Adaptive nonparametric estimation of a component density in a two-class mixture model

Gaëlle Chagny[*], Antoine Channarond[†], Van Hà Hoang[‡], Angelina Roche[§]

July 30, 2020

Abstract

A two-class mixture model, where the density of one of the components is known, is considered. We address the issue of the nonparametric adaptive estimation of the unknown probability density of the second component. We propose a randomly weighted kernel estimator with a fully data-driven bandwidth selection method, in the spirit of the Goldenshluger and Lepski method. An oracle-type inequality for the pointwise quadratic risk is derived as well as convergence rates over Hölder smoothness classes. The theoretical results are illustrated by numerical simulations.

1 Introduction

The following mixture model with two components:

$$g(x) = \theta + (1-\theta)f(x), \quad \forall x \in [0,1], \tag{1}$$

where the mixing proportion $\theta \in (0,1)$ and the probability density function f on $[0,1]$ are unknown, is considered in this article. It is assumed that n independent and identically distributed (*i.i.d.* in the sequel) random variables $X_1, \ldots, X_n$ drawn from density g are observed. The main goal is to construct an adaptive estimator of the nonparametric component f and to provide non-asymptotic upper bounds of the pointwise risk. As an intermediate step, the estimation of the parametric component θ is addressed as well.

Model (1) appears in some statistical settings, robust estimation and multiple testing among others. The one chosen in the present article, as described above, comes from the multiple testing framework, where a large number n of independent hypotheses tests are performed simultaneously. p-values $X_1, \ldots, X_n$ generated by these tests can be modeled by (1). Indeed these are uniformly distributed on $[0,1]$ under null hypotheses while their distribution under alternative hypotheses, corresponding to f, is unknown. The unknown parameter θ is the asymptotic proportion of true null hypotheses. It can be needed to estimate f, especially to evaluate and control different types of expected errors of the testing procedure, which is a major issue in this context. See for instance Genovese and Wassermann [15], Storey [28], Langaas *et al.* [20], Robin *et al.* [26], Strimmer [29], Nguyen and Matias [23], and more fundamentally, Benjamini *et al.* [1] and Efron *et al.* [14].

In the setting of robust estimation, different from the multiple testing one, model (1) can be thought of as a contamination model, where the unknown distribution of interest f is contaminated by the uniform distribution on $[0,1]$ with the proportion θ. This is a very specific case of the Huber contamination model [18]. The statistical task considered consists in robustly estimating f from contaminated observations $X_1, \ldots, X_n$. But unlike our setting, the contamination distribution is not necessarily known while the contamination proportion θ is assumed to be known and the theoretical investigations aim at providing minimax rates as functions of both n and θ. See for instance the preprint of Liu and Gao [22], which addresses pointwise estimation in this framework.

[*]LMRS, UMR CNRS 6085, Université de Rouen Normandie, gaelle.chagny@univ-rouen.fr
[†]LMRS, UMR CNRS 6085, Université de Rouen Normandie, antoine.channarond@univ-rouen.fr
[‡]LMRS, UMR CNRS 6085, Université de Rouen Normandie, van-ha.hoang@univ-rouen.fr
[§]CEREMADE, UMR CNRS 7534, Université Paris Dauphine, roche@ceremade.dauphine.fr

Back to the setting of multiple testing, the estimation of f in model (1) has been addressed in several works. Langaas *et al.* [20] proposed a Grenander density estimator for f, based on a nonparametric maximum likelihood approach, under the assumption that f belongs to the set of decreasing densities on $[0,1]$. Following a similar approach, Strimmer [29] also proposed a modified Grenander strategy to estimate f. However, the two aforementioned papers do not investigate theoretical features of the proposed estimators. Robin *et al.* [26] and Nguyen and Matias[23] proposed a randomly weighted kernel estimator of f, where the weights are estimators of the posterior probabilities of the mixture model, that is, the probabilities of each individual i being in the nonparametric component given the observation X_i. [26] proposes an EM-like algorithm, and proves the convergence to an unique solution of the iterative procedure, but they do not provide any asymptotic property of the estimator. Note that their model $g(x) = \theta\phi(x) + (1-\theta)f(x)$, where ϕ is a known density, is slightly more general, but our procedure is also suitable for this model under some assumptions on ϕ. Besides, [23] achieves a nonparametric rate of convergence $n^{-2\beta/(2\beta+1)}$ for their estimator, where β is the smoothness of the unknown density f. However, their estimation procedure is not adaptive since the choice of their optimal bandwidth still depends on β.

In the present work, a new randomly weighted kernel estimator is proposed. Unlike the usual approach in mixture models, the weights of the estimator are not estimates of the posterior probabilities. A function w is derived instead such that $f(x) = w(\theta, g(x))g(x)$, for all $\theta, x \in [0,1]$. This kind of equation, linking the target distribution (one of the conditional distribution given hidden variables) to the distribution of observed variables, is remarkable in the framework of mixture models. It is a key idea of our approach, since it implies a crucial equation for controlling the bias term of the risk, see Subsection 2.1 for more details. Thus oracle weights are defined by $w(\theta, g(X_i))$, $i = 1, \dots, n$, but g and θ are unknown. These oracle weights are estimated by plug-in, using preliminary estimators of g and θ, based on an additional sample $X_{n+1}, \dots, X_{2n}$. Note that procedures of [23] and [26] actually require preliminary estimates of g and θ as well, but they do not deal with possible biases caused by the multiple use of the same observations in the estimates of θ, g and f.

Furthermore a data-driven bandwidth selection rule is also constructed in this paper, using the Goldenshluger and Lespki (GL) approach [17], which has been applied in various contexts, see for instance, Comte *et al.* [11], Comte and Lacour [9], Doumic *et al.* [13], Reynaud-Bouret *et al.* [25] who apply GL method in kernel density estimation, and Bertin *et al.* [3], Chagny [6], Chichignoud *et al.* [7] or Comte and Rebafka [12]. Our selection rule is then adaptive to unknown smoothness of the target function, which is new in this context. The main original results derived in this paper are the oracle-type inequality in Theorem 1, and the rates of convergence over Hölder classes, which are adapted to the control of pointwise risk of kernel estimators, in Corollary 1.

Some assumptions on the preliminary estimators for g and θ are needed to prove the results on the estimator of f; this paper also provides estimators of g and θ which satisfy these assumptions. The choice of a nice estimator for θ requires identifiability of the model (1). g being given, the couple (θ, f) such that $g = \theta + (1-\theta)f$ is uniquely determined under additional assumptions on f (in particular monotonicity and zero set of f), see a review about this issue in Section 1.1 in Nguyen and Matias [24]. Nonetheless, note that these additional assumptions on f are not needed to obtain the results on the nonparametric estimation procedure of f.

The paper is organized as follows. Our randomly weighted estimator of f is constructed in Section 2.1. Assumptions on f and on preliminary estimators of g and θ required for proving the theoretical results are in this section too. In Section 2, a bias-variance decomposition for the pointwise risk of the estimator of f is given as well as the convergence rate of the kernel estimator with a fixed bandwidth. In Section 3, an oracle inequality which justifies our adaptive estimation procedure. Construction of the preliminary estimators of g and θ are to be found in Section 4. Numerical results illustrate the theoretical results in Section 5. Proofs of theorems, propositions and technical lemmas are postponed to Section 6.

2 Collection of kernel estimators for the target density

In this section, a family of kernel estimators for the density function f based on a sample $(X_i)_{i=1,\dots,n}$ of i.i.d. variables with distribution g is defined. It is assumed that two preliminary estimators $\tilde{\theta}_n$ of the mixing proportion θ and $\hat{g}$ of the mixture density g are available, and defined from an additional sample $(X_i)_{i=n+1,\dots,2n}$ of independent variables also drawn from g but independent of the first sample

$(X_i)_{i=1,\dots,n}$. The definition of these preliminary estimates is the subject of Section 4.

2.1 Construction of the estimators

To define estimators for f, the challenge is that observations $X_1, \dots, X_n$ are not drawn from f but from the mixture density g. Hence the density f cannot be estimated directly by a classical kernel density estimator. We will thus build weighted kernel estimates, using a methodology inspired for example by Comte and Rebafka [12]. The starting point is the following lemma whose proof is straightforward.

Lemma 1. *Let X be a random variable from the mixture density g defined by (1) and Y be an (unobservable) random variable from the component density f. Then for any measurable bounded function φ we have*

$$\mathbb{E}\big[\varphi(Y)\big] = \mathbb{E}\big[w(\theta, g(X))\varphi(X)\big], \tag{2}$$

with

$$w(\theta, g(x)) := \frac{1}{1-\theta}\left(1 - \frac{\theta}{g(x)}\right), \ x \in [0,1].$$

This result will be used as follows. Let $K : \mathbb{R} \to \mathbb{R}$ be a kernel function, that is an integrable function such that $\int_{\mathbb{R}} K(x)dx = 1$ and $\int_{\mathbb{R}} K^2(x)dx < +\infty$. For any $h > 0$, let $K_h(\cdot) = K(\cdot/h)/h$. Then the choice $\varphi(\cdot) = K_h(x - \cdot)$ in Lemma 1 leads to

$$\mathbb{E}\big[K_h(x - Y_1)\big] = \mathbb{E}\big[w(\theta, g(X_1))K_h(x - X_1)\big],$$

where Y is drawn from g. Thus,

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} w(\tilde{\theta}_n, \hat{g}(X_i))K_h(x - X_i), \ x \in [0,1], \tag{3}$$

is well-suited to estimate f, with

$$w(\tilde{\theta}_n, \hat{g}(X_i)) = \frac{1}{1-\tilde{\theta}_n}\left(1 - \frac{\tilde{\theta}_n}{\hat{g}(X_i)}\right), \quad i = 1, \dots, n.$$

Therefore, $\hat{f}_h$ is a randomly weighted kernel estimator of f. However, the total sum of the weights may not equal 1, in comparison with the estimators proposed in Nguyen and Matias [23] and Robin *et al.* [26]. The main advantage of our estimate is that, if we replace $\hat{g}$ and $\tilde{\theta}_n$ by their theoretical unknown counterparts g and θ in (3), we obtain, $\mathbb{E}[\hat{f}_h(x)] = K_h \star f(x)$, where $\star$ stands for the convolution product. This relation, classical in nonparametric kernel estimation, is crucial for the study of the bias term in the risk of the estimator.

2.2 Risk bounds of the estimator

Here, we establish upper bounds for the pointwise mean-squared error of the estimator $\hat{f}_h$, defined in (3), with a fixed bandwidth $h > 0$. Our objective is to study the pointwise risk for the estimation of the density f at a point $x_0 \in [0,1]$. Throughout the paper, the kernel K is chosen compactly supported on an interval $[-A, A]$ with A a positive real number. We denote by $V_n(x_0)$ the neighbourhood of x_0 used in the sequel and defined by

$$V_n(x_0) = \left[x_0 - \frac{2A}{\alpha_n}, x_0 + \frac{2A}{\alpha_n}\right],$$

where $(\alpha_n)_n$ is a positive sequence of numbers larger than 1, only depending on n such that $\alpha_n \to +\infty$ as $n \to +\infty$. For any function u on $\mathbb{R}$, and any interval $I \subset \mathbb{R}$, let $\|u\|_{\infty,I} = \sup_{t\in I} |u(t)|$.

