless a loss of logarithmic order), and to dispose completely of our semi-parametric assumptions on the noise smoothness, which makes more sense in model selection procedure leading to non-asymptotic oracle inequalities.

Numerical illustration In practice we use the following expression of our estimator:

$$\hat{f}_m = \sum_{l\in\mathbb{Z}} \hat{a}_{m,l}\varphi_{m,l} \quad \text{with} \quad \hat{a}_{m,l} = \frac{1}{n}\sum_{j=1}^{n} \tilde{v}_{\varphi_{m,l}}(Y_j) \tag{17}$$

where $\varphi_{m,j}(x) = \sqrt{m}\varphi(mx-j)$, $\varphi(x) = \sin(\pi x)/(\pi x)$, and $\tilde{v}_t^*(u) = t^*(u)/\tilde{f}_\varepsilon^*(-u)$. Since the coefficients $\hat{a}_{m,l}$ can be computed using the Inverse Fast Fourier Transform, that clearly makes the procedure fast.

Let us first compare our estimator to the one of Neumann (1997a). He denotes by $f_0(x) = e^{-|x|}/2$ and he considers two examples:
- example 1: $f = f_0 \star f_0 \star f_0 \star f_0$ and $f_\varepsilon = f_0 \star f_0$
- example 2: $f = f_0 \star f_0$ and $f_\varepsilon = f_0 \star f_0 \star f_0 \star f_0$

We set, as in Neumann (1997a), $n = 200$ and $M = 10$ and the $\mathbb{L}^2$ risk is computed with 100 random samples. In these examples, the signal and the noise are ordinary smooth ($r = \rho = 0$): this induces the rates of convergence $n^{-\frac{15}{24}} + M^{-1}$ and $n^{-\frac{7}{24}} + M^{-\frac{7}{16}}$ for examples 1 and 2 respectively. We also compute the estimator with known noise, replacing $\tilde{f}_\varepsilon$ by f_ε in the procedure. The integrated $\mathbb{L}^2$ risks for 100 replications are given below in Table 1 and show our improvement of the results of Neumann (1997a).

	ex 1	ex 2
f_ε known	0.00257	0.01904
f_ε unknown	0.00828	0.06592

	ex 1	ex 2
f_ε known	0.00243	0.01791
f_ε unknown	0.00612	0.03427

Table 1: MISE for the estimators of Neumann (1997a) (left) and for the penalized estimator (right).

An example of estimation for supersmooth functions is given in Johannes (2009). In his example 5.1, he considers a standard Gaussian noise and $X \sim \mathcal{N}(5, 9)$. In this case $r = 2$, $b = 1/2$ and $\rho = 2$, $\beta = 0$ and the rate of convergence is $n^{-\frac{9}{10}}(\log n)^{-1/2} + M^{-1}$. The improvement brought by our method is striking.

In [L10] different signal densities and different noises are also considered. We notice that the estimation of the characteristic function of the noise does not spoil so much the procedure. It even happens that the estimation with unknown noise works better, which is likely due to the truncation (13).

Figure 12 illustrates these results for two cases: a mixed Gamma density estimated through Laplace noise and a Laplace density estimated through Gaussian noise. The curves for $M = 5$ show that our method is still very satisfactory for small values of M.

	$n=100$	$n=250$	$n=500$
f_ε known	2.0	0.9	0.6
$M=100$	2.0	1.0	0.7
$M=250$	1.9	1.0	0.6
$M=500$	1.9	0.9	0.6

	$n=100$	$n=250$	$n=500$
f_ε known	0.71	0.23	0.12
$M=100$	0.24	0.11	0.07
$M=250$	0.21	0.12	0.07
$M=500$	0.21	0.12	0.07

Table 2: Third quartile of the MISE $\times 100$ for the estimators of Johannes (2009) (left) and for the penalized estimator (right).

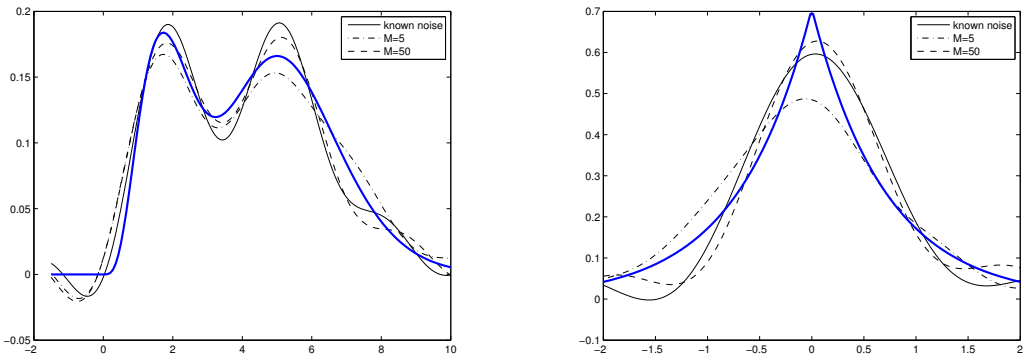


Figure 12: True function f (bold line) and estimators for $n=500$. Left: mixed Gamma density with Laplace noise. Right : Laplace density with Gaussian noise.

3.1.4 Estimation for pure-jump Lévy processes

Consider $(L_t, t \geqslant 0)$ a real-valued Lévy process with characteristic function given by:

$$\psi_t(u) = \mathbb{E}(\exp iuL_t) = \exp\left(t \int_{\mathbb{R}} (e^{iux} - 1)N(x)dx\right). \tag{18}$$

We assume that the Lévy measure admits a density N and that the function $g(x) = xN(x)$ is integrable. Under these assumptions, $(L_t, t \geqslant 0)$ is a pure jump Lévy process without drift and with finite variation on compact sets. Suppose that we have discrete observations $(L_{k\Delta}, k = 1, ..., n)$ with sampling interval Δ. Our aim here is the nonparametric adaptive kernel estimation of the function $xN(x)$ based on these observations under the asymptotic framework $n \to \infty$ and $\Delta \to 0$.

Estimation for Lévy processes is very linked to convolution model, due to the use of characteristic functions. By the way, papers in this domain quite often refer to deconvolution literature. Indeed, as we just see, deconvolution study is based on the equality $f^* = f_Y^*/f_\varepsilon^*$. Here in our context of Lévy process $g^* = (-i/\Delta)(f_{L_\Delta}^*)'/f_{L_\Delta}^*$ where $g(x) = xN(x)$. We have then to estimate the right member in order to reconstruct the Fourier transform of the target. In general, since both numerator and denominator have to be estimated, methods are similar to deconvolution with unknown error. Here, we only consider the high frequency context, then $f_{L_\Delta}^* = \psi_\Delta \approx 1$ and only the numerator needs to be estimate.

This subject has been recently investigated by several authors. Figueroa-López and Houdré (2006) use a penalized projection method to estimate the Lévy density on a compact set separated from 0. Other authors develop an estimation procedure based on empirical estimations of the characteristic function $\psi_\Delta(u)$ of the increments $(Z_k^\Delta = L_{k\Delta} - L_{(k-1)\Delta}, k = 1, \ldots, n)$ and its derivatives followed by a Fourier inversion to recover the Lévy density. For low frequency data

(Δ is fixed), we can quote Watteel and Kulperger (2003), or Jongbloed and van der Meulen (2006) for a parametric study. Still in the low frequency framework, Neumann and Reiß (2009) estimate $\nu(x) = x^2 N(x)$ in the more general case with drift and volatility, and Comte and Genon-Catalot (2010) use model selection to build an adaptive estimator. An adaptive method to estimate linear functionals is also given in Kappus (2012). Belomestny (2011) addresses the issue of inference for time-changed Lévy processes with results in term of uniform and pointwise distance. One can also cite Gugushvili (2012) or Nickl and Reiß (2012) for recent works at fixed Δ. Here we introduce a kernel estimator with local bandwidth selection. Note that a pointwise study involving a kernel estimator can be found in van Es et al. (2007) for more specific compound Poisson processes, but the estimator is different from ours, as well as the observation scheme. In Figueroa-López (2011) a pointwise central limit theorem is given for the estimation of the Lévy density, as well as confidence intervals. Still in the high frequency context, but for integrated distance, we can cite Ueltzhöfer and Klüppelberg (2011), and Duval (2012) for the estimation of a compound Poisson process with low conditions on Δ. Bücher and Vetter (2013) deal with the multivariate case.

Here we shall study local adaptive bandwidth selection (which the previous authors do not consider). For a given non-zero real x_0, we select a bandwidth $\hat{h}(x_0)$ using Goldenshluger-Lepski method, such that the resulting adaptive estimator $\hat{g}_{\hat{h}(x_0)}(x_0)$ automatically reaches the optimal rate of convergence corresponding to the unknown regularity of the function g. The advantage of our kernel method is that it allows us to estimate the Lévy density at a given point, with a local adaptive choice.

Estimator We start from the equality:

$$\mathbb{E}\left[Z_k^\Delta e^{iuZ_k^\Delta}\right] = -i\psi'_\Delta(u) = \Delta\psi_\Delta(u)g^*(u), \tag{19}$$

obtained by differentiating (18). Here $g^*(u) = \int e^{iux}g(x)dx$ is the Fourier transform of g, well defined since we assume g integrable. Then, as $\psi_\Delta(u) \simeq 1$, equation (19) writes $\mathbb{E}\left[Z_k^\Delta e^{iuZ_k^\Delta}\right] \simeq \Delta g^*(u)$. This gives an estimator of $g^*(u)$ as follows:

$$\frac{1}{n\Delta}\sum_{k=1}^n Z_k^\Delta e^{iuZ_k^\Delta}.$$

Now, to recover g, we introduce a kernel to make inversion possible:

$$\frac{1}{n\Delta}\sum_{k=1}^n Z_k^\Delta K^*(uh)e^{iuZ_k^\Delta}$$

which is in fact the Fourier transform of $1/(nh\Delta)\sum_{k=1}^n Z_k^\Delta K((x - Z_k^\Delta)/h)$. At the end, in the high frequency context, a direct method without Fourier inversion can be applied. Indeed, a consequence of (19) is that the empirical measure:

$$\hat{\mu}_n(dz) = \frac{1}{n\Delta}\sum_{k=1}^n Z_k^\Delta \delta_{Z_k^\Delta}(dz)$$

weakly converges to $g(z)dz$ (note that the idea of exploiting this weak convergence is already present in Figueroa-López (2009b)). This suggests to consider kernel estimators of g of the form

$$\hat{g}_h(x) = K_h \star \hat{\mu}_n(x) = \frac{1}{n\Delta}\sum_{k=1}^n Z_k^\Delta K_h(x - Z_k^\Delta)$$

where $K_h(x) = (1/h)K(x/h)$ and K is a kernel such that $\int K = 1$. Below, we study the quadratic pointwise risk of the estimators $\hat{g}_h(x)$ and evaluate the rate of convergence of this risk as n tends to infinity, $\Delta = \Delta(n)$ tends to 0 and $h = h(n)$ tends to 0. This is done under Hölder regularity assumptions for the function g.

Pointwise risk Let us now define the assumptions concerning the target function g, defined by $g(x) = xN(x)$, where N is the Lévy density. We shall assume that g belongs to the Hölder class $H(\beta, L)$ i.e.

$$|g^{(l)}(x) - g^{(l)}(y)| \leqslant L|x-y|^{\beta-l}, \quad \forall x, y \in \mathbb{R}.$$

where $l = \lfloor \beta \rfloor$ is the largest integer strictly smaller than β. Moreover we define

Assumption G(p) $g \in \mathbb{L}^2$, g^* is differentiable almost everywhere and its derivative belongs to $\mathbb{L}^1$, g' exists and is uniformly bounded. Moreover, for p integer, $\int |x|^{p-1}|g(x)|dx < \infty$.

This assumption ensures that $\mathbb{E}|Z_1^\Delta|^p < \infty$. It also implies $g \in H(1, L')$ so we can assume $\beta \geqslant 1$.

Now let us describe which kind of kernel we choose for our estimator. Let us define the following condition

K(β): K belongs to $\mathbb{L}^1 \cap \mathbb{L}^2 \cap \mathbb{L}^\infty$ and $K^* \in \mathbb{L}^1$. Moreover the kernel K is of order $\lfloor \beta \rfloor$ (i.e. $\int K(u)du = 1, \quad \int u^j K(u)du = 0, \quad j \in \{1, ..., \lfloor \beta \rfloor\}$) and $\int |x|^\beta |K(x)|dx < +\infty$.

These assumptions are standard when working on problems of estimation by kernel methods. Note that there is a way to build a kernel of order l (see Kerkyacharian et al. (2001) and Goldenshluger and Lepski (2011)).

In all the following $x_0 \in \mathbb{R}$ is fixed, with $x_0 \neq 0$. The usual bias variance decomposition of the Mean Squared Error yields:

$$\mathbb{E}[(\widehat{g}_h(x_0) - g(x_0))^2] = \mathbb{E}[(\widehat{g}_h(x_0) - \mathbb{E}[\widehat{g}_h(x_0)])^2] + (\mathbb{E}[\widehat{g}_h(x_0)] - g(x_0))^2.$$

We can prove that the variance term is of order $1/(nh\Delta)$. The bias needs further decomposition: $\mathbb{E}[\widehat{g}(x_0)] - g(x_0) = b_1(x_0) + b_2(x_0)$ with the usual bias,

$$b_1(x_0) = K_h \star g(x_0) - g(x_0),$$

bounded by h^β, and the bias resulting from the approximation of $\psi_\Delta(u)$ by 1,

$$b_2(x_0) = \mathbb{E}[\widehat{g}(x_0)] - K_h \star g(x_0).$$

bounded by Δ. Then, under **G(2)** and if K satisfies **K(β)**, we have

$$\mathbb{E}[(\widehat{g}_h(x_0) - g(x_0))^2] \leqslant c_1 h^{2\beta} + c_2 \frac{1}{nh\Delta} + c'_1 \Delta^2, \tag{20}$$

with c_1, c'_1 depending on g, K and $c_2 = (2\pi)^{-1}\|K\|_2^2 \left(\|(g^*)'\|_1 + \|g^*\|_2^2\right)$. For the two first terms the optimal choice of h is $h_{opt} \propto (n\Delta)^{-\frac{1}{2\beta+1}}$ and the associated rate has classical order $O\left((n\Delta)^{-\frac{2\beta}{2\beta+1}}\right)$. Next, a sufficient condition for $\Delta^2 \leqslant (n\Delta)^{-\frac{2\beta}{2\beta+1}}$ for all β is

$$\Delta = O(n^{-1/3}). \tag{21}$$

Under this condition, choosing $h_{opt} \propto (n\Delta)^{-\frac{1}{2\beta+1}}$ gives $\mathbb{E}[(\widehat{g}_{h_{opt}}(x_0) - g(x_0))^2] = O\left((n\Delta)^{-\frac{2\beta}{2\beta+1}}\right)$ and as a consequence $\mathbb{E}[(\widehat{g}_{h_{opt}}(x_0)/x_0 - N(x_0))^2] = O\left((n\Delta)^{-\frac{2\beta}{2\beta+1}}\right)$.

This rate turns out to be the optimal minimax rate of convergence over the class $H(\beta, L)$. This result is proved in Figueroa-López (2009a) in the more general case of estimators based on the whole path of the process up to time $n\Delta$. In our case of discrete sampling, we give another proof in [L14].