The following assumptions will be required for our theoretical results.

(A1) The density f is uniformly bounded on $V_n(x_0)$ for some n: $\|f\|_{\infty,V_n(x_0)} < \infty$.

(A2) The preliminary estimator $\hat{g}$ is bounded away from 0 on $V_n(x_0)$:

$$\hat{\gamma} := \inf_{t \in V_n(x_0)} |\hat{g}(t)| > 0. \tag{4}$$

(A3) The preliminary estimate $\hat{g}$ of g satisfies, for all $\nu > 0$

$$\mathbb{P}\left(\sup_{t \in V_n(x_0)} \left| \frac{\hat{g}(t) - g(t)}{\hat{g}(t)} \right| > \nu \right) \leqslant C_{g,\nu} \exp\left\{ -(\log n)^{3/2} \right\}, \tag{5}$$

with $C_{g,\nu}$ a constant only depending on g and ν.

(A4) The preliminary estimator $\tilde{\theta}_n$ is constructed such that $\tilde{\theta}_n \in [\delta/2, 1-\delta/2]$, for a fixed $\delta \in (0,1)$.

(A5) For any bandwidth $h > 0$, we assume that

$$\alpha_n \leqslant \frac{1}{h} \quad \text{and} \quad \frac{1}{h} \leqslant \frac{\hat{\gamma} n}{\log^3(n)}.$$

(A6) f belongs to the Hölder class of smoothness β and radius $\mathcal{L}$ on $[0,1]$, defined by

$$\Sigma(\beta, \mathcal{L}) = \left\{ \phi : \phi \text{ has } \ell = \lfloor \beta \rfloor \text{ derivatives and } \forall x, y \in [0,1], |\phi^{(\ell)}(x) - \phi^{(\ell)}(y)| < \mathcal{L}|x-y|^{\beta - \ell} \right\},$$

(A7) K is a kernel of order $\ell = \lfloor \beta \rfloor$: $\int_{\mathbb{R}} x^j K(x) dx = 0$ for $1 \leqslant j \leqslant \ell$ and $\int_{\mathbb{R}} |x|^\ell |K(x)| dx < \infty$.

Since $g = \theta + (1-\theta) f$, Assumption **(A1)** implies that $\|g\|_{\infty, V_n(x_0)} < \infty$. The latter condition is needed to control the variance term, among others, of the bias-variance decomposition of the risk. Notice that the density g is automatically bounded from below by a positive constant in our model (1). Assumption **(A2)** is required to bound the term $1/\hat{g}(\cdot)$ that appears in the weight $w(\tilde{\theta}_n, \hat{g}(\cdot))$. Assumption **(A3)** means that the preliminary $\hat{g}$ has to be rather accurate. Assumptions**(A2)** and **(A3)** are also introduced by Bertin *et al.* [2] for conditional density estimation purpose : see (3.2) and (3.3) p.946. The methodology used in our proofs is inspired from their work : the role played by g here corresponds to the role plays by the marginal density of their paper. They have also shown that an estimator of g satisfying these properties can be built, see Theorem 4, p. 14 of [2] and some details at Section 4.1. We also build an estimator $\tilde{\theta}_n$ that satisfied Assumption **(A4)** in Section 4.2. Assumption **(A5)** deals with the order of magnitude of the bandwidths and is also borrowed from [2] (see Assumption (CK) p.947). Assumptions **(A6)** and **(A7)** are classical for kernel density estimation, see [30] or [10]. The index β in Assumption **(A6)** is a measure of the smoothness of the target function. It permits to control the bias term of the bias-variance decomposition of the risk, and thus to derive convergence rates.

We first state an upper bound for the pointwise risk of the estimator $\hat{f}_h$. The proof can be found in Section 6.1.

Proposition 1. *Assume that Assumptions* **(A1)** *to* **(A5)** *are satisfied. Then, for any* $x_0 \in [0,1]$ *and* $\delta \in (0,1)$*, the estimator* $\hat{f}_h$ *defined by* (3) *satisfies*

$$\mathbb{E}\left[\left(\hat{f}_h(x_0) - f(x_0) \right)^2 \right] \leqslant C_1^* \left\{ \|K_h \star f - f\|^2_{\infty, V_n(x_0)} + \frac{1}{\delta^2 \gamma^2 n h} \right\} + \frac{C_2^*}{\delta^6} \mathbb{E}\left[\left| \tilde{\theta}_n - \theta \right|^2 \right] + \frac{C_3^*}{\delta^2 \gamma^2} \mathbb{E}\left[\|\hat{g} - g\|^2_{\infty, V_n(x_0)} \right] + \frac{C_4^*}{\delta^2 n^2}, \tag{6}$$

where C_ℓ^*, $\ell = 1, \ldots, 4$ *are positive constants such that :* C_1^* *depends on* $\|K\|_2$ *and* $\|g\|_{\infty, V_n(x_0)}$, C_2^* *depends on* $\|g\|_{\infty, V_n(x_0)}$ *and* $\|K\|_1$, C_3^* *depends on* $\|K\|_1$ *and* C_4^* *depends on* $\|f\|_{\infty, V_n(x_0)}$, g *and* $\|K\|_\infty$.

Proposition 1 is a bias-variance decomposition of the risk. The first term in the right-hand-side (*r.h.s.* in the sequel) of (6) is a bias term which decreases when the bandwidth h goes to 0 whereas the second one corresponds to the variance term and increases when h goes to 0. There are two additional terms $\mathbb{E}[\|\hat{g} - g\|^2_{\infty, V_n(x_0)}]$ and $\mathbb{E}[|\hat{\theta}_n - \theta|^2]$ in the *r.h.s.* of (6). They are unavoidable since the estimator $\hat{f}_h$ depends on the plug-in estimators $\hat{g}$ and $\tilde{\theta}_n$. The term $C_4^*/(\delta^2 n^2)$ is a remaining term and is also negligible. However, the convergence rate that we derive in Corollary 1 below will prove that these last three terms in (6) are negligible if g and θ are estimated accurately : Section 4 proves that it is possible.

3 Adaptive pointwise estimation

Let $\mathcal{H}_n$ be a finite family of possible bandwidths $h > 0$, whose cardinality is bounded by the sample size n. The best estimator in the collection $(\hat{f}_h)_{h\in\mathcal{H}_n}$ defined in (3) at the point x_0 is the one that have the smallest risk, or similarly, the smallest bias-variance decomposition. But since f is unknown, in practice it is impossible to minimize over $\mathcal{H}_n$ the r.h.s. of inequality (6) in order to select the best estimate. Thus, we propose a data-driven selection, with a rule in the spirit of Goldenshluger and Lepski (GL in the sequel) [17]. The idea is to mimic the bias-variance trade-off for the risk, with empirical counterparts for the unknown quantities. We first estimate the variance term of the trade-off by setting, for any $h \in \mathcal{H}_n$

$$V(x_0, h) = \frac{\kappa \|K\|_1^2 \|K\|_2^2 \|g\|_{\infty,V_n(x_0)}}{\hat{\gamma}^2 nh} \log(n), \tag{7}$$

with $\kappa > 0$ a tuning parameter. The principle of the GL method is then to estimate the bias term $\|K_h \star f - f\|^2_{\infty,V_n(x_0)}$ of $\hat{f}_h(x_0)$ for any $h \in \mathcal{H}_n$ with

$$A(x_0, h) := \max_{h'\in\mathcal{H}_n} \left\{ \left(\hat{f}_{h,h'}(x_0) - \hat{f}_{h'}(x_0)\right)^2 - V(x_0, h') \right\}_+ ,$$

where, for any $h, h' \in \mathcal{H}_n$,

$$\hat{f}_{h,h'}(x_0) = \frac{1}{n}\sum_{i=1}^{n} w(\tilde{\theta}_n, \hat{g}(X_i))(K_h \star K_{h'})(x_0 - X_i) = (K_{h'} \star \hat{f}_h)(x_0).$$

Heuristically, since $\hat{f}_h$ is an estimator of f then $\hat{f}_{h,h'} = K_{h'} \star \hat{f}_h$ can be considered as an estimator of $K_{h'} \star f$. The proof of Theorem 1 below in Section 6.4 then justifies that $A(x_0, h)$ is a good approximation for the bias term of the pointwise risk. Finally, our estimate at the point x_0 is

$$\hat{f}(x_0) := \hat{f}_{\hat{h}(x_0)}(x_0), \tag{8}$$

where the bandwidth $\hat{h}(x_0)$ minimizes the empirical bias-variance decomposition :

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

The constants that appear in the estimated variance $V(x_0, h)$ are known, except κ, which is a numerical constant calibrated by simulation (see practical tuning in Section 5), and except $\|g\|_{\infty,V_n(x_0)}$, which is replaced by an empirical counterpart in practice (see also Section 5). It is also possible to justify the substitution from a theoretical point of view, but it adds cumbersome technicalities. Moreover, the replacement does not change the result of Theorem 1 below. We thus refer to Section 3.3 p.1178 in [8] for example, for the details of a similar substitution. The risk of this estimator is controlled in the following result.

Theorem 1. *Assume that Assumptions* **(A1)** *to* **(A5)** *are fulfilled, and that the sample size n is larger than a constant that only depends on the kernel K. For any $\delta \in (0, 1)$, the estimator $\hat{f}(x_0)$ defined in (8) satisfies*

$$\begin{aligned}\mathbb{E}\left[\left(\hat{f}(x_0) - f(x_0)\right)^2\right] \leqslant\ & C_5^* \min_{h\in\mathcal{H}_n} \left\{ \|K_h \star f - f\|^2_{\infty,V_n(x_0)} + \frac{\log(n)}{\delta^2\gamma^2 nh} \right\} \\ & + \frac{C_6^*}{\delta^6} \sup_{\theta\in[\delta,1-\delta]} \mathbb{E}\left[\left|\tilde{\theta}_n - \theta\right|^2\right] + \frac{C_7^*}{\delta^2\gamma^2}\mathbb{E}\left[\|\hat{g} - g\|^2_{\infty,V_n(x_0)}\right] + \frac{C_8^*}{\delta^2\gamma^2 n^2},\end{aligned} \tag{9}$$

where C_ℓ^*, $\ell = 5, \ldots, 8$ are positive constants such that : C_5^* depends on $\|g\|_{\infty,V_n(x_0)}$, $\|K\|_1$ and $\|K\|_2$, C_6^* depends on $\|K\|_1$, C_7^* depends on $\|g\|_{\infty,V_n(x_0)}$ and $\|K\|_1$, and C_8^* depends on $\|f\|_{\infty,V_n(x_0)}$, g, $\|K\|_2$ and $\|K\|_\infty$.

Theorem 1 is an oracle-type inequality. It holds whatever the sample size, larger than a fixed constant. It shows that the optimal bias variance trade-off is automatically achieved: the selection rule

permits to select in a data-driven way the best estimator in the collection of estimators $(\hat{f}_h)_{h\in\mathcal{H}_n}$, up to a multiplicative constant C_6^*. The last three remainder terms in the *r.h.s.* of (9) are the same as the ones in Proposition 1, and are unavoidable, as aforementioned. We have an additional logarithmic term in the second term of the *r.h.s.*, compared to the analogous term in (6). It is classical in adaptive pointwise estimation (see for example [12] or [4]). In our framework, it does not deteriorate the adaptive convergence rate. The risk of the estimator $\hat{f}(x_0)$ with data-driven bandwidth decreases at the optimal minimax rate of convergence (up to a logarithmic term) if the bandwidth is well-chosen : the upper bound of Corollary 1 matches with the lower-bound for the minimax risk established by Ibragimov and Hasminskii [19].

Corollary 1. *Assume that* **(A6)** *and* **(A7)** *are satisfied, for $\beta > 0$ and $\mathcal{L} > 0$, and for an index $\ell > 0$ such that $\ell \geqslant \lfloor\beta\rfloor$. Suppose also that the assumptions of Theorem 1 are satisfied, and that the preliminary estimates $\tilde{\theta}_n$ and $\hat{g}$ are such that*

$$\mathbb{E}\left[|\tilde{\theta}_n - \theta|^2\right] \leqslant C\left(\frac{\log n}{n}\right)^{\frac{2\beta}{2\beta+1}}, \quad \mathbb{E}\Big[\,\|\hat{g}-g\|^2_{\infty,V_n(x_0)}\,\Big] \leqslant C\left(\frac{\log n}{n}\right)^{\frac{2\beta}{2\beta+1}}. \tag{10}$$

Then,

$$\mathbb{E}\left[\left(\hat{f}(x_0) - f(x_0)\right)^2\right] \leqslant C_9^*\left(\frac{\log n}{n}\right)^{\frac{2\beta}{2\beta+1}}, \tag{11}$$

where C_9^* is a constant depending on $\|g\|_{\infty,V_n(x_0)}$, $\|K\|_1$, $\|K\|_2$, $\mathcal{L}$ and $\|f\|_{\infty,V_n(x_0)}$.

The estimator $\hat{f}$ now achieves the convergence rate $(\log n/n)^{2\beta/(2\beta+1)}$ over the class $\Sigma(\beta,\mathcal{L})$ as soon as $\beta \leqslant \ell$. It automatically adapts to the unknown smoothness of the function to estimate : the bandwidth $\hat{h}(x_0)$ is computed in a fully data-driven way, without using the knowledge of the regularity index β, contrary to the estimator $\hat{f}_n^{rwk}$ of Nguyen and Matias [23] (corollary 3.4). Section 4 below permits to also build $\hat{g}$ and $\tilde{\theta}_n$ without any knowledge of β, to obtain an automatic adaptive estimation procedure.

Remark 1. *In the present work, we focus on Model (1). However, the estimation procedure we develop can easily be extended to the model*

$$g(x) = \theta\phi(x) + (1-\theta)f(x), \quad x \in \mathbb{R}, \tag{12}$$

where the function ϕ is a known density, but not necessarily equal to the uniform one. In this case, a family of kernel estimates can be defined like in (3) replacing the weights $w(\tilde{\theta}_n, \hat{g}(\cdot))$ by

$$w(\tilde{\theta}_n, \hat{g}(\cdot), \phi(x_0)) = \frac{1}{1-\tilde{\theta}_n}\left(1 - \frac{\tilde{\theta}_n\phi(x_0)}{\hat{g}(\cdot)}\right).$$

If the density function ϕ is uniformly bounded on $\mathbb{R}$, it is then possible to obtain analogous results (bias-variance trade-off for the pointwise risk, adaptive bandwidth selection rule leading to oracle-type inequality and optimal convergence rate) as we established for model (1).

4 Estimation of the mixture density g and the mixing proportion θ

This section is devoted to the construction of the preliminary estimators $\hat{g}$ and $\tilde{\theta}_n$, required to build (3). To define them, we assume that we observe an additional sample $(X_i)_{i=n+1,\dots,2n}$ distributed with density function g, but independent of the sample $(X_i)_{i=1,\dots,n}$. We explain how estimators $\hat{g}$ and $\tilde{\theta}_n$ can be defined to satisfy the assumptions described at the beginning of Section 2.2, and also how we compute them in practice. The reader should bear in mind that other constructions are possible, but our main objective is the adaptive estimation of the density f. Thus, further theoretical studies are beyond the scope of this paper.

4.1 Preliminary estimator for the mixture density g

As already noticed, the role plays by g to estimate f in our framework finds an analogue in the work of Bertin *et al.* [3] : the authors propose a conditional density estimation method that involves a preliminary estimator of the marginal density of a couple of real random variables. The assumptions **(A2)** and **(A3)** are borrowed to their paper. From a theoretical point of view, we thus also draw inspiration from them to build $\hat{g}$.

Since we focus on kernel methods to recover f, we also use kernels for the estimation of g. Let $L : \mathbb{R} \to \mathbb{R}$ be a function such that $\int_{\mathbb{R}} L(x)dx = 1$ and $\int_{\mathbb{R}} L^2(x)dx < \infty$. Let $L_b(\cdot) = b^{-1}L(\cdot/b)$, for any $h > 0$. The function L is a kernel, but can be chosen differently from the kernel K used to estimate the density f. The classical kernel density estimate for g is

$$\hat{g}_b(x_0) = \frac{1}{n}\sum_{i=n+1}^{2n} L_b(x_0 - X_i), \tag{13}$$

Theorem 4 p.14 of [2] proves that it is possible to select an adaptive bandwidth b of $\hat{g}_b$ in such a way that Assumptions **(A2)** and **(A3)** are fulfilled, and that the resulting estimate $\hat{g}_{\hat{b}}$ satisfies

$$\mathbb{E}\Big[\, \|\hat{g}_{\hat{b}} - g\|^2_{\infty,V_n(x_0)} \,\Big] \leqslant C\left(\frac{\log n}{n}\right)^{\frac{2\beta}{2\beta+1}},$$

if $g \in \Sigma(\beta, \mathcal{L}')$, where $C, \mathcal{L}' > 0$ are some constants, and if the kernel L has an order $\ell = \lfloor\beta\rfloor$. The idea of the result of Theorem 4 in [2] is to select the bandwidth $\hat{b}$ with a classical Lepski method, and to apply results from Giné and Nickl [16]. Notice that, in our model, Assumption **(A6)** permits to obtain directly the required smoothness assumption, $g \in \Sigma(\beta, \mathcal{L}')$. This guarantees that both the assumptions **(A2)** and **(A3)** on $\hat{g}$ can be satisfied and that the additional term $\mathbb{E}[\|\hat{g} - g\|^2_{\infty,V_n(x_0)}]$ can be bounded as required in the statement of Corollary 1.

For the simulation study below now, we start from the kernel estimators $(\hat{g}_b)_{b>0}$ defined in (13) and rather use a procedure in the spirit of the pointwise GL method to automatically select a bandwidth b. First, this choice permits to be coherent with the selection method chosen for the main estimators $(\hat{f}_h)_{h\in\mathcal{H}_n}$, see Section 3. Then, the construction also provides an accurate estimate of g, see for example [10]. Let $\mathcal{B}$ be a finite family of bandwidths. For any $b, b' \in \mathcal{B}$, we introduce the auxiliary functions $\hat{g}_{b,b'}(x_0) = n^{-1}\sum_{i=n+1}^{2n}(L_b \star L_{b'})(x_0 - X_i)$. Next, for any $b \in \mathcal{B}$, we set

$$A^g(b, x_0) = \max_{b'\in\mathcal{B}} \Big\{ \big(\hat{g}_{b,b'}(x_0) - \hat{g}_{b'}(x_0)\big)^2 - \Gamma_1(b') \Big\}_+,$$

where $\Gamma_1(b) = \epsilon \|L\|_1^2 \|L\|_2^2 \|g\|_\infty \log(n)/(nb)$, with $\epsilon > 0$ a constant to be tuned. Then, the final estimator of g is given by $\hat{g}(x_0) := \hat{g}_{\hat{b}_g(x_0)}(x_0)$, with $\hat{b}_g(x_0) := \operatorname{argmin}_{b\in\mathcal{B}}\{A^g(b, x_0) + \Gamma_1(b)\}$. The tuning of the constant ϵ is presented in Section 5.

4.2 Estimation of the mixing proportion θ

A huge variety of methods have been investigated for the estimation of the mixing proportion θ of model (1) : see, for instance, [28], [20], [26], [5], [24] and references therein. A common and performant estimator is the one proposed by Storey [28]: θ is estimated by $\hat{\theta}_{\tau,n} = \#\{X_i > \tau; i = n+1, \dots, 2n\}/(n(1-\tau))$ with τ a threshold to be chosen. The optimal value of τ is calculated with a boostrap algorithm. However, it seems difficult to obtain theoretical guarantees on $\hat{\theta}_{\tau,n}$. To our knowledge, most of the other methods in the literature rely on different identifiability constraints for the parameters (θ, f). We refer to Celisse and Robin [5] or Nguyen and Matias [24] for a detailed discussion about possible identifiability conditions of model (1). In the sequel we focus on a particular case of model (1), that permits to obtain the identiability of the parameters (θ, f) (see for example Assumption A in [5], or Section 1.1 in [24]). The density f is assumed to belong to the family

$$\mathcal{F}_\delta = \Big\{ f : [0,1] \to \mathbb{R}_+, f \text{ is a continuously non-increasing density, positive on } [0, 1-\delta) \\ \text{and such that } f_{|[1-\delta,1]} = 0 \Big\}, \tag{14}$$

where $\delta \in (0,1)$. Starting from this set, the main idea to recover θ is that it is the lower bound of the density g in model (1) : $\theta = \inf_{x\in[0,1]} g(x) = g(1)$. Celisse and Robin [5] or Nguyen and Matias [24] then define a histogram-based estimator $\hat{g}$ for g, and estimate θ with the lower bound of $\hat{g}$, or with $\hat{g}(1)$. The procedure we choose is in the spirit of this one, but, to be coherent with the other estimates, we use kernels to recover g instead of histograms.

Nevertheless, we cannot directly use the kernel estimates of g defined in (13): it is well-known that kernel density estimation methods suffers from boundary effects, which lead to an inaccuracy estimate of $g(1)$. To avoid this problem, we apply a simple reflection method inspired by Schuster [27]. From the random sample $X_{n+1}, \dots, X_{2n}$ from density g, we introduce, for $k = i - n$, $i = n+1, \dots, 2n$,

$$Y_k = \begin{cases} X_i & \text{if } \epsilon_i = 1, \\ 2 - X_i & \text{if } \epsilon_i = -1, \end{cases}$$

where $\epsilon_{n+1}, \dots, \epsilon_{2n}$ are n *i.i.d.* random variables drawn from Rademacher distribution with parameter $1/2$, and independent of the X_i's. The random variables $Y_1, \dots, Y_n$ can be regarded as symmetrized version of the X_i's, with support $[0,2]$ (see the first point of Lemma 2 below). Now, suppose that L is a symmetric kernel. For any $b > 0$, define

$$\hat{g}_b^{sym}(x) = \frac{1}{2n} \sum_{k=1}^{n} \left[L_b(x - Y_k) + L_b((2-x) - Y_k) \right], \quad x \in [0,2]. \tag{15}$$

The graph of $\hat{g}_b^{sym}$ is symmetric with respect to the straight-line $x = 1$. Then, instead of evaluating $\hat{g}_b^{sym}$ at the single point $x = 1$, we compute the average of all the values of the estimator $\hat{g}_b$ on the interval $[1-\delta, 1]$, relying on the fact that $\theta = g(x)$, for all $x \in [1-\delta, 1]$ (under the assumption $f \in \mathcal{F}_\delta$), to increase the accuracy of the resulting estimate. Thus, we set

$$\hat{\theta}_{n,b} = \frac{2}{\delta} \int_{1-\delta}^{1} \hat{g}_b^{sym}(x) dx. \tag{16}$$

Finally, for the estimation of f, we use a truncated estimator $\tilde{\theta}_n$ defined as

$$\tilde{\theta}_{n,b} := \max\left(\min(\hat{\theta}_{n,b}, 1 - \delta/2), \delta/2 \right). \tag{17}$$

The definition of $\tilde{\theta}_{n,b}$ permits to ensure that $\tilde{\theta}_{n,b} \in [\delta/2, 1-\delta/2]$: this is Assumption **(A4)**. This permits to avoid possible difficulties in the estimation of f when $\hat{\theta}_{n,b}$ is close to zero, see (3). The following lemma establishes some properties of all these estimates. Its proof can be found in Section 6.2.