Adaptation As β is unknown, we need a data-driven selection of the bandwidth. We introduce a set of bandwidth of the form $\mathcal{H} = \{\frac{j}{M}, 1 \leqslant j \leqslant M\}$ with M an integer to be specified later. Actually it is sufficient to control $\sum_{h\in\mathcal{H}} h^{-w}$ for some w so that more general set of bandwidths are possible. We set:

$$V(h) = C_0 \frac{\log(n\Delta)}{nh\Delta}$$

with C_0 to be specified later. Note that $V(h)$ has the same order as the variance multiplied by $\log(n\Delta)$. We also define $\hat{g}_{h,h'}(x_0) = K_{h'} \star \hat{g}_h(x_0) = K_h \star \hat{g}_{h'}(x_0)$. Lastly we set, as an estimator of the bias,

$$A(h, x_0) = \sup_{h'\in\mathcal{H}} \left[|\hat{g}_{h,h'}(x_0) - \hat{g}_{h'}(x_0)|^2 - V(h')\right]_+ .$$

Then, the adaptive bandwidth h is chosen as follows:

$$\hat{h} = \hat{h}(x_0) \in \operatorname*{argmin}_{h\in\mathcal{H}} \{A(h, x_0) + V(h)\}.$$

In light of c_2 in (20), a good choice for C_0 is

$$C_0 = \frac{c}{2\pi}\|K\|_2^2 \left(\|(g^*)'\|_1 + \|g^*\|_2^2\right).$$

However, $\|(g^*)'\|_1$ and $\|g^*\|_2^2$ are unknown, then these quantities have to be estimated with a preliminar estimator of g^*. More precisely, we set $K_0^* = \mathbb{1}_{[-1,1]}$ and

$$\widehat{\|(g^*)'\|}_1 = \int \left|\frac{1}{n\Delta}\sum_{k=1}^{n} (Z_k^\Delta)^2 K_0^*(uh_1) e^{iuZ_k^\Delta}\right| du \quad \text{with } h_1 = (n\Delta)^{-1/3},$$

$$\widehat{\|g^*\|_2^2} = \|\hat{g}_{h_2}^*\|_2^2 = \int \left|\frac{1}{n\Delta}\sum_{k=1}^{n} Z_k^\Delta K_0^*(uh_2) e^{iuZ_k^\Delta}\right|^2 du \quad \text{with } h_2 = (n\Delta)^{-1/3}.$$

We introduce the following smoothness condition: a function ψ belongs to the Sobolev space $\mathcal{S}(1)$ if $\int |\psi^*(u)|^2(1+u)^2 du < \infty$ (this means that ψ has a derivative which is square-integrable).

Before to study the performance of our final estimator $\hat{g}_{\hat{h}}(x_0)$, let us clarify the observation context. We still work in the high frequency framework, and we have seen that we need condition (21). Thus, the assumption on the observation step is the following

S $\Delta \to 0$ and $n\Delta \to \infty$. Moreover $\Delta \leqslant 1$ and $\Delta = O(n^{-1/3})$

Theorem 29 ([L14])**.** *We use a kernel satisfying* $\mathbf{K(1)}$ *and* $M = O((n\Delta)^{1/3})$*. Assume that* g *satisfies* $\mathbf{G(32)}$ *and that* g *and* $xg(x)$ *belong to* $\mathcal{S}(1)$*. In the definition of* $\hat{h}$*, replace* $V(h)$ *by* $\hat{V}(h) = \widehat{C_0}\log(n\Delta)/(nh\Delta)$ *where*

$$\widehat{C_0} = \frac{c}{2\pi}\|K\|^2 \left(\widehat{\|(g^*)'\|}_1 + \widehat{\|g^*\|_2^2}\right)$$

with $c \geqslant 32\max(1, \|K\|_\infty)$. *Then, under scheme* **S**,

$$\mathbb{E}[|g(x_0) - \hat{g}_{\hat{h}}(x_0)|^2] \leqslant C \left\{ \inf_{h\in\mathcal{H}} \left\{ \operatorname{ess\,sup} |g - \mathbb{E}[\hat{g}_h]|^2 + V(h) \right\} + \frac{\log(n\Delta)}{n\Delta} \right\}.$$

Thus our estimator $\hat{g}_{\hat{h}}$ has a risk as good as any of the collection $(\hat{g}_h)_{h\in\mathcal{H}}$, up to a logarithmic term. The pointwise control of the bias has been replaced with a uniform control. Actually, it is possible to keep the pointwise risk in the right term at the cost of a supplementary term $\sup_{h'\in\mathcal{H}} |K_{h'} \star b_h(x_0)|^2$. Although our estimator is not linear (we have an extra bias), it is exactly the same situation as in Goldenshluger and Lepski (2013), and we can conjecture it is in some sense unavoidable.

Note that the theorem is valid for c large enough, say $c \geqslant c_0$. In the proof, we obtain the upper bound $32\max(1, \|K\|_\infty)$ for c_0, unfortunately we can conjecture that this bound is not the optimal one. To obtain a sharper bound we have tuned c_0 in the simulation study.

Let us now conclude with the consequence of this theorem in term of rate of convergence. As already explained, as we need assumption $\mathbf{G}(p)$ to control the bias, we can assume $\beta \geqslant 1$. Then

$$h_{opt} \propto (\log(n\Delta)/n\Delta)^{1/(2\beta+1)} \geqslant (n\Delta)^{-1/3}$$

belongs to $\mathcal{H}$ as soon as M is larger than a constant times $(n\Delta)^{1/3}$. Hence we can state the following corollary.

Corollary 30. *Assume that* g *belongs to* $H(\beta, L)$ *with* $\beta \geqslant 1$. *We use a kernel satisfying* $\mathbf{K}(\beta)$ *and* $M = \lfloor (n\Delta)^{1/3} \rfloor$. *Take* C_0 *as in Theorem 29 with assumptions of this latter theorem. Then, under scheme* **S**,

$$\mathbb{E}[|g(x_0) - \hat{g}_{\hat{h}}(x_0)|^2] = O\left((\log(n\Delta)/n\Delta)^{-\frac{2\beta}{2\beta+1}} \right).$$

Then the price to pay to adaptivity is a logarithmic loss in the rate. Nevertheless this phenomenon is known to be unavoidable in pointwise estimation (see Butucea (2001)). Thus $\hat{g}_{\hat{h}}(x_0)$ (resp. $\hat{g}_{\hat{h}}(x_0)/x_0$) is an adaptive estimator for $g(x_0)$ (resp. $N(x_0)$).

The numerical performance of our method can be observed in [L14] where we give simulations for various examples of Lévy processes. We also detail the extension of our procedure to irregular sampling, i.e. to the case where the interval Δ is not necessarily the same at each time.

3.1.5 Some prospects in deconvolution

- There are still a lot of problems where the issue of noisy data has not been completely solved: estimation of the intensity of a Poisson process, adaptation in the problem of uniform deconvolution, Berkson model, etc.. An issue (suggested by Elizabeth Gassiat) which I find particularly interesting involves the semiparametric model of deconvolution, with unknown noise variance. By assuming a dependence in the signal, e.g. X is a Markov chain, one could ensure identifiability and easier estimation than in the independent case.

- Previous works have considered that the coordinate axes were preferential directions, but this is not necessarily the case. Imagine a Gaussian noise with covariance, or a signal with particular geometry. Then a convenient kernel has the form $K_H(x) = (\det H)^{-1} K(H^{-1}x)$ where H is an invertible matrix. More generally, there is a lot to do in geometrical inference, in line with Dedecker et al. (2015). In particular, assume that one observe a dynamical system on a manifold $X_{n+1} = f(X_n) + \varepsilon_{n+1}$. The question is then to estimate f and to find properties of the manifold.

- While there is now a huge literature about estimation for Lévy process, similar considerations for Lévy fields are very rare. However it is worth considering applications to spatial statistics.

3.2 Goodness-of-fit test for spherical data

3.2.1 Model and motivation

Convolution model has been studied in other frameworks than $\mathbb{R}^d$, like the hyperbolic plane (Huckemann et al., 2010), or compact Lie groups (Kim and Richards, 2001). Here we present the case of spherical data, which finds a natural motivation in astrophysics and geography. The model is then:

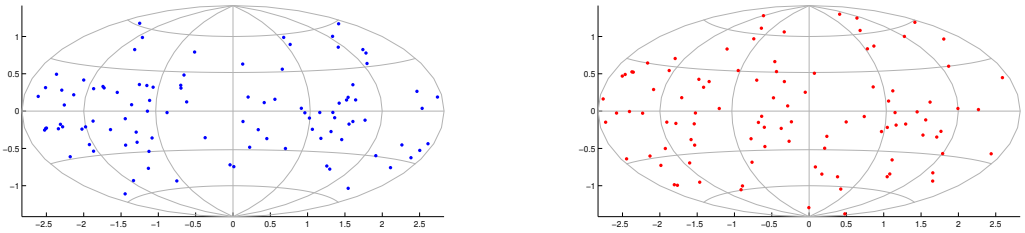
$$Z_i = \varepsilon_i X_i, \quad i = 1, \dots, N \tag{22}$$

where the ε_i are i.i.d. random variables of SO(3) the rotation group in $\mathbb{R}^3$ and the X_i's are i.i.d. random variables on $\mathbb{S}^2$, the unit sphere in $\mathbb{R}^3$. We suppose that X_i and ε_i are independent. We also assume that the distributions of Z_i and X_i are absolutely continuous with respect to the uniform measure on $\mathbb{S}^2$ and we set f_Z and f the densities of Z_i and X_i respectively. The distribution of ε_i is absolutely continuous with respect to the probability Haar measure on SO(3) and we will denote it f_ε. Then we have

$$f_Z = f_\varepsilon \star f := \int_{\mathrm{SO}(3)} f_\varepsilon(u) f(u^{-1}\omega) du$$

where $\star$ denotes the convolution product.

The figure below gives an example of a 100-sample on the sphere : X left, Z right (to visualize points on the sphere, we use Hammer projection, because of its equal-area property).



Here, instead of estimating f, we want to provide a nonparametric adaptive minimax goodness-of-fit testing procedure on f from the noisy observations Z_i. More precisely, let f_0 being the uniform density on $\mathbb{S}^2$, we consider the problem of testing the null hypothesis $f = f_0$ with alternatives expressed in $\mathbb{L}^2$ norm over Sobolev classes.

Spherical data arise in many areas of scientific experimentation and observation. As examples of directional data from various fields, we instance in astrophysics the arrival directions of the Ultra High Energy Cosmic rays (UHECR), from structural geology the facing directions of conically folded planes, from paleomagnetism the measurements of magnetic remanence in rocks, from meteorology the observed wind directions at a given place and from physical oceanography the measurements of current ocean directions. In this dissertation, we will particularly focus on the UHECR study as application of our statistic procedure.

In astrophysics, a burning issue consists in understanding the behaviour of the so-called UHECR. These latter are cosmic rays with an extreme kinetic energy (of the order of 10^{19} eV) and the rarest particles in the universe. The source of those most energetic particles remains a mystery and the stake lies in finding out their origins and which process produces them. Astrophysicists have at their disposal directional data which are measurements of the incoming directions of the UHECR on Earth. Needless to say that finding out more about the law of probability of those incoming directions is crucial to gain an insight into the mechanisms generating the UHECR. Faÿ

et al. (2012) recently developed isotropy goodness-of-fit tests based on the so-called needlets for the non perturbated case. Their study is focused on the practical aspect with nice simulations connected to realistic cosmic rays scenarios. But the difficulty lies in the fact that the observed UHECR do not come necessarily from the genuine direction as specified by Faÿ et al. (2012). Their trajectories are deflected by Galactic and intergalactic fields. As this deflection is inevitable in the measurements, it is quite challenging and essential to take into account this uncertainty in the statistical modelling. A first way to model the deflection in the incoming directions can be done thanks to the model (22) with random rotations. Concerning the hypotheses about the underlying probability of the incoming directions, several are made. A uniform density would suggest that the UHECR are generated by cosmological effects, such as the decay of relic particles from the Big Bang. On the contrary, if these UHECR are generated by astrophysical phenomena (such as acceleration into Active Galactic Nuclei (AGN)), then we should observe a density function which is highly non-uniform et tightly correlated with the the local distribution of extragalactic supermassive black holes at the center of nearby galaxies (AGN). First results seemed to favour a non-uniform density but as underlined by Faÿ et al. (2012), a more recent analysis based on 69 observations of UHECR softens this conclusion of anisotropy. To this prospect, these relevant considerations lead naturally to goodness-of-fit testing on the uniform density in the noisy model (22).

Considering goodness-of-fit testing in the spherical convolution model not only finds its interest in the above important applications, but it also fills a gap both in the noisy setup testing literature and the spherical convolution one. So far, only estimation has been treated in the spherical setup. For the nonparametric estimation problem, one is interested in recovering the underlying density f from noisy observations Z_i. The pioneer works of Healy et al. (1998), Kim and Koo (2002), Kim et al. (2004) introduced a minimax estimation procedure based on the Fourier basis of $\mathbb{L}^2(\mathbb{S}^2)$. Recently, Kerkyacharian et al. (2011) proposed an optimal and adaptive hard thresholding estimation procedure based on needlets.

Nonparametric goodness-of-fit testing has aroused a lot of interest. For minimax testing, we refer to the work of Ingster (1993) which is the main reference in the field. Spokoiny (1996) first established adaptive testing procedure based on wavelets over Besov bodies. Nonetheless, goodness-of-fit testing has mainly focused on the case of direct observations. Indeed, very few works have been devoted to the case of indirect observations. Let us cite the works of Bissantz et al. (2009) for the inverse regression problem and Holzmann et al. (2007) for the multivariate convolution density model. Butucea (2007) built minimax nonparametric goodness-of-fit testing for convolution models based on kernels methods and Butucea et al. (2009) made a step forward by building an adaptive testing procedure in the noisy setup.

For the uniform density of probability on the sphere namely $f_0 = (4\pi)^{-1}\mathbb{1}_{\mathbb{S}^2}$, we want to test the hypothesis

$$H_0 : \quad f = f_0,$$

from observations $Z_1, \ldots, Z_N$ given by model (22). We consider the alternative

$$H_1(s, R, \mathcal{C}\psi_N) : f \in W_s(\mathbb{S}^2, R) \text{ et } \|f - f_0\|_2^2 \geqslant \mathcal{C}\psi_N$$

where $\mathcal{C}$ is a constant and ψ_N is the testing rate. The Sobolev space $W_s(\mathbb{S}^2, R)$ is defined below.

We would also like to bring to the reader's attention some interesting facts when encountering testing problems with indirect observations. Indeed, there is a natural connection between the following approaches : to test $f = f_0$ or to test $f_\varepsilon * f = f_\varepsilon * f_0$. This question has been the object of the recent works of Laurent et al. (2011); Loubes and Marteau (2014) and has been previously evoked by Butucea et al. (2009). In the case of the convolution model on the real

line, Laurent, Loubes and Marteau prove that if a test procedure is minimax for testing problem : $H_0^D : f_\varepsilon * f = f_\varepsilon * f_0$ versus $H_1^D : f_\varepsilon * (f - f_0) \in \mathcal{F}_D$ where

$$\mathcal{F}_D = \{g \text{ with smoothness } s' \text{ and } \|g\|^2 \geqslant C' n^{-4s'/(4s'+1)}, \text{ with } s' = s + \nu\},$$

then it is minimax for $H_0^I : f = f_0$ versus $H_1^I : f - f_0 \in \mathcal{F}_I$ where

$$\mathcal{F}_I = \{f \text{ with smoothness } s \text{ and } \|f\|^2 \geqslant C n^{-4s/(4s+4\nu+1)}\}$$

but the reverse is not true (here n is the number of data and ν the smoothness index of the noise). This interesting conclusion (that we can conjecture true in our context also) does not make it any the less necessary to study the inverse problem here. Indeed, until the present work, the minimax rates were not known in the context of noisy spherical data. Moreover, when dealing with adaptive procedures, the link between the direct and inverse problems is not established yet.