Lemma 2.

- *The estimator $\hat{g}_b^{sym}$ defined in (15) has the same distribution as a classical kernel estimator from the symmetrized density of g. More precisely, let $(R_i)_{i\in\{1,\dots,n\}}$ be i.i.d. random variables with density*

$$r : x \longmapsto \begin{cases} g(x)/2 \text{ if } x \in [0,1] \\ g(2-x)/2 \text{ if } x \in [1,2], \end{cases}$$

  and $\hat{r}_b(x) = n^{-1} \sum_{i=1}^{n} L_b(x - R_i)$. Then, the estimators $\hat{g}_b^{sym}$ and $\hat{r}_b$ have the same distribution.

- *We have*

$$|\hat{\theta}_{n,b} - \theta| \leqslant \left\| \hat{g}_b^{sym} - g \right\|_{\infty,[1-\delta,1]}. \tag{18}$$

- *Moreover,*

$$\mathbb{P}\left(\tilde{\theta}_{n,b} \neq \hat{\theta}_{n,b} \right) \leqslant \frac{4}{\delta^2} \mathbb{E}\left[|\hat{\theta}_{n,b} - \theta|^2 \right]. \tag{19}$$

The first property of Lemma 2 permits to deal with $\hat{g}_b^{sym}$ as with a classical kernel density estimate. The second property (18) allows us to control the estimation risk of $\hat{\theta}_{n,b}$, while the third one, (19), justifies that the introduction of $\tilde{\theta}_{n,b}$ is reasonable.

To obtain a fully data-driven estimate $\tilde{\theta}_{n,b}$, it remains to define a bandwidth selection rule for the kernel estimator $\hat{g}_b^{sym}$. In view of (18), we introduce a data-driven procedure under sup-norm loss, inspired from

Lepski [21]. For any $x \in [0,2]$ and any bandwidth b, b' in a collection $\mathcal{B}'$, we set $\hat{g}^{sym}_{b,b'}(x) = (L_b \star \hat{g}^{sym}_{b'})(x)$, and $\Gamma_2(b) = \lambda \|L\|_\infty \log(n)/(nb)$, with λ a tuning parameter. As for the other bandwidth selection device, we now define

$$\Delta(b) = \max_{b' \in \mathcal{B}'} \left\{ \sup_{x \in [1-\delta,1]} \left(\hat{g}^{sym}_{b,b'}(x) - \hat{g}^{sym}_{b'}(x) \right)^2 - \Gamma_2(b') \right\}_+ ,$$

Finally, we choose $\tilde{b} = \operatorname{argmin}_{b \in \mathcal{B}'} \{\Delta(b) + \Gamma_2(b)\}$, which leads to $\hat{g}^{sym} := \hat{g}^{sym}_{\tilde{b}}$ and $\tilde{\theta}_n := \tilde{\theta}_{n,\tilde{b}}$. The results of [21], combined with Lemma 2 ensure that $\tilde{\theta}_n$ satisfies (10), if g is smooth enough. Numerical simulations in Section 5 justify that our estimator has a good performance from the practical point of view, in comparison with those proposed in [24] and [28].

5 Numerical study

5.1 Simulated data

We briefly illustrate the performance of the estimation method over simulated data, according the following framework. We simulate observations with density g defined by model (1) for sample size $n \in \{500, 1000, 2000\}$. Three different cases of (θ, f) are considered:

- $f_1(x) = 4(1-x)^3 \mathbb{1}_{[0,1]}(x)$, $\theta_1 = 0.65$.
- $f_2(x) = \dfrac{s}{1-\delta}\left(1 - \dfrac{x}{1-\delta}\right)^{s-1} \mathbb{1}_{[0,1-\delta]}(x)$ with $(\delta, s) = (0.3, 1.4)$, $\theta_2 = 0.45$.
- $f_3(x) = \lambda e^{-\lambda x}\left(1 - e^{-\lambda b}\right)^{-1} \mathbb{1}_{[0,b]}(x)$ the density of truncated exponential distribution on $[0,b]$ with $(\lambda, b) = (10, 0.9)$, $\theta_3 = 0.35$.

The density f_1 is borrowed from [23] while the shape of f_2 is used both by [5] and [24]. Figure 1 represents those three cases with respect to each design density and associated proportion θ.

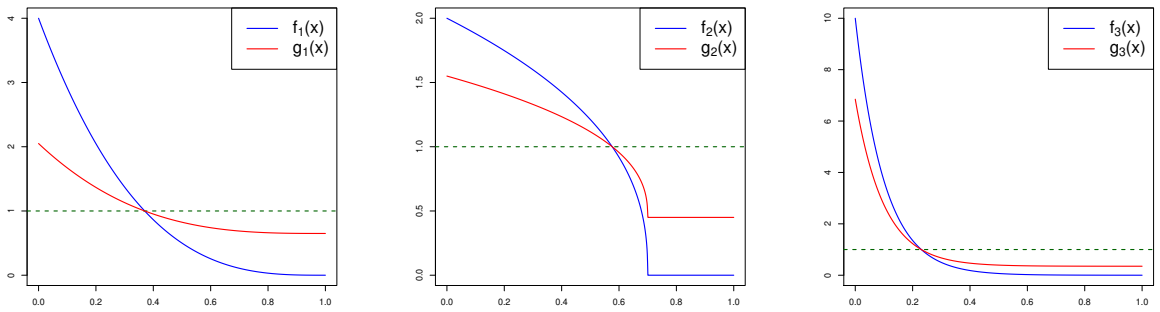
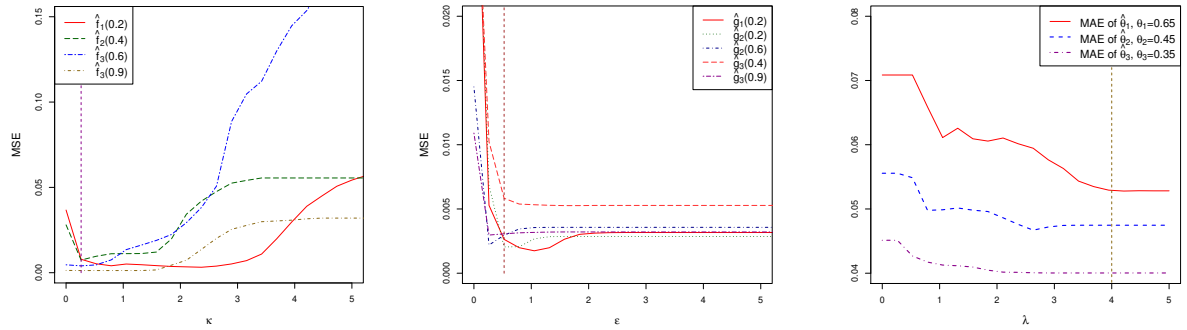


Figure 1: Representation of f_j and the corresponding g_j in model (1) for $(\theta_1 = 0.65, f_1)$ (left), $(\theta_2 = 0.45, f_2)$ (middle) and $(\theta_3 = 0.35, f_3)$ (right).

5.2 Implementation of the method

To compute our estimates, we choose $K(x) = L(x) = (1-|x|)\mathbb{1}_{\{|x| \leqslant 1\}}$ the triangular kernel. In the variance term (7) of the GL method used to select the bandwidth of the kernel estimator of f, we replace $\|g\|_{\infty, V_n(x_0)}$ by the 95^{th} percentile of $\left\{ \max_{t \in V_n(x_0)} \hat{g}_h(t), h \in \mathcal{H}_n \right\}$. Similarly, in the variance term Γ_1 used to select the bandwidth of the kernel estimate of g, we use the 95^{th} percentile of $\left\{ \max_{t \in [0,1]} \hat{g}_h(t), h \in \mathcal{H}_n \right\}$ instead of $\|g\|_\infty$. The collection of bandwidths $\mathcal{H}_n, \mathcal{B}, \mathcal{B}'$ are equal to $\left\{1/k, k = 1, \ldots, \lfloor \sqrt{n} \rfloor\right\}$ where $\lfloor x \rfloor$ denotes a smallest integer which is strictly smaller than the real number x.



(a) (b) (c)

Figure 2: values of the mean-squared error for (a) $\hat{f}(x_0)$ with respect to κ, (b) $\hat{g}(x_0)$ with respect to ϵ. (c) : Values of the mean-absolute error for $\hat{\theta}_n$ with respect to λ. The sample size is $n = 2000$ for all computations. The vertical line corresponds to the chosen value of κ (figure (a)), ϵ (figure (b)) and λ (figure (c)).

We shall settle the values of the constants κ, ϵ and λ involved in the penalty terms $V(x_0, h), \Gamma_1(h)$ and $\Gamma_2(b)$ respectively, to compute the selected bandwidths. Since the calibrations of these tuning parameters are carried out in the same fashion, we only describe the calibration for κ. Denote by $\hat{f}_\kappa$ the estimator of f depending on the constant κ to be calibrated. We approximate the mean-squared error for the estimator $\hat{f}_\kappa$, defined by $\mathrm{MSE}(\hat{f}_\kappa(x_0)) = \mathbb{E}[(\hat{f}_\kappa(x_0) - f(x_0))^2]$, over 100 Monte-Carlo runs, for different possible values $\{\kappa_1, \ldots, \kappa_K\}$ of κ, for the three densities f_1, f_2, f_3 calculated at several test points x_0. We choose a value for κ that leads to small risks in all investigated cases. Figure 2(a) shows that $\kappa = 0.26$ is an acceptable choice even if other values can be also convenient. Similarly, we set $\epsilon = 0.52$ and $\lambda = 4$ (see Figure 2(b) and 2(c)) for the calibrations of ϵ and λ.

5.3 Simulation results

5.3.1 Estimation of the mixing proportion θ

We compare our estimator $\hat{\theta}_n$ with the histogram-based estimator $\hat{\theta}_n^{\text{Ng-M}}$ proposed in [24] and the estimator $\hat{\theta}_n^S$ introduced in [28]. Boxplots in Figure 3 represent the absolute errors of $\hat{\theta}_n$, $\hat{\theta}_n^{\text{Ng-M}}$ and $\hat{\theta}_n^S$, labeled respectively by "Sym-Ker", "Histogram" and "Bootstrap". The estimators $\hat{\theta}_n$ and $\hat{\theta}_n^{\text{Ng-M}}$ have comparable performances, and outperform $\hat{\theta}_n^S$.

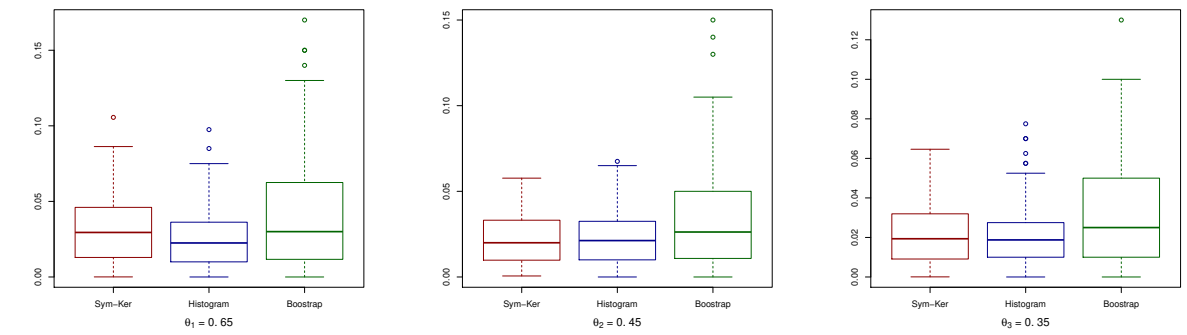


Figure 3: errors for the estimation of θ in the three simulated settings (with sample size $n = 2000$).

5.3.2 Estimation of the target density f

We present in Tables 1, 2 and 3 the mean-squared error (MSE) for the estimation of f according to the three different models and the different sample sizes introduced in Section 5.1. The MSEs' are approximated over 100 Monte-Carlo replications. We shall choose the estimation points (to compute the pointwise risk): we propose $x_0 \in \{0.1, 0.4, 0.6, 0.9\}$. The choices of $x_0 = 0.4$ and $x_0 = 0.6$ are standard. The choices of $x_0 = 0.1$ and $x_0 = 0.9$ allows to test the performance of $\hat{f}$ close to the boundaries of the domain of definition of f and g. We compare our estimator $\hat{f}$ with the randomly weighted estimator proposed in Nguyen and Matias [23]. In the sequel, the label "AWKE" (Adaptive Weighted Kernel Estimator) refers to our estimator $\hat{f}$, whose bandwidth is selected by the Goldenshluger-Lepski method and "Ng-M" refers to the one proposed by [23]. Resulting boxplots are displayed in Figure 4 for $n = 2000$.

Sample size	Estimator	$x_0 = 0.1$	$x_0 = 0.4$	$x_0 = 0.6$	$x_0 = 0.9$
$n = 500$	AWKE	0.1848	0.0121	0.0286	0.0057
	Ng-M	0.2869	0.0450	0.1046	0.0433
$n = 1000$	AWKE	0.0860	0.0088	0.0146	0.0070
	Ng-M	0.1643	0.0469	0.0651	0.0279
$n = 2000$	AWKE	0.0437	0.0119	0.0107	0.0050
	Ng-M	0.0982	0.0246	0.0326	0.0164

Table 1: mean-squared error of the reconstruction of f_1, for our estimator $\hat{f}$ (AWKE), and for the estimator of Nguyen and Matias [23] (Ng-M).

Sample size	Estimator	$x_0 = 0.1$	$x_0 = 0.4$	$x_0 = 0.6$	$x_0 = 0.9$
$n = 500$	AWKE	0.0453	0.0136	0.0297	0.0024
	Ng-M	0.0560	0.0540	0.0306	0.0138
$n = 1000$	AWKE	0.0190	0.0061	0.0237	0.0006
	Ng-M	0.0277	0.0209	0.0123	0.0069
$n = 2000$	AWKE	0.0063	0.0036	0.0075	0.0001
	Ng-M	0.0164	0.0159	0.0113	0.0038

Table 2: mean-squared error of the reconstruction of f_2, for our estimator $\hat{f}$ (AWKE), and for the estimator of Nguyen and Matias [23] (Ng-M).

Sample size	Estimator	$x_0 = 0.1$	$x_0 = 0.4$	$x_0 = 0.6$	$x_0 = 0.9$
$n = 500$	AWKE	0.0806	0.0096	0.0045	0.0016
	Ng-M	0.1308	0.0247	0.0207	0.0096
$n = 1000$	AWKE	0.0321	0.0054	0.0029	0.0010
	Ng-M	0.0566	0.0106	0.0096	0.0060
$n = 2000$	AWKE	0.0239	0.0025	0.0015	0.0005
	Ng-M	0.0342	0.0059	0.0062	0.0021

Table 3: mean-squared error of the reconstruction of f_3, for our estimator $\hat{f}$ (AWKE), and for the estimator of Nguyen and Matias [23] (Ng-M).

Tables 1, 2, 3 and boxplots show that our estimator outperforms the one of [23]. Notice that the errors are relatively large at the point $x_0 = 0.1$, for both estimators, which was expected (boundary effect).

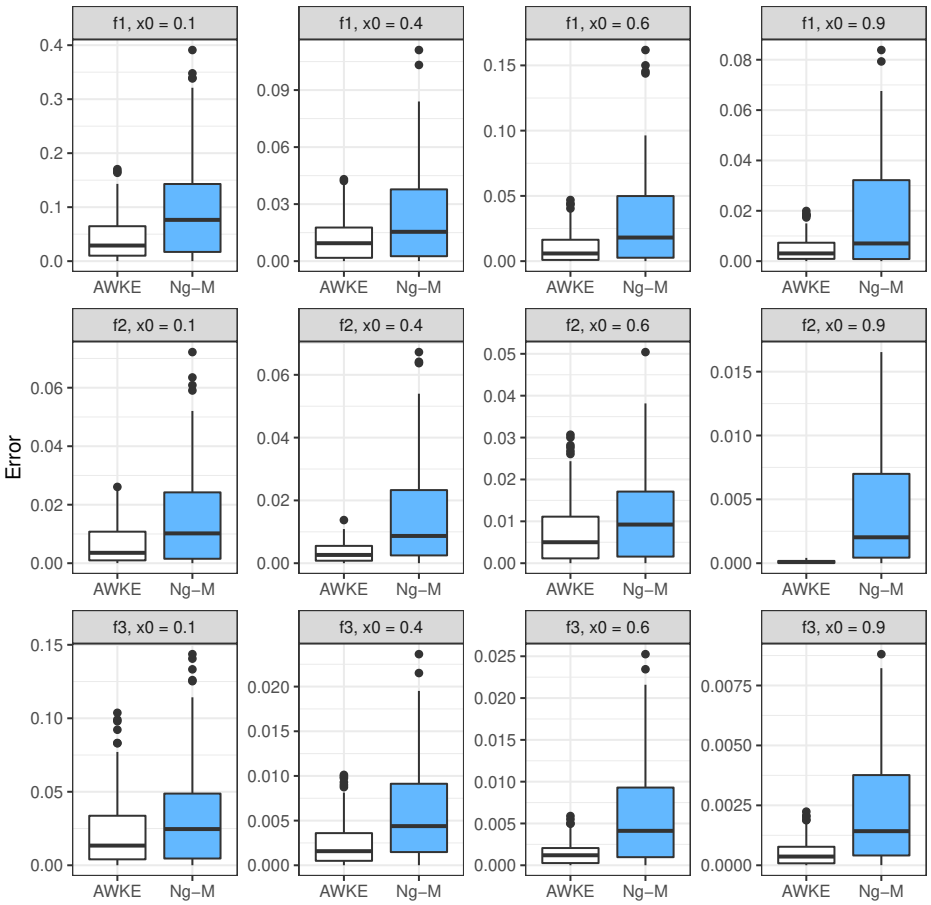


Figure 4: errors for the estimation of f_1, f_2 and f_3 for $x_0 \in \{0.1, 0.4, 0.6, 0.9\}$ and sample size $n = 2000$.

6 Proofs

In the sequel, the notations $\tilde{\mathbb{P}}$, $\tilde{\mathbb{E}}$ and $\tilde{\mathbb{V}}ar$ respectively denote the probability, the expectation and the variance associated with $X_1, \ldots, X_n$, conditionally on the additional random sample $X_{n+1}, \ldots, X_{2n}$.

6.1 Proof of Proposition 1

Let $\rho > 1$, introduce the event

$$\Omega_\rho = \left\{ \rho^{-1}\gamma \leqslant \hat{\gamma} \leqslant \rho\gamma \right\}.$$

such as

$$\hat{f}_h(x_0) - f(x_0) = \big(\hat{f}_h(x_0) - f(x_0)\big)\mathbb{1}_{\Omega_\rho} + \big(\hat{f}_h(x_0) - f(x_0)\big)\mathbb{1}_{\Omega_\rho^c}. \tag{20}$$

We first evaluate the term $\big(\hat{f}_h(x_0) - f(x_0)\big)\mathbb{1}_{\Omega_\rho}$. Suppose now that we are on Ω_ρ, then for any $x_0 \in [0,1]$, we have

$$\big(\hat{f}_h(x_0) - f(x_0)\big)^2 \leqslant 3\Big(\big(\hat{f}_h(x_0) - K_h \star \check{f}(x_0)\big)^2 + \big(K_h \star \check{f}(x_0) - \check{f}(x_0)\big)^2 + \big(\check{f}(x_0) - f(x_0)\big)^2\Big), \tag{21}$$

where we define

$$\check{f}(x) = w(\tilde{\theta}_n, \hat{g}(x))g(x) = \frac{1}{1-\tilde{\theta}_n}\left(1 - \frac{\tilde{\theta}_n}{\hat{g}(x)}\right) g(x).$$

Note that by definition of $\check{f}$, we have $K_h \star \check{f}(x_0) = \tilde{\mathbb{E}}\big[\hat{f}_h(x_0)\big]$. Hence,

$$\big(\hat{f}_h(x_0) - K_h \star \check{f}(x_0)\big)^2 = \big(\hat{f}_h(x_0) - \tilde{\mathbb{E}}\big[\hat{f}_h(x_0)\big]\big)^2.$$

It follows that

$$\tilde{\mathbb{E}}\left[\left(\hat{f}_h(x_0)-\tilde{\mathbb{E}}[\hat{f}_h(x_0)]\right)^2\right]=\tilde{\mathbb{V}}ar\left(\hat{f}_h(x_0)\right)=\tilde{\mathbb{V}}ar\left(\frac{1}{n}\sum_{i=1}^{n}w(\tilde{\theta}_n,\hat{g}(X_i))K_h(x_0-X_i)\right)$$
$$=\frac{1}{n}\tilde{\mathbb{V}}ar\left(w(\tilde{\theta}_n,\hat{g}(X_1))K_h(x_0-X_1)\right)$$
$$\leqslant\frac{1}{n}\tilde{\mathbb{E}}\left[\left(w(\tilde{\theta}_n,\hat{g}(X_1))K_h(x_0-X_1)\right)^2\right].$$

On the other hand, for all $i\in\{1,\dots,n\}$, thanks to **(A4)** and **(A2)**,

$$w(\tilde{\theta}_n,\hat{g}(X_i))K_h(x_0-X_i)=\frac{1}{1-\tilde{\theta}_n}\left(1-\frac{\tilde{\theta}_n}{\hat{g}(X_i)}\right)K_h(x_0-X_i)\leqslant\frac{2}{\delta}\left(1+\frac{\tilde{\theta}_n}{|\hat{g}(X_i)|}\right)K_h(x_0-X_i)$$
$$\leqslant\frac{2}{\delta}\left(1+\frac{1}{\hat{\gamma}}\right)K_h(x_0-X_i)\leqslant\frac{4}{\delta\hat{\gamma}}K_h(x_0-X_i).\tag{22}$$

Indeed, as we use compactly supported kernel to construct the estimator $\hat{f}_h$, condition $\alpha_n\leqslant h^{-1}$ in **(A5)** ensures that $\left(\hat{g}(X_i)\right)^{-1}K_h(x_0-X_i)$ is upper bounded by $\hat{\gamma}^{-1}K_h(x_0-X_i)$ even though we have no observation in the neighbourhood of x_0.