3.2.2 Fourier analysis on the sphere

Let us provide here some elements of Fourier analysis on SO(3) and $\mathbb{S}^2$. For a square integrable function on $\mathbb{S}^2$, we can write

$$f(\omega) = \sum_{l \geqslant 0} \sum_{m=-l}^{l} f_m^{\star l} Y_m^l(\omega),$$

where (Y_m^l) is the spherical harmonic basis, and $f_m^{\star l} = \int_{\mathbb{S}^2} f(x)\overline{Y_m^l(x)}dx$ is the spherical Fourier transform on $\mathbb{S}^2$, considered at each level l as a $(2l+1)$ vector. Then it is sufficient to estimate the Fourier coefficients $f_m^{\star l}$ to retrieve f. In the same way, the eigenfunctions of the Laplace Beltrami operator on SO(3) lead to an orthonormal basis on $\mathbb{L}^2(\mathrm{SO}(3))$: $(\sqrt{2l+1}D_{mn}^l,\ -l \leqslant m,\ n \leqslant l,\ l = 0,\ 1, \ldots)$. Then, for a function g in $\mathbb{L}^2(\mathrm{SO}(3))$ (with respect to the Haar measure),

$$g(u) = \sum_{l \geqslant 0} \sum_{-l \leqslant m,n \leqslant l} g_{mn}^{\star l}(2l+1)\overline{D_{mn}^l(u)},$$

where $g_{mn}^{\star l} = \int_{\mathrm{SO}(3)} g(u) D_{mn}^l(u)du$, is the Fourier transform, considered at each level l as a $(2l+1) \times (2l+1)$ matrix. Moreover, the Fourier coefficients of a convolution product are described by the following matrix product: for all $-l \leqslant m \leqslant l$, $l = 0,\ 1, \ldots$,

$$(f_\varepsilon \star f)_m^{\star l} = \sum_{n=-l}^{l} (f_\varepsilon^{\star l})_{mn} f_n^{\star l} = (f_\varepsilon^{\star l} f^{\star l})_m.$$

Observing this formula, we see that it is sufficient to estimate $(f_Z)_m^{\star l}$ from $Z_1, \ldots, Z_N$ and to inverse the matrices $f_\varepsilon^{\star l}$. We shall assume here that these matrices are invertible. Moreover, denoting $\|A\|_{op} = \sup_{h \neq 0} \|Ah\|_2 / \|h\|_2$, we will say that the distribution of ε is

ordinary smooth of order ν if: $\quad \forall l \geqslant 0 \qquad \|(f_\varepsilon^{\star l})^{-1}\|_{op} \lesssim l^\nu \quad$ and $\quad \|f_\varepsilon^{\star l}\|_{op} \lesssim l^{-\nu}$,

supersmooth of order β if: $\quad \forall l \geqslant 0 \qquad \|(f_\varepsilon^{\star l})^{-1}\|_{op} \lesssim l^{-\nu_0} \exp(l^\beta/\delta) \quad$ and $\quad \|f_\varepsilon^{\star l}\|_{op} \lesssim l^{\nu_1} \exp(-l^\beta/\delta)$.

Recall that we assume that f_ε is known, consequently ν or β are also considered known. Let us give two examples of noise distribution. The rotational Laplace distribution is the rotational analogue of the well-known Euclidean Laplace distribution (known also as double exponential distribution). Its expanded form in terms of rotational harmonics is the following $f_\varepsilon = \sum_{l\geqslant 0}\sum_{m=-l}^{l}(1+\sigma^2 l(l+1))^{-1}(2l+1)\overline{D^l_{mm}}$, for some $\sigma^2>0$ which is a variance parameter. Hence we have

$$(f_\varepsilon^{\star l})_{mn} = (1+\sigma^2 l(l+1))^{-1}\delta_{mn},$$

for $l=0,1,\dots$ and where $\delta_{mn}=1$ if $m=n$ and is 0 otherwise. The Laplace distribution is ordinary smooth with a smoothness index $\nu=2$. Let us present now the Gaussian distribution. The distribution can be written as follows

$$f_\varepsilon = \sum_{l\geqslant 0}\sum_{m=-l}^{l}\exp(-\sigma^2 l(l+1)/2)(2l+1)\overline{D^l_{mm}},$$

for $\sigma>0$. This is an example of a supersmooth distribution with $\delta = 2/\sigma^2$ and $\beta=2$.

Let us now precise what are the regularity assumptions on the signal. For some fixed constant $R>0$, let $W_s(\mathbb{S}^2,R)$ denote the smoothness class of densities f which satisfy

$$\|f\|^2_{W_s} := \sum_{l\geqslant 0}\sum_{m=-l}^{l}(1+l(l+1))^s|f_m^{\star l}|^2 \leqslant \frac{1}{4\pi}+R^2.$$

3.2.3 Test procedure

For the uniform density of probability on the sphere namely $f_0 = (4\pi)^{-1}\mathbb{1}_{\mathbb{S}^2}$, we want to test the hypothesis

$$H_0: \quad f = f_0,$$

from observations $Z_1,\dots,Z_N$ given by model (22). We consider the alternative

$$H_1(s,R,\mathcal{C}\psi_N): f\in W_s(\mathbb{S}^2,R) \text{ and } \|f-f_0\|_2^2 \geqslant \mathcal{C}\psi_N$$

where $\mathcal{C}$ is a constant and ψ_N is the testing rate.

In order to build a test statistic, as usual, we first have to construct an unbiased estimator of the quadratic functional $\int_{\mathbb{S}^2}(f-f_0)^2 = \|f-f_0\|_2^2$. To do so, we remark that thanks to Parseval equality:

$$\int_{\mathbb{S}^2}(f-f_0)^2 = \sum_{l\geqslant 0}\sum_{m=-l}^{l}|f_m^{\star l}-f_{0m}^{\star l}|^2 = \sum_{l\geqslant 1}\sum_{m=-l}^{l}|f_m^{\star l}|^2,$$

the last equality coming from the fact that $(f_0)_m^{\star l}\neq 0$ only for $(l,m)=(0,0)$. From now on we denote $f^{\star l}_{\varepsilon^{-1}} = (f_\varepsilon^{\star l})^{-1}$. Since $f^{\star l} = f^{\star l}_{\varepsilon^{-1}}f_Z^{\star l}$ for $l=0,1,\dots$, we can write under our assumptions

$$f_m^{\star l} = \sum_{n=-l}^{l}(f^{\star l}_{\varepsilon^{-1}})_{mn}(f_Z^{\star l})_n.$$

A natural estimator of $f_m^{\star l}$ is given by

$$\hat{f}_m^{\star l} = \frac{1}{N}\sum_{i=1}^{N}\sum_{n=-l}^{l}(f^{\star l}_{\varepsilon^{-1}})_{mn}\overline{Y_n^l(Z_i)}.$$

If we denote by $\Phi_{lm}(x) = \sum_{n=-l}^{l} (f_{\varepsilon^{-1}}^{\star l})_{mn} \overline{Y_n^l}(x)$ then $\hat{f}_m^{\star l} = \frac{1}{N} \sum_{i=1}^{N} \Phi_{lm}(Z_i)$. Consequently, we can derive an unbiased estimator T_{lm} of $|f_m^{\star l}|^2$

$$T_{lm} = \frac{2}{N(N-1)} \sum_{i_1<i_2} \Phi_{lm}(Z_{i_1}) \overline{\Phi_{lm}(Z_{i_2})},$$

and finally an estimator of $\|f - f_0\|_2^2$

$$T_L = \sum_{l=1}^{L} \sum_{m=-l}^{l} \frac{2}{N(N-1)} \sum_{i_1<i_2} \Phi_{lm}(Z_{i_1}) \overline{\Phi_{lm}(Z_{i_2})}.$$

We can now define a test procedure

$$\Delta = \begin{cases} 1 & \text{if } |T_L| > t^2, \\ 0 & \text{otherwise,} \end{cases}$$

for a threshold $t^2 \sim \sqrt{\mathrm{Var}(T_L)}$ ($L^{2\nu+1}/N$ for ordinary smooth noise and $\exp(2L^\beta/\delta)/N$ for supersmooth noise). The choice of L is crucial but the optimal choice depends on s (the unknown regularity of f) a priori. This point will be solved in Section 3.2.4.

As one may have noticed, the noise smoothness hypothesis and hence the test procedure only rely on the Fourier transform of the noise density f_ε. Consequently, we do not need the existence of the density f_ε but only the existence of the characteristic function $\mathbb{E}(D_{mn}^l(\varepsilon))$ of the variable ε.

3.2.4 Rates of convergence

It is known that the separation rate in case of direct observations in dimension two is $N^{-4s/(4s+2)}$ when one considers for the alternative functions belonging to Sobolev ellipsoid in dimension 2 (see Ingster and Sapatinas (2009)). We can prove (see [L13]) that for our case of indirect observations spoiled by an ordinary smooth noise, this rate is modified in $\psi_N = N^{-2s/(2s+2\nu+1)}$.

This means that testing with a faster rate than $\psi_N = N^{-2s/(2s+2\nu+1)}$ is impossible. If the distance between f_0 and the alternative is smaller than $\psi_N = N^{-2s/(2s+2\nu+1)}$, the sum of the error of the two kinds is close to 1. Nevertheless, it requires the knowledge of the smoothness index s. That is why we want to build on a so-called adaptive test procedure which does not depend on s. But we prove in the next statement that we have to face a phenomenon of "lack of adaptability" for our problem, i.e. it is not possible to test adaptively with the same rate. Indeed, in the context where s is unknown and belongs to some set $\mathcal{S}$, there is not any universal test with small error for each $s \in \mathcal{S}$. The price to pay for adaptivity is an extra factor $\sqrt{\log\log N}$ in the separation rate.

Theorem 31 ([L13]). *Assume that the noise is ordinary smooth with order ν. For all $s \geqslant 1$, let $\psi_N^{ad}(s) = (N/\sqrt{\log\log(N)})^{-2s/(2s+2\nu+1)}$. Let $\mathcal{S}$ be a set such that $\mathcal{S} \cap [1, \infty)$ contains an interval. If $\mathcal{C} \leqslant KR^2$ then,*

$$\liminf_{N\to\infty} \inf_{\Delta_N} \left\{ \mathbb{P}_{f_0}(\Delta_N = 1) + \sup_{s\in\mathcal{S}} \sup_{f\in H_1(s,R,\mathcal{C}\psi_N^{ad}(s))} \mathbb{P}_f(\Delta_N = 0) \right\} \geqslant 1$$

where the infimum is taken over all test procedures Δ_N based on the observations $Z_1, \ldots, Z_N$.

Note on the proof:
The proof is based on the construction of a set of random functions which are far from f_0 is $\mathbb{L}^2$ distance but corresponding to close statistical models. These functions are of the form

$$f_\theta = f_0 + c \sum_L \sum_{l=L/2}^{L-1} \sum_{m=-l}^{l} \theta_{Llm} \varphi_{lm}$$

where φ_{lm} is such that $f_\varepsilon \star \varphi_{lm} = Y_{lm}$. Here, in order to have a result on adaptive estimators, we need to choose a grid of values for s: $s_1 < \cdots < s_{k_N}$ with $k_N \sim \log N$, and corresponding values for J defined by $2^{J_j(2\nu+2s_j+1)} \sim N/\sqrt{\log\log N}$. Then the θ_{Llm} are chosen following a prior $\mu = {k_N}^{-1} \sum_{j=1}^{k_N} \mu_j$ and $\mu_j(\theta_{Llm} = \pm 2^{-J_j(\nu+s_j+1)}) = 1/2$ if $L = 2^{J_j}, 2^{J_j-1} \leqslant l < 2^{J_j}$, $-l \leqslant m \leqslant l$, $\mu_j(\theta_{Llm} = 0) = 1$ otherwise. ■

The next result show that our procedure achieves this rate. For the adaptation in s, a simple maximum over a set of levels L is sufficient.

Theorem 32 ([L13]). *Assume that the noise is ordinary smooth with order ν. Assume $s \geqslant 1$ and $\psi_N^{ad} = (N/\sqrt{\log\log N})^{-2s/(2s+2\nu+1)}$. We consider the set $\mathcal{L} = \{2^{j_0}, \ldots, 2^{j_m}\}$ where $j_0 = \lceil \log_2(\log\log N) \rceil$, $j_m = \lceil \log_2(N(\log\log N)^{-3/2}) \rceil$ and the adaptive test statistic*

$$D_N = \mathbb{1}_{\{\max_{L\in\mathcal{L}}(|T_L|/t_L^2) > \sqrt{2/K_0}\}}$$

with $t_L^2 = L^{2\nu+1}\sqrt{\log\log N}/N$. Then, if $\mathcal{C} > \sqrt{2K_0^{-1}} + ((4\pi)^{-1} + R^2)2^{2s}$,

$$\lim_{N\to\infty} \left\{ \mathbb{P}_{f_0}(D_N = 1) + \sup_{f \in H_1(s,R,\mathcal{C}\psi_N^{ad})} \mathbb{P}_f(D_N = 0) \right\} = 0.$$

Note on the proof:
Usual deconvolution methods on $\mathbb{R}$ use kernel estimators and Fourier transform. But on the sphere, the Fourier analysis (Fourier series instead of Fourier transform) leads to projection estimators. Consequently, the approach proves to be quite different than the one on the real line. The difficulty of testing in a spherical deconvolution model can be seen in the following way. If you use an orthogonal basis (ψ_k) to estimate the unknown function f, then using U-statistics requires that the "deconvolved" basis φ_k (such that $\psi_k = f_\varepsilon \star \varphi_k$) is also (almost) orthogonal, which is delicate to realize. This explains why we choose to use spherical harmonics and their good properties in terms of orthogonality. ■

This result shows that our procedure achieves the minimax rate of testing, and the limiting distribution of the asymptotically minimax test statistic is degenerate. Note that the direct case (without noise) is included in this result, taking $\varepsilon = Id$, $f_\varepsilon^{\star l} = Id$, $\nu = 0$. In this case, the separation rate is $(N/\sqrt{\log\log N})^{-2s/(2s+1)}$. To our knowledge, even in this simpler case, this result was not established yet.

In the case of a supersmooth noise, as usual in deconvolution problems, the rate is degraded into $\psi_N = (\log N)^{-2s/\beta}$ (see the lower bound theorem in [L13]). However, this rate is reached without any knowledge on the smoothness of f:

Theorem 33 ([L13]). *Assume that the noise is supersmooth with order β. Assume $s \geqslant 1/2$ and $\psi_N = (\log N)^{-2s/\beta}$ and $K_0 > 0$. We consider $L^* = \left\lfloor (\delta \log(N)/8)^{1/\beta} \right\rfloor$ and the test statistic*

$$D_N = \mathbb{1}_{\{|T_{L^*}|/t^2_{L^*} > K_0\}}$$

with $t_L^2 = L^{-2\nu_0+1} \exp(2L^\beta/\delta)/N$. Then, if $\mathcal{C} > K_0 + ((4\pi)^{-1} + R^2)(\delta/16)^{-2s/\beta}$,

$$\lim_{N\to\infty} \left\{ \mathbb{P}_{f_0}(D_N = 1) + \sup_{f \in H_1(s,R,\mathcal{C}\psi_N^{ad})} \mathbb{P}_f(D_N = 0) \right\} = 0.$$

A posterior work of Kim et al. (2015) studies the case of a supermooth density: in this case the rate is greatly improved.

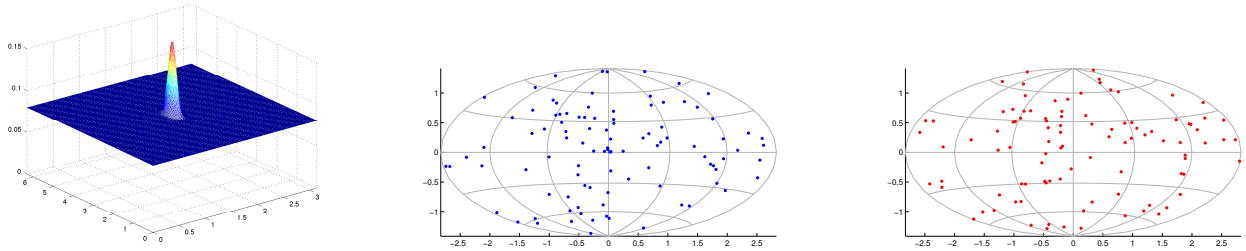
3.2.5 Numerical illustrations

Simulations We have investigated the performances of our testing procedure for two kind of alternatives. These alternatives aim at describing different relevant scenarios in practice.

The first family of alternatives is non isotropic, unimodal with a Gaussian shape. More precisely, it is a mixture of a Gaussian-like density with the uniform density f_0. We will denote this alternative by H_1^a. The H_1^a density has the following form

$$f(x) = (1-\delta)f_0 + \delta h_\gamma(x),$$

where $h_\gamma(x) := C_\gamma \exp(-d(x,x_0)^2/(2\gamma^2))$, d is the spherical distance, C_γ is a normalization constant such that $\int_{\mathbb{S}^2} f(x)dx = 1$ and x_0 is $(\pi/2, 0)$ in spherical coordinates. In the sequel, we chose $\delta = 0.08$ and $\gamma = 5\pi/180$ i.e. $\gamma = 5^o$. Remark that with this choice of parameters, the dose of uniformness injected in H_1^a is high and complicates the detection of the alternative from the null hypothesis. This density is particularly meaningful in the field of astrophysics since very often one seeks for some departure from isotropy and some principal direction. Figure 13 allows to visualize this alternative. The density is represented in spherical coordinates as a surface $z = f(\theta, \phi)$.



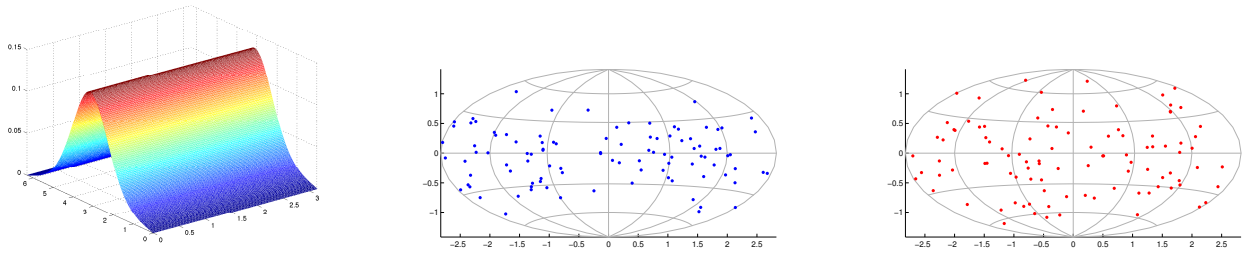
a b c

Figure 13: a/ Representation of the H_1^a density in spherical coordinates. b/ 100 random draws X_i from H_1^a distribution, c/ 100 random draws Z_i from H_1^a convolved with a Laplace noise with variance 0.1

The second alternative that we consider and which is denoted by H_1^b is the Watson distribution (Watson, 1965). Its density is

$$f(\theta, \phi) = C \exp(-2\cos^2(\theta))$$

with C such that $\int_0^{2\pi}\int_0^{\pi} f(\theta,\phi)\sin(\theta)d\theta d\phi = 1$. This distribution has a girdle form, distributed around the equator. This choice is motivated by two reasons : first, this gives an alternative very different from H_1^a, second, it plays a role in applications. For example, in the case of gamma-ray bursts (see Vedrenne and Atteia, 2009), many theories assumed that the sources of these flashes were located around the galactic plane (then a girdle distribution), whereas other proposed that gamma-ray bursts come from beyond the Milky Way (rather a uniform distribution). Figure 14 presents this alternative. Notice that the presence of noise (Figure 14 c/) prevents from seeing the equatorial nature of the distribution.



a b c

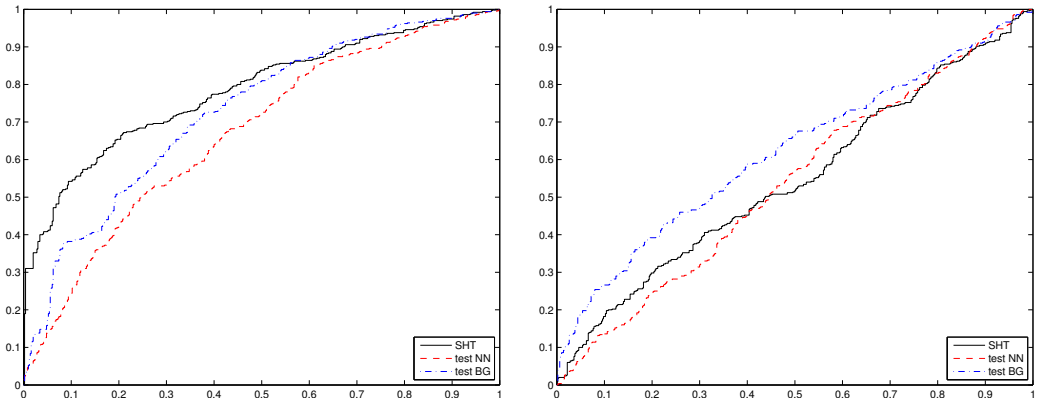
Figure 14: a/ Representation of the Watson density in spherical coordinates. b/ 100 random draws X_i from Watson distribution, c/ 100 random draws Z_i from Watson distribution convolved with a Gaussian noise with variance 0.2

We computed the ROC curves for the three methods for different noise and numbers of observations settings. Let us recall that the *Receiver Operating Characteristic* curves allow to illustrate the performance of a test by plotting the true positive rate vs. the false positive rate, at various threshold settings. Roughly speaking, greater the area under the ROC curve, better the test.

Our adaptive testing procedure is denoted by SHT (as Spherical Harmonics Test). For the quantile K_0, we generate 1000 times N observations uniformly under H_0. Then, we compute by 1000 Monte Carlo runs the 5% quantile of the statistics $\max_{L\in\mathcal{L}}(|T_L|/t_L^2)$ defined in the theorems. We point out that our numerical procedure is notably fast all the more so as we are in dimension 2. Furthermore, we do not have any tuning parameter. To compare our results, we have implemented two other procedures (designed for non-noisy data). The first one is called *the Nearest Neighbour test* and was proposed by Quashnock and Lamb (1993), it will be denoted NN in the sequel. The second procedure was introduced by Beran (1968) and Giné (1975). We precise that on Figure 16a the solid line corresponding to the performance of our procedure SHT is mixed up with the axes passing through the points $(0,0),(0,1)$ and $(1,1)$.

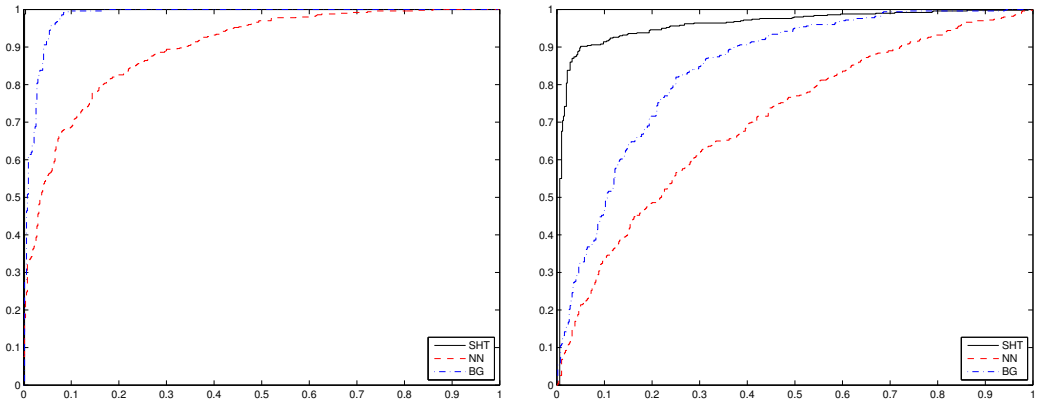
Real data: UHECR To apply our procedure to UHECR data of observatory Pierre Auger (The Pierre AUGER Collaboration (2010)), we need to take into account the observatory exposure. Indeed, only cosmic rays with zenith angle of arrival less that 60^o can be observed. Then, a coverage function over the years of observation can be computed from geometrical considerations and it is displayed in Figure 17.

In addition to the noise due to extragalactic magnetic fields, a selection is done depending on whether the ray is in the observation area. Denoting the coverage density by g_0, the observations are now $V_1,\ldots,V_N$ where the density of V is proportional to g_0 times f_Z: $f_V = cg_0f_Z$, with c



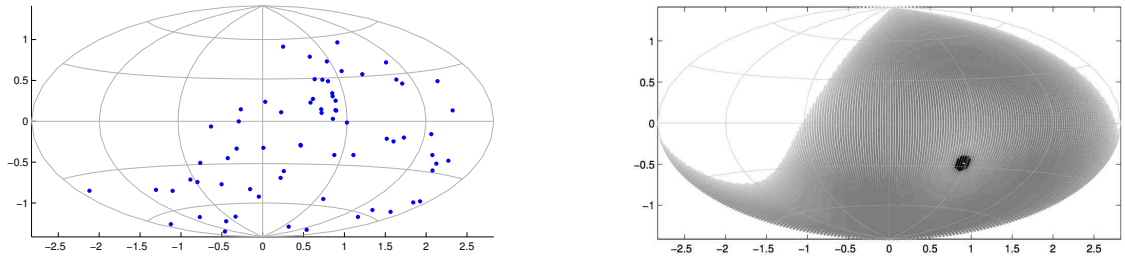
a b

Figure 15: ROC Curves for the three methods and for the alternative H_1^a: a/ No noise and $N = 100$. b/ Laplace noise with variance 0.1 and $N = 100$.



a b

Figure 16: ROC Curves for the three methods and for the alternative H_1^b: a/ No noise and $N = 100$. b/ Laplace noise with variance 0.1 and $N = 100$.



a b

Figure 17: a/ Representation of the 69 arrival directions of highest energy cosmic rays (Pierre Auger data) b/ Coverage function g_0 for the Pierre Auger observatory (the darker the more observed, white area non-observed)

such that f_V is a density. The relevant test is then $f = f_0 \Leftrightarrow f_V = g_0$. Although we do not extend our theorems to this case, we nevertheless implement an extended method. Our initial test procedure is based on the estimation of $(f_Z)_n^{*l}$ by $N^{-1}\sum_i \overline{Y_n^l(Z_i)}$. Then it is sufficient to apply the same procedure but with estimator $N^{-1}\sum_{i=1}^N (\overline{Y_n^l}/(cg_0))(V_i)$. Indeed this quantity approximates $\int f_V \overline{Y_n^l}/(cg_0) = \int f_Z \overline{Y_n^l} = (f_Z)_n^{*l}$. Using this method, we obtain the following p-values, assuming different kinds of possible noise.

Noise type	No noise	Laplace			Gaussian		
variance		0.05	0.1	0.2	0.05	0.1	0.2
p	0.003	0.014	0.034	0.092	0.016	0.001	0.076

Then our method confirms what was already noticed by astrophycisists: there seems to be some kind of anisotropy in the UHECR phenomenon.

3.2.6 Some prospects

Here we present some prospects for statistical works in view of astrophysical applications.

• In this section 3.2, we only dealt with the case of the uniform density, and some tools developed in the proofs are specific to the uniform distribution. The case of other given densities is of course of great interest in practice.

• A model which appears naturally in astrophysics in the one where the observations come from a mixture between a signal and a background noise with an unknown proportion p (close to 1):

$$Z \sim pf_\varepsilon + (1-p)f.$$

The noise distribution f_ε is already identified. The question is then to estimate simultaneously the proportion of the mixture and the signal density.

• Consider the following model: we observe $X_i, \dots, X_n$ and $Y_1, \dots, Y_m$ where the X_i's are i.i.d. variables on the sphere $\mathbb{S}^2$ with density f, and the Y_i's are i.i.d. on $\mathbb{S}^2$ with density g, the two sample being independent. The aim is then to test the hypothesis $H_0 : f = g$ versus the alternative

$$H_1(s, \mathcal{C}\psi_{nm}) : f \neq g, \quad f, g \text{ with smoothness } s, \quad \|f-g\|_2^2 \geqslant \mathcal{C}\psi_{nm}.$$

This homogeneity test has application in astrophysics, in particular for the study of neutrinos (Antares or Icecube Experiments). The data are now numerous, that allows us to hope a great value for n. For the moment, with Gilles Fäy and Thanh Mai Pham Ngoc, we have introduced an adaptive test using needlets, with rate $\psi_{mn} = (N/\sqrt{\log\log N})^{-2s/(2s+1)}$ where $N^{-1} = n^{-1} + m^{-1}$. Following discussions with Bruny Baret of laboratory APC (AstroParticules et Cosmologie, Paris Diderot), we would like to study the case of noisy data, which seems more realistic.

3.3 Nonparametric inference for hidden Markov chain

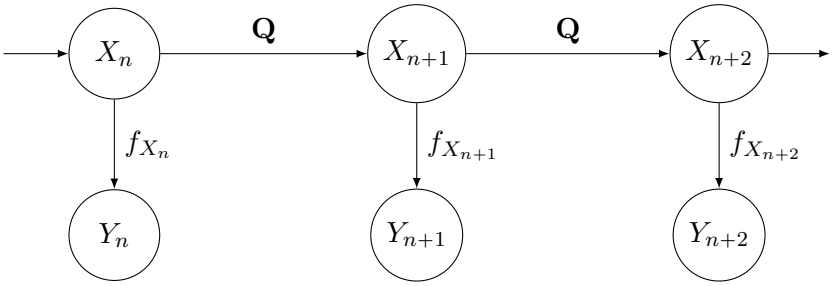
A direct extension of the convolution model consists in the hidden Markov chains (HMM). In my PhD thesis, I had addressed the particular case of Markov chain only observed through an additive noise, which is exactly the convolution model when the signal is Markovian ([L4], [L5]). Here, I prefer to detail my work on general hidden Markov models with finite state space. These models seem to be reliable to depict practical situations in a variety of applications such as economics, genomics, signal processing and image analysis, ecology, environment, speech recognition, to name but a few. In this framework a huge literature is concerning the case where the observations distribution belongs to a parametric family. Inference is then achieved via Monte Carlo methods, maximum likelihood estimators, EM (Expectation-Minimization) algorithm, see Cappé et al. (2005). The nonparametric case requires another approach.

More precisely, the model that we study here is the following. From latent variables $(X_n)_{n\geqslant 1}$ which form a Markov chain with K possible values, the observations $(Y_n)_{n\geqslant 1}$ are independent conditionally to $(X_n)_{n\geqslant 1}$:

$$\mathcal{L}\left((Y_n)_{n\geqslant 1}|(X_n)_{n\geqslant 1}\right) = \bigotimes_{n\geqslant 1} \mathcal{L}\left(Y_n|X_n\right).$$

We assume moreover that the distribution of Y_n given $X_n = x$ has a density with respect to the Lebesgue measure on $\mathbb{R}$, denoted by f_x. From observations $Y_1, \ldots, Y_N$, the model parameters to be inferred are then:

- the transition matrix $\mathbf{Q}$ of the Markov chain on $\{1, \ldots, K\}$ (K is assumed to be known),
- the emission densities $f_1, \ldots, f_K$.