Moreover, since $\hat{\gamma}\geqslant\rho^{-1}\gamma$ on Ω_ρ, we have that $w(\tilde{\theta}_n,\hat{g}(X_i))\leqslant 4\rho\delta^{-1}\gamma^{-1}$. Thus we obtain

$$\tilde{\mathbb{E}}\left[\left(\hat{f}_h(x_0)-\tilde{\mathbb{E}}[\hat{f}_h(x_0)]\right)^2\right]\leqslant\frac{16\rho^2}{\delta^2\gamma^2 n}\tilde{\mathbb{E}}\left[K_h^2(x_0-X_1)\right]\leqslant\frac{16\rho^2\|K\|_2^2\,\|g\|_{\infty,V_n(x_0)}}{\delta^2\gamma^2 nh}.\tag{23}$$

For the last two terms of (21), we apply the following proposition:

Proposition 2. *Assume* **(A1)** *and* **(A3)**. *On the set Ω_ρ, we have the following results for any $x_0\in[0,1]$*

$$\left(\check{f}(x_0)-f(x_0)\right)^2\leqslant C_1\delta^{-2}\gamma^{-2}\|\hat{g}-g\|_{\infty,V_n(x_0)}^2+C_2\delta^{-6}\left|\tilde{\theta}_n-\theta\right|^2,\tag{24}$$
$$\left(K_h\star\check{f}(x_0)-\check{f}(x_0)\right)^2\leqslant 6\|K_h\star f-f\|_{\infty,V_n(x_0)}^2+C_3\delta^{-2}\gamma^{-2}\|\hat{g}-g\|_{\infty,V_n(x_0)}^2+C_4\delta^{-6}\left|\tilde{\theta}_n-\theta\right|^2,\tag{25}$$

where C_1 and C_2 respectively depend on ρ and $\|g\|_{\infty,V_n(x_0)}$, C_3 depends on ρ and $\|K\|_1$ and C_4 depends on $\|g\|_{\infty,V_n(x_0)}$ and $\|K\|_1$.

Combining (23), (24) and (25), we obtain

$$\mathbb{E}\left[\left(\hat{f}_h(x_0)-f(x_0)\right)^2\mathbb{1}_{\Omega_\rho}\right]\leqslant 18\|K_h\star f-f\|_{\infty,V_n(x_0)}^2+3(C_1+C_3)\delta^{-2}\gamma^{-2}\mathbb{E}\left[\|\hat{g}-g\|_{\infty,V_n(x_0)}^2\right]$$
$$+3(C_2+C_4)\delta^{-6}\mathbb{E}\left[\left|\tilde{\theta}_n-\theta\right|^2\right]+\frac{48\rho^2\|K\|_2^2\,\|g\|_{\infty,V_n(x_0)}}{\delta^2\gamma^2 nh}.$$

It remains to study the risk bound on Ω_ρ^c. To do so, we successively apply the following lemmas whose proofs are postponed to the end of Theorem 1's proof.

Lemma 3. *Suppose that Assumption* **(A3)** *is satisfied. Then we have for $\rho>1$*

$$\mathbb{P}\left(\Omega_\rho^c\right)\leqslant C_{g,\rho}\exp\left\{-(\log n)^{3/2}\right\},$$

with $C_{g,\rho}$ a positive constant depending on g and ρ.

Lemma 4. *Assume* **(A2)** *and* **(A5)**. *For any $h\in\mathcal{H}_n$, we have*

$$\mathbb{E}\left[\left(\hat{f}_h(x_0)-f(x_0)\right)^2\mathbb{1}_{\Omega_\rho^c}\right]\leqslant\frac{C_4^*}{\delta^2 n^2},$$

with C_4^* a positive constant depending on $\|f\|_{\infty,V_n(x_0)}$, $\|K\|_\infty$, g and ρ.

This concludes the proof of Proposition 1.

6.2 Proof of Lemma 2

First, we prove that $\hat{g}_b^{sym}$ has the same distribution as $\hat{r}_b$. First, Y_i has the same distribution as $2 - Y_i$. Thus, $\hat{g}_b^{sym}$ has the same distribution as $x \mapsto n^{-1} \sum_{i=1}^n L_b(x - Y_i)$. It is thus sufficient to show that Y_i has the same distribution as R_i. To this aim, let φ be a measurable bounded function defined on $\mathbb{R}$. We compute

$$\begin{aligned}
\mathbb{E}[\varphi(Y_i)] &= \mathbb{E}[\mathbb{E}[\varphi(X_i)|\varepsilon_i]\mathbb{1}_{\{\epsilon_1=1\}}] + \mathbb{E}[\mathbb{E}[\varphi(2-X_i)|\varepsilon_i]\mathbb{1}_{\{\epsilon_1=-1\}}], \\
&= \frac{1}{2}\left(\mathbb{E}[\varphi(X_i)] + \mathbb{E}[\varphi(2-X_i)]\right), \\
&= \frac{1}{2}\left(\int_0^1 \varphi(x)g(x)dx + \int_0^1 \varphi(2-x)g(x)dx\right), \\
&= \frac{1}{2}\left(\int_0^1 \varphi(x)g(x)dx + \int_1^2 \varphi(x)g(2-x)dx\right), \\
&= \int_0^2 \varphi(x)r(x)dx = \mathbb{E}[\varphi(R_i)].
\end{aligned}$$

This allows us to obtain the first assertion of the lemma.

We prove now (18). Under the identifiability condition, we have $\theta = g(x)$ for all $x \in [1-\delta, 1]$. Hence we have

$$\begin{aligned}
|\hat{\theta}_{n,b} - \theta| &= \left|\frac{1}{\delta}\int_{1-\delta}^1 \hat{g}_b^{sym}(x)dx - \frac{1}{\delta}\int_{1-\delta}^1 g(x)dx\right| = \frac{1}{\delta}\left|\int_{1-\delta}^1 (\hat{g}_b^{sym}(x) - g(x))dx\right| \\
&\leqslant \frac{1}{\delta}\int_{1-\delta}^1 \left|\hat{g}_b^{sym}(x) - g(x)\right| dx \\
&\leqslant \frac{1}{\delta}\int_{1-\delta}^1 \left\|\hat{g}_b^{sym} - g\right\|_{\infty,[1-\delta,1]} dx \\
&= \left\|\hat{g}_b^{sym} - g\right\|_{\infty,[1-\delta,1]}.
\end{aligned}$$

Moreover, thanks to the Markov Inequality

$$\begin{aligned}
\mathbb{P}\left(\tilde{\theta}_{n,b} \neq \hat{\theta}_{n,b}\right) &= \mathbb{P}\left(\hat{\theta}_{n,b} \notin \left[\frac{\delta}{2}, 1-\frac{\delta}{2}\right]\right) \\
&\leqslant \mathbb{P}\left(|\hat{\theta}_{n,b} - \theta| > \frac{\delta}{2}\right) \leqslant \frac{4}{\delta^2}\mathbb{E}\left[|\hat{\theta}_{n,b} - \theta|^2\right],
\end{aligned}$$

which is (19). This concludes the proof of Lemma 2.

6.3 Proof of Proposition 2

Let us introduce the function

$$\tilde{f}(x) := w(\tilde{\theta}_n, g(x))g(x) = \frac{1}{1-\tilde{\theta}_n}\left(1 - \frac{\tilde{\theta}_n}{g(x)}\right)g(x). \tag{26}$$

Then we have for $x_0 \in [0,1]$

$$\left(\check{f}(x_0) - f(x_0)\right)^2 \leqslant 2\left(\left(\check{f}(x_0) - \tilde{f}(x_0)\right)^2 + \left(\tilde{f}(x_0) - f(x_0)\right)^2\right).$$

For the first term, on $\Omega_\rho = \{\rho^{-1}\gamma \leqslant \hat{\gamma} \leqslant \rho\gamma\}$ we have, by using **(A4)**,

$$\begin{aligned}
\big(\check{f}(x_0) - \tilde{f}(x_0)\big)^2 &= \Big(w(\tilde{\theta}_n, \hat{g}(x_0))g(x_0) - w(\tilde{\theta}_n, g(x_0))g(x_0)\Big)^2 \\
&= \left(\frac{1}{1-\tilde{\theta}_n}\left(1 - \frac{\tilde{\theta}_n}{\hat{g}(x_0)}\right) - \frac{1}{1-\tilde{\theta}_n}\left(1 - \frac{\tilde{\theta}_n}{g(x_0)}\right)\right)^2 |g(x_0)|^2 \\
&= \frac{\tilde{\theta}_n^2}{(1-\tilde{\theta}_n)^2}\left(\frac{1}{\hat{g}(x_0)} - \frac{1}{g(x_0)}\right)^2 |g(x_0)|^2 \\
&\leqslant \frac{4}{\delta^2}\left(\frac{\hat{g}(x_0) - g(x_0)}{\hat{g}(x_0)g(x_0)}\right)^2 |g(x_0)|^2 \\
&\leqslant 4\rho^2\delta^{-2}\gamma^{-2}\,\|\hat{g} - g\|^2_{\infty, V_n(x_0)}\,. \qquad (27)
\end{aligned}$$

Moreover, thanks to **(A1)**,

$$\begin{aligned}
\big(\tilde{f}(x_0) - f(x_0)\big)^2 &= \Big(w(\tilde{\theta}_n, g(x_0))g(x_0) - w(\theta, g(x_0))g(x_0)\Big)^2 \\
&= \left(\frac{1}{1-\tilde{\theta}_n}\left(1 - \frac{\tilde{\theta}_n}{g(x_0)}\right)g(x_0) - \frac{1}{1-\theta}\left(1 - \frac{\theta}{g(x_0)}\right)g(x_0)\right)^2 \\
&= \left(\frac{1}{1-\tilde{\theta}_n} - \frac{1}{1-\theta} + \left(\frac{\theta}{1-\theta} - \frac{\tilde{\theta}_n}{1-\tilde{\theta}_n}\right)\frac{1}{g(x_0)}\right)^2 |g(x_0)|^2 \\
&= \frac{|g(x_0)|^2}{(1-\theta)^2(1-\tilde{\theta}_n)^2}\left(\tilde{\theta}_n - \theta + \frac{\theta - \tilde{\theta}_n}{g(x_0)}\right)^2 \\
&\leqslant \frac{4\,\|g\|^2_{\infty, V_n(x_0)}}{\delta^4}\left(\tilde{\theta}_n - \theta + \frac{\theta - \tilde{\theta}_n}{g(x_0)}\right)^2 \\
&\leqslant 16\,\|g\|^2_{\infty, V_n(x_0)}\,\delta^{-6}\big|\tilde{\theta}_n - \theta\big|^2. \qquad (28)
\end{aligned}$$