Until very recently, asymptotic performances of estimators were proved theoretically only in the parametric frame (that is, with finitely many unknown parameters). Though, nonparametric methods for HMMs have been considered in applied papers: see references in [L18]. Recent papers that contain theoretical results on different kinds of nonparametric HMMs are Gassiat and Rousseau (2015), where the emitted distributions are translated of each other, and Dumont and Le Corff (2012) in which the authors consider regression models with hidden regressor variables that can be Markovian on a continuous state space.

The preliminary obstacle to obtain theoretical results on general finite state space nonparametric HMMs was to understand when such models are indeed identifiable. The papers Allman et al. (2009), Hsu et al. (2012) and Anandkumar et al. (2012) paved the way to obtain identifiability under reasonable assumptions. In Anandkumar et al. (2012) the authors point out a structural link between multivariate mixtures with conditionally independent observations and finite state space HMMs. In Hsu et al. (2012) the authors propose a spectral method to estimate all parameters for finite state space HMMs (with finitely many observations), under the assumption that the transition matrix of the hidden chain is non singular, and that the (finitely valued)

emission distributions are linearly independent. Those spectral methods have the extremely interesting characteristic that to compute the estimator the algorithms do not require initialization as is usual in latent variable models estimation when using the EM algorithm. They may be used under the linear independence assumption. Extension to emission distributions on any space, under the linear independence assumptions (and keeping the assumption of non singularity of the transition matrix), allowed to prove the general identifiability result for finite state space HMMs. Gassiat et al. (2015) have established the following result: if the probability densities $f_1, \ldots, f_K$ are linearly independent, and if $\mathbf{Q}$ has full rank, then the parameters $\mathbf{Q}$ and $f_1, \ldots, f_K$ are identifiable from the distribution of three consecutive observations Y_1, Y_2, Y_3, up to label swapping of the hidden states (later, Alexandrovich and Holzmann (2014) obtained identifiability when the emission distributions are all distinct, not necessarily linearly independent). Thus our assumptions are the following.

Assumption (H)

- The Markov chain $(X_n)_{n\geqslant 1}$ is irreducible and aperiodic,
- The initial distribution $\pi = (\pi_1, \ldots, \pi_K)$ is the stationary distribution,
- The transition matrix $\mathbf{Q}$ has full rank,
- The family of emission densities $\{f_1, \ldots, f_K\}$ is linearly independent.

3.3.1 The spectral method

From now on we assume that the emission densities are in $\mathbb{L}^2(\mathbb{R})$ and we shall use this Hilbertian structure. The first step is to choose a sieve of finite dimensional subspaces with orthonormal basis $\Phi_M = \{\varphi_1, \ldots, \varphi_M\}$: for example splines or Fourier basis or wavelets. Since we can set

$$\widehat{f}_{M,k} = \sum_{m=1}^{M} \widehat{\langle f_k, \varphi_m \rangle}\, \varphi_m,$$

the problem is reduced to estimate $\mathbf{Q}$ and $\langle f_k, \varphi_m \rangle = \mathbb{E}(\varphi_m(Y_n)|X_n = k)$, $k = 1, \ldots, K$, $m = 1, \ldots, M$ on the basis of the empirical distribution of the three-dimensional marginal, i.e. the distribution of (Y_1, Y_2, Y_3). Then we face a parametric problem and we can use the works of Song et al. (2014) and Hsu et al. (2012) (see also references therein). They propose an algorithm which uses only one SVD (singular value decomposition), matrix inversions and one diagonalization. We need the following notation: $\forall (a,b,c) \in \{1, \ldots, M\}^3, \forall (m,k) \in \{1, \ldots, M\} \times \{1, \ldots, K\}$

$$\begin{aligned}
\mathbf{P}_{123}(a,b,c) &= \mathbb{E}(\varphi_a(Y_1)\varphi_b(Y_2)\varphi_c(Y_3)) \\
\mathbf{P}_{13}(a,c) &= \mathbb{E}(\varphi_a(Y_1)\varphi_c(Y_3)) \\
\mathbf{A}_M(m,k) &= \mathbb{E}(\varphi_m(Y_1)|X_1 = k) = \langle f_k, \varphi_m \rangle
\end{aligned}$$

Note that $\mathbf{P}_{13}$ and $\mathbf{P}_{123}$ can be estimated by their empirical version, and $\mathbf{A}_M$ is the unknown parameter. The crucial lemma is the following one

Lemma 34. *Let U be any $M \times K$ matrix such that $\mathbf{P}_{13}U$ has rank K. Then $U^\top \mathbf{P}_{13} U$ is invertible and there exists an invertible matrix R such that*

$$\forall b \in \{1, \ldots, M\}, \quad B(b) := (U^\top \mathbf{P}_{13} U)^{-1} U^\top \mathbf{P}_{123}(.\,, b, .\,) U = R \mathrm{Diag}[\mathbf{A}_M(b, .\,)] R^{-1}.$$

This lemma seems technical but it permits to link observable (and easy to estimate) quantities $\mathbf{P}_{13}$, $\mathbf{P}_{123}$ with the unknown target $\mathbf{A}_M$. It is then sufficient to estimate $B(b)$ and to diagonalize it to retrieve the bth row of $\mathbf{A}_M$. An important point is that the change-of-basis matrix R does not depend on b. Actually a slightly modification is necessary in order to separate the eigenvalues. Indeed the replacement of $\mathbf{P}_{13}$ and $\mathbf{P}_{123}$ by their estimates induce a modification of the eigenspace which is under control only for sufficiently separated eigenvalues. The use of a random matrix rotation Θ can fix this problem with a quantifiable cost. The final algorithm is described in Algorithm 1 below.

To state the result about this estimation method, we need to introduce the following quantity:

$$\eta^2(\Phi_M) := \sup_{y,y'} \sum_{a,b,c=1}^{M} (\varphi_a(y_1)\varphi_b(y_2)\varphi_c(y_3) - \varphi_a(y_1')\varphi_b(y_2')\varphi_c(y_3'))^2.$$

Note that in classical examples (Spline, Fourier, Wavelets) we have: $\eta(\Phi_M) \leqslant C_\eta M^{\frac{3}{2}}$ where $C_\eta > 0$ is a constant. To control our estimators performance, we use $\mathbb{L}^2$ norm for the densities and the spectral norm for matrices. We shall denote $f_{M,k}$ the orthogonal projection of f_k on $\mathrm{Span}\{\Phi_M\}$ and $f_M = (f_{M,1}, \ldots, f_{M,K})$. As usual in nonparametric estimation, the risk for f_k is decomposed in a bias term $\|f_k - f_{M,k}\|_2$, which comes from the approximating properties of the spaces $\mathrm{Span}\{\Phi_M\}$ and decreases when M increases, and in a variance term which comes from the estimation, and increases when M increases. A good choice of M has to balance those two terms. The aim of the following result is to bound the so-called variance term, and what is important is to get a precise behavior of the upper bound with respect to both N and M.

Theorem 35 ([L18])**.** *Assume* **(H)**. *Then, there exist positive constant numbers $\mathcal{C}$ and $N^\star$ such that the following holds. Let $x > \log 6$, M large enough and $N \geqslant N^\star \eta(\Phi_M)^2 x$. With probability greater than $1 - 6e^{-x}$,* up to label switching*,*

$$\|f_{M,k} - \hat{f}_{M,k}\|_2 \leqslant \mathcal{C}\frac{\eta(\Phi_M)}{\sqrt{N}}x\,, \qquad \|\mathbf{Q} - \hat{\mathbf{Q}}\| \leqslant \mathcal{C}\frac{\eta(\Phi_M)}{\sqrt{N}}x\,.$$

Moreover, if $M = M_N$ is such that $\eta(\Phi_{M_N}) = o(\sqrt{N})$,

$$\mathbb{E}\big[\|f_{M_N,k} - \hat{f}_{N,k}\|_2^2\big] = O\Big(\frac{\eta^2(\Phi_{M_N})}{N}\Big), \qquad \mathbb{E}\big[\|\mathbf{Q} - \hat{\mathbf{Q}}\|^2\big] = O\Big(\frac{\eta^2(\Phi_{M_N})}{N}\Big).$$

Here, the expectations are taken on the observations and on the random unitary matrix drawn at [Step 4] of the spectral algorithm.

Note on the proof:
The proof is entirely based on perturbation matrix theory: it is about controlling singular values, eigenvalues, eigenvectors when a small perturbation is applied. ■

According to this theorem, concerning the parametric part, if we choose M_N such as $\eta(\Phi_{M_N}) = (\log N)^\delta$ for some positive δ, we get that

$$\mathbb{E}\left[\|\mathbf{Q} - \hat{\mathbf{Q}}\|^2\right] = O\left(\frac{(\log N)^{2\delta}}{N}\right)$$

which is quasi the parametric rate.

Input: $Y_{1:N} = (Y_1, \dots, Y_N)$ observation; Basis Φ_M of the projection space;
Output: Estimation of the transition $\hat{\mathbf{Q}}$ (and its stationary distribution $\hat{\pi}$) and the emission laws $\hat{f}$.

[Step 1] Consider the following empirical estimators: for any a, b, c in $\{1, \dots, M\}$,

$$\hat{\mathbf{P}}_1(a) = N^{-1} \sum_{s=1}^{N} \varphi_a(Y_s), \quad \hat{\mathbf{P}}_{123}(a,b,c) := \frac{1}{N} \sum_{s=1}^{N-2} \varphi_a(Y_s)\varphi_b(Y_{s+1})\varphi_c(Y_{s+2})$$

$$\hat{\mathbf{P}}_{13}(a,c) := \frac{1}{N} \sum_{s=1}^{N-2} \varphi_a(Y_s)\varphi_c(Y_{s+2}), \quad \hat{\mathbf{P}}_{12}(a,b) = N^{-1} \sum_{s=1}^{N-1} \varphi_a(Y_s)\varphi_b(Y_{s+1})$$

[Step 2] Let $\hat{\mathbf{U}}$ be the $M \times K$ matrix of orthonormal right singular vectors of $\hat{\mathbf{P}}_{13}$ corresponding to its top K singular values.

[Step 3] Form the matrices:

$$\forall b \in \{1, \dots, M\}, \quad \hat{\mathbf{B}}(b) := (\hat{\mathbf{U}}^\top \hat{\mathbf{P}}_{13} \hat{\mathbf{U}})^{-1} \hat{\mathbf{U}}^\top \hat{\mathbf{P}}_{123}(.\,, b, .\,) \hat{\mathbf{U}}\,.$$

[Step 4] Set Θ a $(K \times K)$ random unitary matrix uniformly drawn and form the matrices:

$$\forall k \in \{1, \dots, K\}, \quad \hat{\mathbf{C}}(k) := \sum_{b=1}^{M} (\hat{\mathbf{U}}\Theta)(b,k) \hat{\mathbf{B}}(b)\,.$$

[Step 5] Compute $\hat{\mathbf{R}}$ a $(K \times K)$ unit Euclidean norm columns matrix that diagonalizes the matrix $\hat{\mathbf{C}}(1)$:

$$\hat{\mathbf{R}}^{-1} \hat{\mathbf{C}}(1) \hat{\mathbf{R}} = \mathrm{Diag}[(\hat{\Lambda}(1,1), \dots, \hat{\Lambda}(1,K))]\,.$$

[Step 6] Set:

$$\forall k, k' \in \{1, \dots, K\}, \quad \hat{\Lambda}(k,k') := (\hat{\mathbf{R}}^{-1} \hat{\mathbf{C}}(k) \hat{\mathbf{R}})(k',k')\,,$$

and $\hat{\mathbf{A}}_M := \hat{\mathbf{U}}\Theta\hat{\Lambda}$.

[Step 7] Consider the emission laws estimator $(\hat{f}_{M,k})_{1 \leqslant k \leqslant K}$ defined by:

$$\forall k \in \{1, \dots, K\}, \quad \hat{f}_{M,k} := \sum_{m=1}^{M} \hat{\mathbf{A}}_M(m,k) \varphi_m\,,$$

[Step 8] Set

$$\tilde{\pi} := \left(\hat{\mathbf{U}}^\top \hat{\mathbf{A}}_M\right)^{-1} \hat{\mathbf{U}}^\top \hat{\mathbf{P}}_1.$$

[Step 9] Consider the transition matrix estimator:

$$\hat{\mathbf{Q}} := \Pi_{\mathrm{TM}}\Big(\left(\hat{\mathbf{U}}^\top \hat{\mathbf{A}}_M \mathrm{Diag}[\tilde{\pi}]\right)^{-1} \hat{\mathbf{U}}^\top \hat{\mathbf{P}}_{12} \hat{\mathbf{U}} \left(\hat{\mathbf{A}}_M^\top \hat{\mathbf{U}}\right)^{-1} \Big)\,,$$

where Π_{TM} denotes the projection (with respect to the scalar product given by the Frobenius norm) onto the convex set of transition matrices, and define $\hat{\pi}$ as the stationary distribution of $\hat{\mathbf{Q}}$.

Algorithm 1: Nonparametric Spectral Estimation

Concerning the parametric part, the theorem states that the variance term is typically of order M^3/N. To get a control on the risk $\|f_k - \hat{f}_{M,k}\|_2$ one has to make a trade-off with the bias term $\|f_k - f_{M,k}\|_2$, which has order $O(M^{-\alpha})$ where α is the minimal regularity of the emission laws. Choosing $M^{3+2\alpha} \sim N$, this leads to to the rate $N^{-\alpha/(2\alpha+3)}$ for the non parametric estimation. This is similar to the rate of estimation of a density in dimension 3 with smoothness α. However our target densities are in dimension 1, so that we should find estimators with variance M/N. Here, the variance term of the spectral estimator has order M^3/N because it comes from the nonparametric estimation of a density of dimension 3: that of the distribution of (Y_1, Y_2, Y_3). To get a variance term of order M/N, we shall use the fact that the intrinsic complexity of the statistical model for the distribution of (Y_1, Y_2, Y_3) is not that of a distribution on $\mathbb{R}^3$ but of K distributions on $\mathbb{R}$.

3.3.2 The penalized least-squares method

For $f = (f_1, \dots, f_K)$ densities on $\mathbb{R}$ and Q a transition matrix with stationary distribution π, set $g^{Q,f}$ a possible density of (Y_1, Y_2, Y_3)

$$g^{Q,f}(x_1, x_2, x_3) = \sum_{k_1,k_2,k_3=1}^{K} \pi(k_1)Q(k_1, k_2)Q(k_2, k_3)f_{k_1}(x_1)f_{k_2}(x_2)f_{k_3}(x_3).$$

We can estimate $g = g^{\mathbf{Q},f}$, which is the density of the observations, with standard methods in density estimation. Next the following proposition is crucial. Unfortunately it is proved only for $K = 2$ hidden states. In such a case, $f = (f_1, f_2)$, and

$$\mathbf{Q} = \begin{pmatrix} 1 - p^\star & p^\star \\ q^\star & 1 - q^\star \end{pmatrix}$$

for some $p^\star, q^\star$ in $[0, 1]$. We shall assume that the coefficients $p^\star$ and $q^\star$ verify $0 < p^\star < 1,\ 0 < q^\star < 1,\ p^\star \neq 1 - q^\star$.