Thus we obtain by gathering (27) and (28),

$$\big(\check{f}(x_0) - f(x_0)\big)^2 \leqslant 8\rho^2\delta^{-2}\gamma^{-2}\,\|\hat{g} - g\|^2_{\infty, V_n(x_0)} + 32\,\|g\|^2_{\infty, V_n(x_0)}\,\delta^{-6}\big|\tilde{\theta}_n - \theta\big|^2.$$

Next, the term $\big(K_h \star \check{f}(x_0) - \check{f}(x_0)\big)^2$ can be treated by studying the following decomposition

$$\begin{aligned}
\big(K_h \star \check{f}(x_0) - \check{f}(x_0)\big)^2 \leqslant 3\Big(&\big(K_h \star \check{f}(x_0) - K_h \star \tilde{f}(x_0)\big)^2 + \big(K_h \star \tilde{f}(x_0) - K_h \star f(x_0)\big)^2 \\
&+ \big(K_h \star f(x_0) - \check{f}(x_0)\big)^2\Big) \\
=: 3\big(&A_1 + A_2 + A_3\big).
\end{aligned}$$

For term A_1, we have by using (27)

$$\begin{aligned}
A_1 = \big(K_h \star (\check{f} - \tilde{f})(x_0)\big)^2 &= \left(\int K_h(x_0 - u)(\check{f}(u) - \tilde{f}(u))du\right)^2 \\
&\leqslant \left(\int |K_h(x_0 - u)||\check{f}(u) - \tilde{f}(u)|du\right)^2 \\
&\leqslant 4\rho^2\delta^{-2}\gamma^{-2}\,\|\hat{g} - g\|^2_{\infty, V_n(x_0)}\left(\int |K_h(x_0 - u)|du\right)^2 \\
&\leqslant 4\rho^2\delta^{-2}\gamma^{-2}\,\|K\|_1^2\,\|\hat{g} - g\|^2_{\infty, V_n(x_0)}\,.
\end{aligned}$$

By using (28) and following the same lines as for A_1, we obtain

$$A_2 = \big(K_h \star (\tilde{f} - f)(x_0)\big)^2 \leqslant 16\, \|g\|^2_{\infty, V_n(x_0)}\, \delta^{-6}\, \|K\|_1^2\, \big|\tilde{\theta}_n - \theta\big|^2.$$

For A_3, using the upper bound obtained as above for $(\check{f}(x_0) - f(x_0))^2$, we have

$$\begin{aligned} A_3 &\leqslant 2\big(K_h \star f(x_0) - f(x_0)\big)^2 + 2\big(f(x_0) - \check{f}(x_0)\big)^2 \\ &\leqslant 2\, \|K_h \star f - f\|^2_{\infty, V_n(x_0)} + 16\rho^2\delta^{-2}\gamma^{-2}\, \|\hat{g} - g\|^2_{\infty, V_n(x_0)} + 64\, \|g\|^2_{\infty, V_n(x_0)}\, \delta^{-6} \big|\tilde{\theta}_n - \theta\big|^2. \end{aligned}$$

Finally, combining all the terms A_1, A_2 and A_3, we obtain (25). This ends the proof of Proposition 2.

6.4 Proof of Theorem 1

Suppose that we are on Ω_ρ. Let $\hat{f}$ be the adaptive estimator defined in (8), we have for any $x_0 \in [0,1]$,

$$\big(\hat{f}(x_0) - f(x_0)\big)^2 \leqslant 2\left(\big(\hat{f}(x_0) - \check{f}(x_0)\big)^2 + \big(\check{f}(x_0) - f(x_0)\big)^2\right)$$

The second term is controlled by (24) of Proposition 2. Hence it remains to handle with the first term. For any $h \in \mathcal{H}_n$, we have

$$\begin{aligned} \big(\hat{f}(x_0) - \check{f}(x_0)\big)^2 &\leqslant 3\left(\big(\hat{f}_{\hat{h}(x_0)}(x_0) - \hat{f}_{\hat{h},h}(x_0)\big)^2 + \big(\hat{f}_{\hat{h}(x_0),h}(x_0) - \hat{f}_h(x_0)\big)^2 + \big(\hat{f}_h(x_0) - \check{f}(x_0)\big)^2\right) \\ &= 3\Big(\big(\hat{f}_{\hat{h}(x_0)}(x_0) - \hat{f}_{\hat{h},h}(x_0)\big)^2 - V(x_0,\hat{h}) + \big(\hat{f}_{\hat{h}(x_0),h}(x_0) - \hat{f}_h(x_0)\big)^2 - V(x_0,h) \\ &\qquad + V(x_0,\hat{h}) + V(x_0,h) + \big(\hat{f}_h(x_0) - \check{f}(x_0)\big)^2\Big) \\ &\leqslant 3\left(A(x_0,\hat{h}) + A(x_0,h) + V(x_0,\hat{h}) + V(x_0,h) + \big(\hat{f}_h(x_0) - \check{f}(x_0)\big)^2\right) \\ &\leqslant 6A(x_0,h) + 6V(x_0,h) + 3\big(\hat{f}_h(x_0) - K_h \star \check{f}(x_0)\big)^2 + 3\big(K_h \star \check{f}(x_0) - \check{f}(x_0)\big)^2. \end{aligned} \tag{29}$$

Next, we have

$$\begin{aligned} A(x_0,h) &= \max_{h' \in \mathcal{H}_n} \left\{\big(\hat{f}_{h,h'}(x_0) - \hat{f}_{h'}(x_0)\big)^2 - V(x_0,h')\right\}_+ \\ &\leqslant 3 \max_{h' \in \mathcal{H}_n} \Big\{\big(\hat{f}_{h,h'}(x_0) - K_{h'} \star (K_h \star \check{f})(x_0)\big)^2 + \big(\hat{f}_{h'}(x_0) - K_{h'} \star \check{f}(x_0)\big)^2 \\ &\qquad + \big(K_{h'} \star (K_h \star \check{f})(x_0) - K_{h'} \star \check{f}(x_0)\big)^2 - \frac{V(x_0,h')}{3}\Big\}_+ \\ &\leqslant 3\left(B(h) + D_1 + D_2\right), \end{aligned}$$

where

$$\begin{aligned} B(h) &= \max_{h' \in \mathcal{H}_n} \Big(K_{h'} \star (K_h \star \check{f})(x_0) - K_{h'} \star \check{f}(x_0)\Big)^2 \\ D_1 &= \max_{h' \in \mathcal{H}_n} \left\{\big(\hat{f}_{h'}(x_0) - K_{h'} \star \check{f}(x_0)\big)^2 - \frac{V(x_0,h')}{6}\right\}_+ \\ D_2 &= \max_{h' \in \mathcal{H}_n} \left\{\big(\hat{f}_{h,h'}(x_0) - K_{h'} \star (K_h \star \check{f})(x_0)\big)^2 - \frac{V(x_0,h')}{6}\right\}_+. \end{aligned}$$

Since

$$\begin{aligned} B(h) &= \max_{h' \in \mathcal{H}_n} \Big(K_{h'} \star (K_h \star \check{f})(x_0) - K_{h'} \star \check{f}(x_0)\Big)^2 = \max_{h' \in \mathcal{H}_n} \Big(K_{h'} \star (K_h \star \check{f} - \check{f})(x_0)\Big)^2 \\ &\leqslant \|K\|_1^2 \sup_{t \in V_n(x_0)} \big|K_h \star \check{f}(t) - \check{f}(t)\big|^2, \end{aligned}$$

then we can rewrite (29) as

$$\Big(\hat{f}(x_0)-\check{f}(x_0)\Big)^2 \leqslant 18D_1+18D_2+6V(x_0,h)+3\big(\hat{f}_h(x_0)-K_h\star\check{f}(x_0)\big)^2 + (18\,\|K\|_1^2+3)\sup_{t\in V_n(x_0)}\big|K_h\star\check{f}(t)-\check{f}(t)\big|^2. \quad (30)$$

The last two terms of (30) are controlled by (23) and (25) of Proposition 2. Hence it remains to deal with terms D_1 and D_2.

For D_1, we recall that $K_h\star\check{f}(x_0)=\tilde{\mathbb{E}}\big[\hat{f}_h(x_0)\big]$ and

$$\begin{aligned}\tilde{\mathbb{E}}[D_1] &= \tilde{\mathbb{E}}\left[\max_{h\in\mathcal{H}_n}\left\{\Big(\hat{f}_h(x_0)-K_h\star\check{f}(x_0)\Big)^2-\frac{V(x_0,h)}{6}\right\}_+\right]\\ &\leqslant \sum_{h\in\mathcal{H}_n}\tilde{\mathbb{E}}\left[\left\{\Big(\hat{f}_h(x_0)-\tilde{\mathbb{E}}\big[\hat{f}_h(x_0)\big]\Big)^2-\frac{V(x_0,h)}{6}\right\}_+\right]\\ &\leqslant \sum_{h\in\mathcal{H}_n}\int_0^{+\infty}\tilde{\mathbb{P}}\left(\left\{\Big(\hat{f}_h(x_0)-\tilde{\mathbb{E}}\big[\hat{f}_h(x_0)\big]\Big)^2-\frac{V(x_0,h)}{6}\right\}_+>u\right)du\\ &\leqslant \sum_{h\in\mathcal{H}_n}\int_0^{+\infty}\tilde{\mathbb{P}}\left(\big|\hat{f}_h(x_0)-\tilde{\mathbb{E}}\big[\hat{f}_h(x_0)\big]\big|>\sqrt{\frac{V(x_0,h)}{6}+u}\right)du. \quad (31)\end{aligned}$$

Now let us introduce the sequence of *i.i.d.* random variables $Z_1,\dots,Z_n$ where we set

$$Z_i=w(\tilde{\theta}_n,\hat{g}(X_i))K_h(x_0-X_i).$$

Then we have

$$\hat{f}_h(x_0)-\tilde{\mathbb{E}}\big[\hat{f}_h(x_0)\big]=\frac{1}{n}\sum_{i=1}^n\big(Z_i-\tilde{\mathbb{E}}[Z_i]\big).$$

Moreover, we have by (22) and (23)

$$|Z_i|=|w(\tilde{\theta}_n,\hat{g}(X_i))K_h(x_0-X_i)|\leqslant\frac{4\,\|K\|_\infty}{h\delta\hat{\gamma}}=:b,$$

and

$$\tilde{\mathbb{E}}\big[Z_1^2\big]=\tilde{\mathbb{E}}\Big[w(\tilde{\theta}_n,\hat{g}(X_i))^2K_h^2(x_0-X_i)\Big]\leqslant\frac{16\,\|K\|_2^2\,\|g\|_{\infty,V_n(x_0)}}{h\delta^2\hat{\gamma}^2}=:v.$$