Proposition 36 ([L18]). *Let $\mathcal{K}$ be a compact subset of $\mathbb{L}^2$ such that if $\mathbf{h} = (h_1, h_2) \in \mathcal{K}$, then $\int h_1 = 0 = \int h_2$. Let $\mathcal{V}$ be a compact neighborhood of $\mathbf{Q}$ such that, for all $Q \in \mathcal{V}$, Q verifies $0 < p < 1,\ 0 < q < 1,\ p \neq 1 - q$. Assume that $f_1 \neq f_2$. Then there exists a positive constant c such that*

$$\forall \mathbf{h} = (h_1, h_2) \in \mathcal{K}^2,\ \forall Q \in \mathcal{V}, \|g^{Q,f+\mathbf{h}} - g^{Q,f}\|_2 \geqslant c\|\mathbf{h}\|_Q,$$

where

$$\|\mathbf{h}\|_{\mathbf{Q}} := \begin{cases} \|h_1\|_2 + \|h_2\|_2 & \text{if } \mathbf{Q}(1,2) \neq \mathbf{Q}(2,1), \\ \min(\|h_1\|_2 + \|h_2\|_2, \|h_1 + f_1 - f_2\|_2 + \|h_2 + f_2 - f_1\|_2 & \text{if } \mathbf{Q}(1,2) \neq \mathbf{Q}(2,1). \end{cases}$$

Then if we bound $\|g^{\mathbf{Q},\hat{f}} - g^{\mathbf{Q},f}\|_2$, we shall bound (up to label switching) $\|\hat{f}_k - f_k\|_2$ for $k = 1, 2$. Now let us explain our estimation of g. Define the classical empirical contrast, for any $t : \mathbb{R}^3 \to \mathbb{R}$:

$$\gamma_N(t) = \|t\|_2^2 - \frac{2}{N}\sum_{s=1}^{N-2} t(Y_s, Y_{s+1}, Y_{s+2}),$$

which is an empirical counterpart of $\|t - g\|_2^2 - \|g\|_2^2 = \|t\|_2^2 - 2\langle t, g\rangle$. We fix a compact subset $\mathcal{F}$ of $\mathbb{L}^2$ such that for any $f \in \mathcal{F}$, $\int f = 1$ and $\|f\|_\infty \leqslant C_{\mathcal{F},\infty}$ for some fixed $C_{\mathcal{F},\infty} > 0$. Define

$\mathcal{S}(Q, M)$ as the set of functions $g^{Q,f}$ such that

$$\forall k = 1, \dots, K,\ \exists (a_{mk})_{1 \leqslant m \leqslant M} \in \mathbb{R}^M,\ f_k = \sum_{m=1}^{M} a_{mk} \varphi_m \text{ and } f_k \in \mathcal{F}.$$

Let now $\hat{\mathbf{Q}}$ be an estimator of $\mathbf{Q}$ (for instance the spectral estimator). For each M, we define $\hat{g}_M = g^{\hat{\mathbf{Q}}, \hat{f}^M}$ as a minimizer of $\gamma_N(t)$ for $t \in \mathcal{S}(\hat{\mathbf{Q}}, M)$. To choose a value for M, we set a penalty function $\mathrm{pen}(M)$ and choose

$$\hat{M} = \underset{M=1,\dots,N}{\operatorname{argmin}} \left\{ \gamma_N(\hat{g}_M) + \mathrm{pen}(M) \right\}.$$

Then the estimator of g is $\hat{g} = \hat{g}_{\hat{M}}$, and the estimator of f is the corresponding $\hat{f}$ such that $\hat{g} = g^{\hat{\mathbf{Q}}, \hat{f}^{\hat{M}}}$ i.e.

$$\hat{f} := \hat{f}^{\hat{M}}.$$

Using model selection machinery, we can prove an oracle inequality for the estimation of g, that is: there exists $\kappa^\star$ such that if

$$\mathrm{pen}(M) \geqslant \kappa^\star \frac{M \log M}{N}$$

then, up to label switching, with probability $1 - (e-1)^{-1} e^{-x}$

$$\|\hat{g} - g\|_2^2 \leqslant 6 \inf_M \left\{ \|g - g^{\mathbf{Q}, f_M}\|_2^2 + \mathrm{pen}(M) \right\} + C_1 \frac{x}{N} + C_2 \big(\|\mathbf{Q} - \hat{\mathbf{Q}}\|^2 + \|\pi - \hat{\pi}\|_2^2 \big).$$

This gives

Theorem 37 ([L18])**.** *Assume assumption* **(H)** *with $K = 2$ hidden states. Then if* $\mathrm{pen}(M) \geqslant \kappa^\star \frac{M \log M}{N}$*, then* up to label switching*, for all $N \geqslant (x \vee x^2) N^\star \log N$, with probability larger than* $1 - 8e^{-x}$*,*

$$\|f_1 - \hat{f}_1\|_2^2 + \|f_2 - \hat{f}_2\|_2^2 \leqslant C \left[\inf_M \left\{ \|f_1 - f_{M,1}\|_2^2 + \|f_2 - f_{M,2}\|_2^2 + \mathrm{pen}(M) \right\} + \frac{x}{N} \right.$$
$$\left. + \|\mathbf{Q} - \hat{\mathbf{Q}}\|^2 + \|\pi - \hat{\pi}\|_2^2 \right].$$

Moreover, if $\mathbf{Q}$ and π are estimated with rate $\sqrt{(\log N)/N}$, when N tends to infinity,

$$\mathbb{E}\left[\|f_1 - \hat{f}_1\|_2^2 + \|f_2 - \hat{f}_2\|_2^2 \right] = O\left(\inf_M \left\{ \|f_1 - f_{M,1}\|^2 + \|f_2 - f_{M,2}\|^2 + \mathrm{pen}(M) \right\} + \frac{\log N}{N} \right).$$

Note on the proof:
We use concentration inequalities for dependent variables of Paulin (2014). Here the model is not a vector space and we have to make a fine work and to use bracketing entropy computations to catch the true complexity (MK instead of M^3). ■

Thus, choosing $\mathrm{pen}(M) = \kappa M \log M / N$ for a large κ leads to the minimax asymptotic rate of convergence up to $\log N$. Indeed, standard results in approximation theory show that one can upper bound the approximation error $\|f_k - f_{M,k}\|_2$ by $O(M^{-\alpha})$ where $\alpha > 0$ denotes a

regularity parameter. Then the trade-off is obtained for $M \sim (N/\log N)^{1/(2\alpha+1)}$, which leads to the quasi-optimal rate $(N/\log N)^{-\alpha/(2\alpha+1)}$ for the nonparametric estimation when the minimal smoothness of the emission densities is α. Notice that the algorithm automatically selects the best M leading to this rate.

To implement the estimator, it remains to choose a value for κ in the penalty. In practice we have used the slope heuristic (see Baudry et al., 2012).

3.3.3 Conclusion and illustration

To sum up, the spectral method have the great advantage to avoid initialization problems. But it does not achieve the minimax rate: it over-estimates the intrinsic complexity of the statistical model. By contrast, least squares estimator improves the quadratic risk at any fixed approximation level, and using model selection with least squares estimators leads to minimax rates. Then our final procedure is to initialize least squares minimization with spectral estimators.

We understand that a crucial step lies in computing least squares estimators $\hat{g}_M$. One may struggle to compute $\hat{g}_M$ since the function γ_N is non-convex. It follows that an acceptable procedure must start from a good approximation of $\hat{g}_M$. This is done by the spectral method. Then we propose a two steps estimation procedure that starts by the spectral estimator. The latter seems to be a good candidate to initialize an iterative scheme that will converge towards $\hat{g}_M$. Hence we compute $\hat{g}_M$ for each $M = 1, \ldots, N$ as follows

- First compute the spectral estimator. This is straightforward using the procedure described in Algorithm 1. In particular, the spectral estimator gives an estimation $\hat{\mathbf{Q}}, \hat{\pi}$ of the transition matrix and its stationary distribution which is used to compute the least squares contrast function.

- Use the spectral estimator of the emission densities as a starting point for "Covariance Matrix Adaptation Evolution Strategy" (CMA-ES), see Hansen (2006). This iterative algorithm may ultimately find a local/global minimum of the contrast function.

To illustrate the performance of our method, we present here some numerical experiments. We consider the regular histogram basis for estimating $K = 2$ emission laws given by beta laws of parameters $(2, 5)$ and $(4, 3)$ from a single chain of size $N = 30,000$.

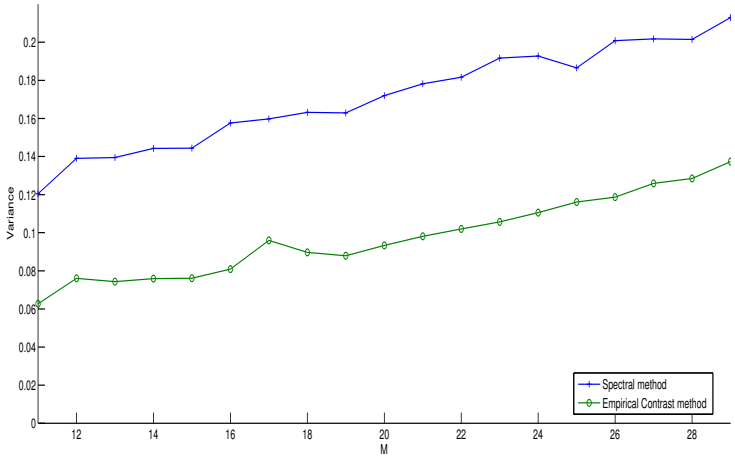


Figure 18: Comparison of the variances of the spectral and the least squares estimators.

A first numerical experiment, depicted in Figure 18, compares, for each M, the variances (i.e. the $\mathbb{L}^2$-distance between the estimator and the orthogonal projection onto the subspace generated by the basis Φ_M) obtained by the spectral method and the empirical least squares method over 100 iterations on chains of length $40,000$. It consolidates the idea that the least square method significantly improves the $\mathbb{L}^2$-distance to the best approximation of the emission laws. Indeed, even for small values of M, one may see that the variance is divided by two in Figure 18.

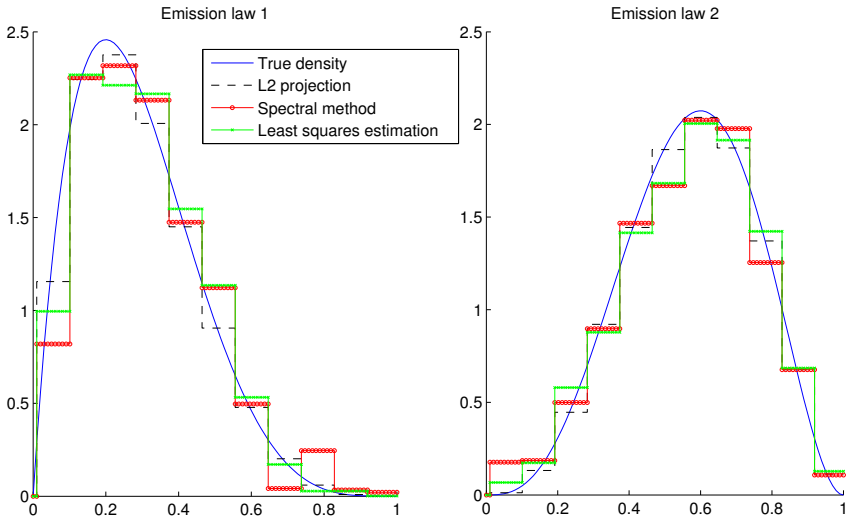


Figure 19: Estimators of the emissions densities using the regular histogram basis

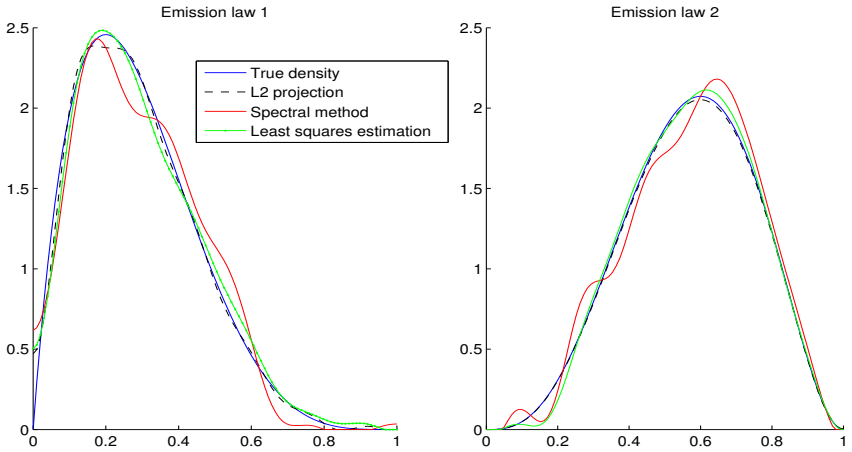


Figure 20: Estimators of the emissions densities using the Fourier basis

One can see on Figures 19 and 20 that our method also qualitatively improves upon the spectral method.

Chapter 4

Bibliography

List of my papers

[L1] Lacour, C. (2006) Rates of convergence for nonparametric deconvolution. *C. R. Acad. Sci. Paris* 342 (11), 877–882

[L2] Lacour, C. (2008) Nonparametric estimation of the stationary density and the transition density of a Markov chain. *Stochastic Processes and their Applications* 118 (2), 232–260

with Erratum in (2012) *Stochastic Processes and their Applications* 122 (6), 2480–2485

[L3] Lacour, C. (2007) Adaptive estimation of the transition density of a Markov chain. *Annales de l'Institut Henri Poincaré Probab. Statist.* 43 (5), 571–597

[L4] Lacour, C. (2008) Adaptive estimation of the transition density of a particular hidden Markov chain. *Journal of Multivariate Analysis* 99 (5), 787-814

[L5] Lacour, C. (2008) Least-square type estimation of the transition density of a particular hidden Markov chain. *Electronic Journal of Statistics* 2, 1–39

[L6] Brunel, E., Comte, F. and Lacour, C. (2007) Adaptive estimation of the conditional density in presence of censoring. *Sankhyā* 69(4), 734–763

[L7] Comte, F., Lacour, C. and Rozenholc, Y. (2010) Adaptive estimation of the dynamics of a discrete time stochastic volatility model. *Journal of Econometrics* 154 (1), 59-73

[L8] Brunel, E., Comte, F. and Lacour, C. (2010) Minimax estimation of the conditional cumulative distribution function. *Sankhyā A* 72 (2), 293–330

[L9] Comte, F. and Lacour, C. (2010) Pointwise deconvolution with unknown error distribution. *C. R. Acad. Sci. Paris* 348, 323–326

[L10] Comte, F. and Lacour, C. (2011) Data driven density estimation in presence of unknown convolution operator. *Journal of the Royal Statistical Society, Ser B.* 73(4), 601–627

[L11] Akakpo, N. and Lacour, C. (2011) Inhomogeneous and anisotropic conditional density estimation from dependent data. *Electronic Journal of Statistics* 5, 1618–1653

[L12] Comte, F. and Lacour, C. (2013) Anisotropic adaptive kernel deconvolution. *Annales de l'Institut Henri Poincaré Probab. Stat.* 49(2), 569–609

[L13] Lacour, C. and Pham Ngoc, T. M. (2014) Goodness-of-fit test for noisy directional data. *Bernoulli* 20(4), 2131–2168

[L14] Bec, M. and Lacour C. (2015) Adaptive kernel estimation of the Lévy density. *Statistical Inference for Stochastic Processes* 18(3), 229–256

[L15] Bertin, K., Lacour, C. and Rivoirard, V. Adaptive pointwise estimation of conditional density function. To appear in *Annales de l'Institut Henri Poincaré Probab. Stat.*

[L16] Chagny, G. and Lacour, C. (2015) Optimal adaptive estimation of the relative density. *TEST* 24(3), 605–631

[L17] Lacour, C. and Massart, P. Minimal penalty for the Goldenshluger-Lepski method. *Submitted*

[L18] De Castro, Y., Gassiat, E. and Lacour, C. Minimax adaptive estimation of nonparametric hidden Markov models. *Submitted*

References

Abbaszadeh, M., Chesneau, C., and Doosti, H. (2013). Multiplicative censoring: estimation of a density and its derivatives under the L_p-risk. *REVSTAT*, 11(3):255–276.

Adamczak, R. (2008). A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.*, 13(34):1000–1034.

Aït-Sahalia, Y. (2001). Transition densities for interest rate and other nonlinear diffusions [J. Finance **54** (1999), no. 4, 1361–1395]. In *Quantitative analysis in financial markets*, pages 1–34. World Sci. Publ., River Edge, NJ.

Akakpo, N. (2009). *Estimation adaptative par sélection de partitions en rectangles dyadiques*. PhD thesis, Université Paris-Sud 11.

Akakpo, N. (2012). Adaptation to anisotropy and inhomogeneity via dyadic piecewise polynomial selection. *Math. Methods Statist.*, 21(1):1–28.

Alexandrovich, G. and Holzmann, H. (2014). Nonparametric identification of hidden Markov models. *arXiv:1404.4210*.

Allman, E. S., Matias, C., and Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.*, 37(6A):3099–3132.

Amann, H. (2000). Compact embeddings of vector-valued Sobolev and Besov spaces. *Glas. Mat. Ser. III*, 35(55)(1):161–177. Dedicated to the memory of Branko Najman.

Anandkumar, A., Hsu, D., and Kakade, S. M. (2012). A method of moments for mixture models and hidden Markov models. *arXiv:1203.0683*.

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical models based on counting processes*. Springer Series in Statistics. Springer-Verlag, New York.

Athreya, K. B. and Atuncar, G. S. (1998). Kernel estimation for real-valued Markov chains. *Sankhyā Ser. A*, 60(1):1–17.

Autin, F., Claeskens, G., and Freyermuth, J.-M. (2014). Hyperbolic wavelet thresholding methods and the curse of dimensionality through the maxiset approach. *Applied and Computational Harmonic Analysis*, 36(2):239–255.

Azzalini, A. and Bowman, A. (1990). A look at some data on the old faithful geyser. *Applied Statistics*, 39(3):357–365.

Barron, A. R., Cohen, A., Dahmen, W., and DeVore, R. A. (2008). Approximation and learning by greedy algorithms. *Ann. Statist.*, 36(1):64–94.

Bashtannyk, D. M. and Hyndman, R. J. (2001). Bandwidth selection for kernel conditional density estimation. *Comput. Statist. Data Anal.*, 36(3):279–298.

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

Beaumont, M., Zhang, W., and Balding, D. (2002). Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.

Belomestny, D. (2011). Statistical inference for time-changed Lévy processes via composite characteristic function estimation. *Ann. Statist.*, 39(4):2205–2242.

Beran, R. J. (1968). Testing for uniformity on a compact homogeneous space. *J. Appl. Probability*, 5:177–195.

Bertin, K. and Lecué, G. (2008). Selection of variables and dimension reduction in high-dimensional non-parametric regression. *Electron. J. Stat.*, 2:1224–1241.

Biau, G., Cérou, F., and Guyader, A. (2012). New insights into approximate bayesian computation. *arXiv:1207.6461*.

Birgé, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete*, 65(2):181–237.

Birgé, L. (2013). Robust tests for model selection. In *From probability to statistics and back: high-dimensional models and processes*, volume 9 of *Inst. Math. Stat. (IMS) Collect.*, pages 47–64. Inst. Math. Statist., Beachwood, OH.

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

Bissantz, N., Claeskens, G., Holzmann, H., and Munk, A. (2009). Testing for lack of fit in inverse regression—with applications to biophotonic imaging. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 71(1):25–48.

Bissantz, N., Dümbgen, L., Holzmann, H., and Munk, A. (2007). Non-parametric confidence bands in deconvolution density estimation. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 69(3):483–506.

Blum, M. (2010). Approximate bayesian computation: a nonparametric perspective. *Journal of the American Statistical Association*, 105(491):1178–1187.

Bouaziz, O. and Lopez, O. (2010). Conditional density estimation in a censored single-index regression model. *Bernoulli*, 16(2):514–542.

Bücher, A. and Vetter, M. (2013). Nonparametric inference on Lévy measures and copulas. *Ann. Statist.*, 41(3):1485–1515.

Butucea, C. (2001). Exact adaptive pointwise estimation on Sobolev classes of densities. *ESAIM Probab. Statist.*, 5:1–31 (electronic).

Butucea, C. (2004). Deconvolution of supersmooth densities with smooth noise. *Canad. J. Statist.*, 32(2):181–192.

Butucea, C. (2007). Goodness-of-fit testing and quadratic functional estimation from indirect observations. *Ann. Statist.*, 35(5):1907–1930.

Butucea, C. and Comte, F. (2009). Adaptive estimation of linear functionals in the convolution model and applications. *Bernoulli*, 15(1):69–98.

Butucea, C. and Matias, C. (2005). Minimax estimation of the noise level and of the deconvolution density in a semiparametric convolution model. *Bernoulli*, 11(2):309–340.

Butucea, C., Matias, C., and Pouet, C. (2008). Adaptivity in convolution models with partially known noise distribution. *Electron. J. Stat.*, 2:897–915.

Butucea, C., Matias, C., and Pouet, C. (2009). Adaptive goodness-of-fit testing from indirect observations. *Ann. Inst. Henri Poincaré Probab. Stat.*, 45(2):352–372.

Butucea, C. and Tsybakov, A. B. (2007). Sharp optimality in density deconvolution with dominating bias. I. *Teor. Veroyatn. Primen.*, 52(1):111–128.

Butucea, C. and Tsybakov, A. B. (2008). Sharp optimality in density deconvolution with dominating bias. II. *Theory Probab. Appl.*, 52(2):237–249.

Cai, Z. W. (1991). Strong consistency and rates for recursive nonparametric conditional probability density estimates under (α, β)-mixing conditions. *Stochastic Process. Appl.*, 38(2):323–333.

Caillerie, C. and Michel, B. (2011). Model selection for simplicial approximation. *Found. Comput. Math.*, 11(6):707–731.

Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in hidden Markov models*. Springer Series in Statistics. Springer, New York.

Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.*, 83(404):1184–1186.

Cator, E. A. (2001). Deconvolution with arbitrarily smooth kernels. *Statist. Probab. Lett.*, 54(2):205–214.

Cavalier, L. and Raimondo, M. (2007). Wavelet deconvolution with noisy eigenvalues. *IEEE Trans. Signal Process.*, 55(6, part 1):2414–2424.

Chagny, G. (2013). Warped bases for conditional density estimation. *Math. Methods Statist.*, 22(4):253–282.

Chagny, G. and Roche, A. (2014). Adaptive and minimax estimation of the cumulative distribution function given a functional covariate. *Electron. J. Stat.*, 8(2):2352–2404.

Chen, X., Linton, O., and Robinson, P. M. (2001). The estimation of conditional densities. *LSE STICERD Research Paper No. EM/2001/415*.

Chernozhukov, V., Fernández-Val, I., and Galichon, A. (2009). Improving point and interval estimators of monotone functions by rearrangement. *Biometrika*, 96(3):559–575.

Chesneau, C. (2011). Wavelet density estimators for the deconvolution of a component from a mixture. *Sankhya A*, 73(2):245–266.

Chesneau, C., Comte, F., Mabon, G., and Navarro, F. (2015). Estimation of convolution in the model with noise. *Preprint*.

Clémençon, S. (2000). Adaptive estimation of the transition density of a regular Markov chain. *Math. Methods Statist.*, 9(4):323–357.

Cohen, S. X. and Le Pennec, E. (2013). Partition-based conditional density estimation. *ESAIM Probab. Stat.*, 17:672–697.

Comminges, L. and Dalalyan, A. S. (2012). Tight conditions for consistency of variable selection in the context of high dimensionality. *Ann. Statist.*, 40(5):2667–2696.

Comte, F. and Genon-Catalot, V. (2010). Nonparametric adaptive estimation for pure jump Lévy processes. *Ann. Inst. Henri Poincaré Probab. Stat.*, 46(3):595–617.

Comte, F., Rozenholc, Y., and Taupin, M.-L. (2006). Penalized contrast estimator for adaptive density deconvolution. *Canad. J. Statist.*, 34(3):431–452.

Comte, F. and Samson, A. (2012). Nonparametric estimation of random-effects densities in linear mixed-effects model. *J. Nonparametr. Stat.*, 24(4):951–975.

Comte, F., Samson, A., and Stirnemann, J. J. (2014). Deconvolution estimation of onset of pregnancy with replicate observations. *Scand. J. Stat.*, 41(2):325–345.

Dattner, I., Goldenshluger, A., and Juditsky, A. (2011). On deconvolution of distribution functions. *Ann. Statist.*, 39(5):2477–2501.

Dattner, I. and Reiser, B. (2013). Estimation of distribution functions in measurement error models. *J. Statist. Plann. Inference*, 143(3):479–493.

Dattner, I., Reiß, M., and Trabs, M. (2013). Adaptive quantile estimation in deconvolution with unknown error distribution. *arXiv:1303.1698*.

Davis, G., Mallat, S., and Avellaneda, M. (1997). Adaptive greedy approximations. *Constr. Approx.*, 13(1):57–98.

De Gooijer, J. G. and Zerom, D. (2003). On conditional density estimation. *Statist. Neerlandica*, 57(2):159–176.

Dedecker, J., Fischer, A., and Michel, B. (2015). Improved rates for Wasserstein deconvolution with ordinary smooth error in dimension one. *Electron. J. Stat.*, 9:234–265.

Dedecker, J. and Michel, B. (2013). Minimax rates of convergence for Wasserstein deconvolution with supersmooth errors in any dimension. *J. Multivariate Anal.*, 122:278–291.

Delaigle, A. and Gijbels, I. (2004). Bootstrap bandwidth selection in kernel density estimation from a contaminated sample. *Ann. Inst. Statist. Math.*, 56(1):19–47.

Delaigle, A., Hall, P., and Meister, A. (2008). On deconvolution with repeated measurements. *Ann. Statist.*, 36(2):665–685.

Devroye, L. (1989). Consistent deconvolution in density estimation. *Canad. J. Statist.*, 17(2):235–239.

Diggle, P. J. and Hall, P. (1993). A Fourier approach to nonparametric deconvolution of a density estimate. *J. Roy. Statist. Soc. Ser. B*, 55(2):523–531.

Donoho, D. L. (1997). CART and best-ortho-basis: a connection. *Ann. Statist.*, 25(5):1870–1911.

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(2):925–950.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. (2014). Privacy aware learning. *J. ACM*, 61(6):Art. 38, 57.

Dumont, T. and Le Corff, S. (2012). Nonparametric regression on hidden phi-mixing variables: identifiability and consistency of a pseudo-likelihood based estimation procedure . *arXiv:1209.0633*.

Duval, C. (2012). Adaptive wavelet estimation of a compound Poisson process. *arXiv:1203.3135*.

Efromovich, S. (1997). Density estimation for the case of supersmooth measurement error. *J. Amer. Statist. Assoc.*, 92(438):526–535.

Efromovich, S. (2007). Conditional density estimation in a regression setting. *Ann. Statist.*, 35(6):2504–2535.

Efromovich, S. (2010a). Dimension reduction and adaptation in conditional density estimation. *J. Amer. Statist. Assoc.*, 105(490):761–774.

Efromovich, S. (2010b). Oracle inequality for conditional density estimation and an actuarial example. *Ann. Inst. Statist. Math.*, 62(2):249–275.

Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.*, 19(3):1257–1272.

Fan, J., Yao, Q., and Tong, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, 83(1):189–206.

Fan, J. and Yim, T. H. (2004). A crossvalidation method for estimating conditional densities. *Biometrika*, 91(4):819–834.

Faugeras, O. P. (2009). A quantile-copula approach to conditional density estimation. *J. Multivariate Anal.*, 100(9):2083–2099.

Faÿ, G., Delabrouille, J., Kerkyacharian, G., and Picard, D. (2012). Testing the isotropy of high energy cosmic rays using spherical needlets. *arXiv:1107.5658v2*.

Ferraty, F., Laksaci, A., and Vieu, P. (2006). Estimating some characteristics of the conditional distribution in nonparametric functional models. *Stat. Inference Stoch. Process.*, 9(1):47–76.

Feuerverger, A., Kim, P. T., and Sun, J. (2008). On optimal uniform deconvolution. *J. Stat. Theory Pract.*, 2(3):433–451.

Figueroa-López, J. E. (2009a). Nonparametric estimation of Lévy models based on discrete-sampling. In *Optimality*, volume 57 of *IMS Lecture Notes Monogr. Ser.*, pages 117–146. Inst. Math. Statist., Beachwood, OH.

Figueroa-López, J. E. (2009b). Nonparametric estimation of time-changed Lévy models under high-frequency data. *Adv. Appl. Probab.*, 41(4):1161–1188.

Figueroa-López, J. E. (2011). Sieve-based confidence intervals and bands for Lévy densities. *Bernoulli*, 17(2):643–670.

Figueroa-López, J. E. and Houdré, C. (2006). Risk bounds for the non-parametric estimation of Lévy processes. In *High dimensional probability*, volume 51 of *IMS Lecture Notes Monogr. Ser.*, pages 96–116. Inst. Math. Statist., Beachwood, OH.

Gassiat, É., Cleynen, A., and Robin, S. (2015). Finite state space non parametric hidden Markov models are in general identifiable. *Stat. Comput.* To appear.

Gassiat, É. and Rousseau, J. (2015). Non parametric finite translation hidden Markov models and extensions. *Bernoulli*. To appear.

Giné, E. (1975). Invariant tests for uniformity on compact Riemannian manifolds based on Sobolev norms. *Ann. Statist.*, 3(6):1243–1266.

Giné, E. and Nickl, R. (2010). Confidence bands in density estimation. *Ann. Statist.*, 38(2):1122–1170.

Goldenshluger, A. (1999). On pointwise adaptive nonparametric deconvolution. *Bernoulli*, 5(5):907–925.

Goldenshluger, A. (2002). Density deconvolution in the circular structural model. *J. Multivariate Anal.*, 81(2):360–375.

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

Goldenshluger, A. and Lepski, O. (2009). Structural adaptation via $\mathbb{L}_p$-norm oracle inequalities. *Probab. Theory Related Fields*, 143(1-2):41–71.

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

Goldenshluger, A. and Lepski, O. (2014). On adaptive minimax density estimation on R^d. *Probab. Theory Related Fields*, 159(3-4):479–543.

Goldenshluger, A., Tsybakov, A., and Zeevi, A. (2006). Optimal change-point estimation from indirect observations. *Ann. Statist.*, 34(1):350–372.

Goldenshluger, A. V. and Lepski, O. V. (2013). General selection rule from a family of linear estimators. *Theory Probab. Appl.*, 57(2):209–226.

Gugushvili, S. (2012). Nonparametric inference for discretely sampled Lévy processes. *Ann. Inst. Henri Poincaré Probab. Stat.*, 48(1):282–307.

Györfi, L. and Kohler, M. (2007). Nonparametric estimation of conditional distributions. *IEEE Trans. Inform. Theory*, 53(5):1872–1879.

Hall, P. and Lahiri, S. N. (2008). Estimation of distributions, moments and quantiles in deconvolution problems. *Ann. Statist.*, 36(5):2110–2134.

Hall, P. and Meister, A. (2007). A ridge-parameter approach to deconvolution. *Ann. Statist.*, 35(4):1535–1558.

Hall, P., Racine, J., and Li, Q. (2004). Cross-validation and the estimation of conditional probability densities. *J. Amer. Statist. Assoc.*, 99(468):1015–1026.

Hall, P., Wolff, R. C. L., and Yao, Q. (1999). Methods for estimating a conditional distribution function. *J. Amer. Statist. Assoc.*, 94(445):154–163.

Hall, P. and Yao, Q. (2005). Approximating conditional distribution functions using dimension reduction. *Ann. Statist.*, 33(3):1404–1421.

Hansen, N. (2006). The CMA evolution strategy: a comparing review. In *Towards a new evolutionary computation*, pages 75–102. Springer.

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

Healy, Jr., D. M., Hendriks, H., and Kim, P. T. (1998). Spherical deconvolution. *J. Multivariate Anal.*, 67(1):1–22.

Hesse, C. H. (1999). Data-driven deconvolution. *J. Nonparametr. Statist.*, 10(4):343–373.

Hoffmann, M. and Reiß, M. (2008). Nonlinear estimation for linear inverse problems with error in the operator. *Ann. Statist.*, 36(1):310–336.

Holmes, M. P., Gray, A. G., and Isbell, C. L. (2012). Fast nonparametric conditional density estimation. *arXiv:1206.5278*.

Holzmann, H., Bissantz, N., and Munk, A. (2007). Density testing in a contaminated sample. *J. Multivariate Anal.*, 98(1):57–75.

Hsu, D., Kakade, S. M., and Zhang, T. (2012). A spectral algorithm for learning hidden Markov models. *J. Comput. System Sci.*, 78(5):1460–1480.

Huckemann, S. F., Kim, P. T., Koo, J.-Y., and Munk, A. (2010). Möbius deconvolution on the hyperbolic plane with application to impedance density estimation. *Ann. Statist.*, 38(4):2465–2498.

Hyndman, R. J., Bashtannyk, D. M., and Grunwald, G. K. (1996). Estimating and visualizing conditional densities. *J. Comput. Graph. Statist.*, 5(4):315–336.

Hyndman, R. J. and Yao, Q. (2002). Nonparametric estimation and symmetry tests for conditional density functions. *J. Nonparametr. Stat.*, 14(3):259–278.

Ingster, Y. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives. I, II, III. *Math. Methods Statist.*, 2(1):85–114.

Ingster, Y. and Sapatinas, T. (2009). Minimax goodness-of-fit testing in multivariate nonparametric regression. *Math. Methods Statist.*, 18:241–269.

Jeon, J. and Taylor, J. W. (2012). Using conditional kernel density estimation for wind power density forecasting. *J. Amer. Statist. Assoc.*, 107(497):66–79.

Johannes, J. (2009). Deconvolution with unknown error distribution. *Ann. Statist.*, 37(5A):2301–2323.

Johannes, J. and Schwarz, M. (2013). Adaptive circular deconvolution by model selection under unknown error distribution. *Bernoulli*, 19(5A):1576–1611.

Jongbloed, G. and van der Meulen, F. H. (2006). Parametric estimation for subordinators and induced OU processes. *Scand. J. Statist.*, 33(4):825–847.

Kappus, J. (2012). Nonparametric adaptive estimation of linear functionals for low frequency observed Lévy processes. *SFB 649 discussion paper, No. 2012-016*.

Kappus, J. and Mabon, G. (2014). Adaptive density estimation in deconvolution problems with unknown error distribution. *Electron. J. Stat.*, 8(2):2879–2904.

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

Kerkyacharian, G., Pham Ngoc, T. M., and Picard, D. (2011). Localized spherical deconvolution. *Ann. Statist.*, 39(2):1042–1068.

Kim, P. T. and Koo, J.-Y. (2002). Optimal spherical deconvolution. *J. Multivariate Anal.*, 80(1):21–42.

Kim, P. T., Koo, J.-Y., and Park, H. J. (2004). Sharp minimaxity and spherical deconvolution for super-smooth error distributions. *J. Multivariate Anal.*, 90(2):384–392.

Kim, P. T. and Richards, D. S. P. (2001). Deconvolution density estimation on compact Lie groups. In *Algebraic methods in statistics and probability (Notre Dame, IN, 2000)*, volume 287 of *Contemp. Math.*, pages 155–171. Amer. Math. Soc., Providence, RI.

Kim, Peter, S., Koo, J.-Y., and Pham Ngoc, T. M. (2015). Supersmooth Testing on the Sphere over Analytic Classes. *Preprint*.

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

Klemelä, J. (2009). Multivariate histograms with data-dependent partitions. *Statist. Sinica*, 19(1):159–176.

Koo, J.-Y. (1999). Logspline deconvolution in Besov space. *Scand. J. Statist.*, 26(1):73–86.

Lafferty, J. and Wasserman, L. (2007). Sparse nonparametric density estimation in high dimensions using the rodeo. *In Proc. 11th International Conf. on Artificial Intelligence and Statistics (AISTATS'07)*.

Lafferty, J. and Wasserman, L. (2008). Rodeo: sparse, greedy nonparametric regression. *Ann. Statist.*, 36(1):28–63.

Laurent, B., Loubes, J.-M., and Marteau, C. (2011). Testing inverse problems: a direct or an indirect problem? *J. Statist. Plann. Inference*, 141(5):1849–1861.

Laurent, B., Ludena, C., and Prieur, C. (2008). Adaptive estimation of linear functionals by model selection. *Electronic Journal of Statistics*, 2:993–1020.

Lebarbier, É. (2005). Detecting multiple change-points in the mean of gaussian process by model selection. *Signal processing*, 85(4):717–736.

Lepski, O. (2014). Adaptive estimation over anisotropic functional classes via oracle approach. *arXiv:1405.4504*.

Li, R. and Liu, Y. (2014). Wavelet optimal estimations for a density with some additive noises. *Appl. Comput. Harmon. Anal.*, 36(3):416–433.

Li, T. and Vuong, Q. (1998). Nonparametric estimation of the measurement error model using multiple indicators. *J. Multivariate Anal.*, 65(2):139–165.

Liu, M. C. and Taylor, R. L. (1989). A consistent nonparametric density estimator for the deconvolution problem. *Canad. J. Statist.*, 17(4):427–438.

Loubes, J. M. and Marteau, C. (2014). Goodness-of-fit testing strategies from indirect observations. *J. Nonparametr. Stat.*, 26(1):85–99.

Lounici, K. and Nickl, R. (2011). Global uniform risk bounds for wavelet deconvolution estimators. *Ann. Statist.*, 39(1):201–231.

Mabon, G. (2015). Adaptive deconvolution on the nonnegative real line. *Preprint*.

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. (2012). Approximate bayesian computation methods. *Statistics and Computing*, 22(6):1167–1180.

Masry, E. (1989). Nonparametric estimation of conditional probability densities and expectations of stationary processes: strong consistency and rates. *Stochastic Process. Appl.*, 32(1):109–127.

Masry, E. (1991). Multivariate probability density deconvolution for stationary random processes. *IEEE Trans. Inform. Theory*, 37(4):1105–1115.

Masry, E. (2003). Deconvolving multivariate kernel density estimates from contaminated associated observations. *IEEE Trans. Inform. Theory*, 49(11):2941–2952.

Masry, E. and Rice, J. A. (1992). Gaussian deconvolution via differentiation. *Canad. J. Statist.*, 20(1):9–21.

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

Matias, C. (2002). Semiparametric deconvolution with unknown noise variance. *ESAIM Probab. Statist.*, 6:271–292 (electronic). New directions in time series analysis (Luminy, 2001).

Maugis, C. and Michel, B. (2011). A non asymptotic penalized criterion for Gaussian mixture model selection. *ESAIM Probab. Stat.*, 15:41–68.

Meister, A. (2004). On the effect of misspecifying the error density in a deconvolution problem. *Canad. J. Statist.*, 32(4):439–449.

Meister, A. (2006). Density estimation with normal measurement error with unknown variance. *Statist. Sinica*, 16(1):195–211.

Meister, A. (2007). Deconvolving compactly supported densities. *Math. Methods Statist.*, 16(1):63–76.

Meister, A. (2008). Deconvolution from Fourier-oscillating error densities under decay and smoothness restrictions. *Inverse Problems*, 24(1):015003, 14.

Meister, A. (2009). *Deconvolution problems in nonparametric statistics*, volume 193 of *Lecture Notes in Statistics*. Springer-Verlag, Berlin.

Meister, A., Stadtmüller, U., and Wagner, C. (2010). Density deconvolution in a two-level heteroscedastic model with unknown error density. *Electron. J. Stat.*, 4:36–57.

Mendelsohn, J. and Rice, J. (1982). Deconvolution of microfluorometric histograms with B-splines. *Journal of the American Statistical Association*, 77(380):748–753.

Neumann, M. H. (1997a). On the effect of estimating the error density in nonparametric deconvolution. *J. Nonparametr. Statist.*, 7(4):307–330.

Neumann, M. H. (1997b). Optimal change-point estimation in inverse problems. *Scand. J. Statist.*, 24(4):503–521.

Neumann, M. H. (2007). Deconvolution from panel data with unknown error distribution. *J. Multivariate Anal.*, 98(10):1955–1968.

Neumann, M. H. and Reiß, M. (2009). Nonparametric estimation for Lévy processes from low-frequency observations. *Bernoulli*, 15(1):223–248.

Nickl, R. and Reiß, M. (2012). A Donsker theorem for Lévy measures. *J. Funct. Anal.*, 263(10):3306–3332.

Odiachi, P. and Prieve, D. (2004). Removing the effects of additive noise in tirm measurements. *J. Colloid Interface Sci.*, 270:113–122.

Paulin, D. (2014). Concentration inequalities for Markov chains by Marton couplings. *arXiv:1212.2015v3*.

Pensky, M. and Vidakovic, B. (1999). Adaptive wavelet estimator for nonparametric density deconvolution. *Ann. Statist.*, 27(6):2033–2053.

Plancade, S. (2013). Adaptive estimation of the conditional cumulative distribution function from current status data. *J. Statist. Plann. Inference*, 143(9):1466–1485.

Plancade, S., Rozenholc, Y., and Lund, E. (2012). Generalization of the normal-exponential model: exploration of a more accurate parametrisation for the signal distribution on illumina beadarrays. *BMC bioinformatics*, 13(1):329.

Quashnock, J. M. and Lamb, D. Q. (1993). Evidence for the galactic origin of gamma-ray bursts. *M.N.R.A.S.*, 265:L45–L50.

Raskutti, G., Wainwright, M. J., and Yu, B. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *J. Mach. Learn. Res.*, 13:389–427.

Resti, Y., Ismail, N., and Jaaman, S. H. (2012). Mathematical modelling for claim severities using normal and t copulas. *Int. J. Appl. Math. Stat.*, 27(3):8–19.

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

Rosenblatt, M. (1969). Conditional probability density and regression estimators. In *Multivariate Analysis, II (Proc. Second Internat. Sympos., Dayton, Ohio, 1968)*, pages 25–31. Academic Press, New York.

Roussas, G. G. (1969). Nonparametric estimation of the transition distribution function of a Markov process. *Ann. Math. Statist.*, 40:1386–1400.

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

Saumard, A. (2012). Optimal upper and lower bounds for the true and empirical excess risks in heteroscedastic least-squares regression. *Electron. J. Stat.*, 6:579–655.

Schmidt-Hieber, J., Munk, A., and Dümbgen, L. (2013). Multiscale methods for shape constraints in deconvolution: confidence statements for qualitative features. *Ann. Statist.*, 41(3):1299–1328.

Schwarz, M. and Van Bellegem, S. (2010). Consistent density deconvolution under partially known error distribution. *Statist. Probab. Lett.*, 80(3-4):236–241.

Scricciolo, C. (2015). Empirical bayes conditional density estimation. *arXiv:1501.01847*.

Söhl, J. and Trabs, M. (2012). A uniform central limit theorem and efficiency for deconvolution estimators. *Electron. J. Stat.*, 6:2486–2518.

Song, L., Anandkumar, A., Dai, B., and Xie, B. (2014). Nonparametric estimation of multi-view latent variable models. In *ICML*.

Spokoiny, V. G. (1996). Adaptive hypothesis testing using wavelets. *Ann. Statist.*, 24(6):2477–2498.

Stefanski, L. and Carroll, R. J. (1990). Deconvoluting kernel density estimators. *Statistics*, 21(2):169–184.

Stefanski, L. A. and Cook, J. R. (1995). Simulation-extrapolation: the measurement error jackknife. *J. Amer. Statist. Assoc.*, 90(432):1247–1256.

Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *Ann. Statist.*, 22(1):118–184. With discussion by Andreas Buja and Trevor Hastie and a rejoinder by the author.

Stute, W. (1986). On almost sure convergence of conditional empirical distribution functions. *Ann. Probab.*, 14(3):891–901.

Takeuchi, I., Nomura, K., and Kanamori, T. (2009). Nonparametric conditional density estimation using piecewise-linear solution path of kernel quantile regression. *Neural Comput.*, 21(2):533–559.

The Pierre AUGER Collaboration (2010). Update on the correlation of the highest energy cosmic rays with nearby extragalactic matter. *Astroparticle Physics*, 34:314–326.

Triebel, H. (2006). *Theory of function spaces. III*, volume 100 of *Monographs in Mathematics*. Birkhäuser Verlag, Basel.

Tropp, J. A. (2004). Greed is good: algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory*, 50(10):2231–2242.

Ueltzhöfer, F. A. and Klüppelberg, C. (2011). An oracle inequality for penalised projection estimation of lévy densities from high-frequency observations. *Journal of Nonparametric Statistics*, 23(4):967–989.

van Es, A. J. and Kok, A. R. (1998). Simple kernel estimators for certain nonparametric deconvolution problems. *Statist. Probab. Lett.*, 39(2):151–160.

van Es, B. (2011). Combining kernel estimators in the uniform deconvolution problem. *Stat. Neerl.*, 65(3):275–296.

van Es, B., Gugushvili, S., and Spreij, P. (2007). A kernel type nonparametric density estimator for decompounding. *Bernoulli*, 13(3):672–694.

Vedrenne, G. and Atteia, J.-L. (2009). *Gamma-Ray Bursts: The brightest explosions in the Universe*. Springer/Praxis Books.

Verzelen, N. (2010). High-dimensional Gaussian model selection on a Gaussian design. *Ann. Inst. Henri Poincaré Probab. Stat.*, 46(2):480–524.

Viennet, G. (1997). Inequalities for absolutely regular sequences: application to density estimation. *Probab. Theory Related Fields*, 107(4):467–492.

Wang, X.-F. and Ye, D. (2015). Conditional density estimation in measurement error problems. *Journal of Multivariate Analysis*, 133(C):38–50.

Watson, G. S. (1965). Equatorial distributions on a sphere. *Biometrika*, 52:193–201.

Watteel, R. N. and Kulperger, R. J. (2003). Nonparametric estimation of the canonical measure for infinitely divisible distributions. *J. Stat. Comput. Simul.*, 73(7):525–542.

Youndjé, É. and Wells, M. T. (2008). Optimal bandwidth selection for multivariate kernel deconvolution density estimation. *TEST*, 17(1):138–162.

Zhang, C.-H. (1990). Fourier methods for estimating mixing densities and distributions. *Ann. Statist.*, 18(2):806–831.