Applying the Bernstein inequality (cf. Lemma 2 of Comte and Lacour [9]), we have for any $u>0$,

$$\begin{aligned}\tilde{\mathbb{P}}\left(\big|\hat{f}_h(x_0)-\tilde{\mathbb{E}}\big[\hat{f}_h(x_0)\big]\big|>\sqrt{\frac{V(x_0,h)}{6}+u}\right) &= \tilde{\mathbb{P}}\left(\Big|\frac{1}{n}\sum_{i=1}^n\big(Z_i-\tilde{\mathbb{E}}[Z_i]\big)\Big|>\sqrt{\frac{V(x_0,h)}{6}+u}\right)\\ &\leqslant 2\max\left\{\exp\left(-\frac{n}{4v}\left(\frac{V(x_0,h)}{6}+u\right)\right),\exp\left(-\frac{n}{4b}\sqrt{\frac{V(x_0,h)}{6}+u}\right)\right\}\\ &\leqslant 2\max\left\{\exp\left(-\frac{n}{24v}V(x_0,h)\right)\exp\left(-\frac{nu}{4v}\right),\exp\left(-\frac{n}{8b}\sqrt{\frac{V(x_0,h)}{6}}\right)\exp\left(-\frac{n\sqrt{u}}{8b}\right)\right\}\end{aligned}$$

On the other hand, by the definition of $V(x_0,h)$ we have

$$\frac{n}{24v}V(x_0,h)=\frac{nh\hat{\gamma}^2\delta^2}{384\rho\,\|K\|_2^2\,\|g\|_{\infty,V_n(x_0)}}\times\frac{\kappa\,\|K\|_1^2\,\|K\|_2^2\,\|g\|_{\infty,V_n(x_0)}}{\hat{\gamma}^2nh}\log(n)=\frac{\kappa\delta^2\,\|K\|_1^2}{384\rho}\log(n)\geqslant\frac{\kappa\delta^2}{384\rho}\log(n).$$

If we choose κ such that $\dfrac{\kappa\delta^2}{384\rho}\geqslant 2$, we get

$$\frac{n}{24v}V(x_0,h)\geqslant 2\log(n).$$

Moreover, using the assumption that $\hat{\gamma}nh \geqslant \log^3(n)$ we have

$$\begin{aligned}\frac{n}{8b}\sqrt{\frac{V(x_0,h)}{6}} &= \frac{nh\hat{\gamma}\delta}{32\sqrt{6}\,\|K\|_\infty} \times \frac{\|K\|_1\,\|K\|_2\,\sqrt{\kappa\,\|g\|_{\infty,V_n(x_0)}\log(n)}}{\hat{\gamma}\sqrt{nh}} \\ &= \frac{\delta\,\|K\|_1\,\|K\|_2\,\|g\|^{1/2}_{\infty,V_n(x_0)}}{32\sqrt{6}\,\|K\|_\infty}\sqrt{\kappa nh\log(n)} \\ &\geqslant \frac{\delta\,\|K\|_1\,\|K\|_2}{32\sqrt{6}\rho^{1/2}\gamma^{1/2}\,\|K\|_\infty}\sqrt{\kappa}\log^2(n) \geqslant 2\log(n),\end{aligned}$$

if

$$\frac{\delta\,\|K\|_1\,\|K\|_2}{32\sqrt{6}\rho^{1/2}\gamma^{1/2}\,\|K\|_\infty}\sqrt{\kappa}\log(n) \geqslant 2$$

which automatically holds for well-chosen value of κ, and n large enough. Then we have by using the conditions $\rho^{-1}\gamma \leqslant \hat{\gamma}$ and $h \geqslant 1/n$,

$$\begin{aligned}\tilde{\mathbb{E}}[D_1] &\leqslant \sum_{h\in\mathcal{H}_n}\int_0^{+\infty} 2n^{-2}\max\left\{\exp\left(-\frac{nu}{4v}\right),\exp\left(-\frac{n\sqrt{u}}{8b}\right)\right\}du \\ &\leqslant 2n^{-2}\sum_{h\in\mathcal{H}_n}\int_0^{+\infty}\max\left\{\exp\left(-nh\frac{\delta^2\hat{\gamma}^2}{64\,\|K\|_2^2\,\|g\|_{\infty,V_n(x_0)}}u\right),\exp\left(-nh\frac{\delta\hat{\gamma}}{32\,\|K\|_\infty}\sqrt{u}\right)\right\}du \\ &\leqslant 2n^{-2}\sum_{h\in\mathcal{H}_n}\int_0^{+\infty}\max\left\{\exp\left(-nh\frac{\delta^2\gamma^2}{64\rho^2\,\|K\|_2^2\,\|g\|_{\infty,V_n(x_0)}}u\right),\exp\left(-nh\frac{\delta\gamma}{32\rho\,\|K\|_\infty}\sqrt{u}\right)\right\}du \\ &\leqslant 2n^{-2}\sum_{h\in\mathcal{H}_n}\int_0^{+\infty}\max\left\{e^{-\pi_1 u},e^{-\pi_2\sqrt{u}}\right\}du \leqslant 2n^{-2}\sum_{h\in\mathcal{H}_n}\max\left\{\frac{1}{\pi_1},\frac{2}{\pi_2^2}\right\}.\end{aligned}$$

with $\pi_1 := \dfrac{\delta^2\gamma^2}{64\rho^2\,\|K\|_2^2\,\|g\|_{\infty,V_n(x_0)}}$ and $\pi_2 := \dfrac{\delta\gamma}{32\rho\,\|K\|_\infty}$.

Since $\mathrm{card}(\mathcal{H}_n) \leqslant n$, we finally obtain

$$\tilde{\mathbb{E}}[D_1] \leqslant C_5\delta^{-2}\gamma^{-2}n^{-1}, \tag{32}$$

where C_5 is a positive constant depending on $\|g\|_{\infty,V_n(x_0)}$, $\|K\|_\infty$, $\|K\|_2$ and ρ.

Similarly, we introduce $U_i = w(\tilde{\theta}_n,\hat{g}(X_i))K_{h'}\star K_h(x_0-X_i)$ for $i=1,\ldots,n$. Then,

$$\hat{f}_{h,h'}(x_0) - K_{h'}\star(K_h\star\check{f})(x_0) = \hat{f}_{h,h'}(x_0) - \tilde{\mathbb{E}}\big[\hat{f}_{h,h'}(x_0)\big] = \frac{1}{n}\sum_{i=1}^n\big(U_i - \tilde{\mathbb{E}}[U_i]\big),$$

and

$$|U_i| \leqslant \frac{4\,\|K\|_1\,\|K\|_\infty}{h'\delta\hat{\gamma}} =: \bar{b}, \quad \text{and} \quad \tilde{\mathbb{E}}\big[U_1^2\big] \leqslant \frac{16\,\|K\|_1^2\,\|K\|_2^2\,\|g\|_{\infty,V_n(x_0)}}{h'\delta^2\hat{\gamma}^2} =: \bar{v}.$$

Following the same lines as for obtaining (32), we get by using Bernstein inequality

$$\tilde{\mathbb{E}}[D_2] \leqslant C_6\delta^{-2}\gamma^{-2}n^{-1}, \tag{33}$$

with C_6 a positive constant depends on $\|g\|_{\infty,V_n(x_0)}$, $\|K\|_\infty$, $\|K\|_1$, $\|K\|_2$ and ρ.

Finally, combining (30), (32), (33) and successively applying Lemma 3 and Lemma 4 allow us to conclude the result stated in Theorem 1.

6.5 Proof of Corollary 1

Assume that Assumptions **(A6)** and **(A7)** are fulfilled. According to Proposition 1.2 of Tsybakov [30], we get for all $x_0 \in [0,1]$

$$|K_h \star f(x_0) - f(x_0)| \leqslant C_7 \mathcal{L} h^\beta,$$

where C a constant depending on K and $\mathcal{L}$. Taking

$$h = \mathcal{L}^{-1/\beta} \Lambda_n^{-1/\beta}, \quad \Lambda_n = \mathcal{L}^{-1/(2\beta+1)} \left(\frac{\delta^2 \gamma^2 n}{\log n} \right)^{\beta/(2\beta+1)},$$

we get

$$\frac{\log(n)}{\delta^2 \gamma^2 n h} = \Lambda_n^{-2}.$$

Hence, we obtain

$$\min_{h \in \mathcal{H}_n} \left\{ \|K_h \star f - f\|^2_{\infty, V_n(x_0)} + \frac{\log(n)}{\delta^2 \gamma^2 n h} \right\} \leqslant (C_7^2 + 1) \mathcal{L}^{2/(2\beta+1)} \left(\frac{\delta^2 \gamma^2 n}{\log n} \right)^{-2\beta/(2\beta+1)}. \tag{34}$$

Finally, since we also assume (10), gathering (9) and (34), we obtain

$$\mathbb{E}\left[\left(\hat{f}(x_0) - f(x_0) \right)^2 \right] \leqslant C_8 \left(\frac{\log n}{n} \right)^{\frac{2\beta}{2\beta+1}},$$

where C_8 is a constant depending on K, $\|f\|_{\infty, V_n(x_0)}$, g, δ, γ, ρ, $\mathcal{L}$ and β. This ends the proof of Corollary 1.

6.6 Proofs of technical lemmas

6.6.1 Proof of Lemma 3

Lemma 3 is a consequence of (5). Indeed, assume that condition **(A3)** is satisfied, then we have for all $t \in V_n(x_0)$, $|\hat{g}(t) - g(t)| \leqslant \nu |\hat{g}(t)|$ with probability $1 - C_{g,\nu} \exp\big(- (\log n)^{3/2} \big)$.

This implies,

$$(1+\nu)^{-1} |g(t)| \leqslant |\hat{g}(t)| \leqslant (1-\nu)^{-1} |g(t)|.$$

Since $\gamma = \inf_{t \in V_n(x_0)} |g(t)|$ and $\hat{\gamma} = \inf_{t \in V_n(x_0)} |\hat{g}(t)|$, by using (5) and taking $\nu = \rho - 1$, $\nu = 1 - \rho^{-1}$, we obtain with probability $1 - C_{g,\nu} \exp\big(- (\log n)^{3/2} \big)$, $(1+\nu)^{-1} \gamma \leqslant \hat{\gamma} \leqslant (1-\nu)^{-1} \gamma$ which completes the proof of Lemma 3.

6.6.2 Proof of Lemma 4

We have for any $x_0 \in [0,1]$,

$$\mathbb{E}\left[\left(\hat{f}_h(x_0) - f(x_0) \right)^2 \mathbb{1}_{\Omega_\rho^c} \right] \leqslant 2\mathbb{E}\big[|\hat{f}_h(x_0)|^2 \mathbb{1}_{\Omega_\rho^c} \big] + 2 \|f\|^2_{\infty, V_n(x_0)} \mathbb{P}(\Omega_\rho^c).$$

Using Assumptions **(A6)** and (22), we obtain

$$\begin{aligned}
\mathbb{E}\big[|\hat{f}_h(x_0)|^2 \mathbb{1}_{\Omega_\rho^c} \big] &= \mathbb{E}\left[\left| \frac{1}{nh} \sum_{i=1}^n w(\tilde{\theta}_n, \hat{g}(X_i)) K\left(\frac{x_0 - X_i}{h} \right) \right|^2 \mathbb{1}_{\Omega_\rho^c} \right] \\
&\leqslant \frac{16}{\delta^2} \mathbb{E}\left[\left| \frac{1}{\hat{\gamma} n h} \sum_{i=1}^n K\left(\frac{x_0 - X_i}{h} \right) \right|^2 \mathbb{1}_{\Omega_\rho^c} \right] \\
&\leqslant \frac{16 \|K\|^2_\infty}{\delta^2} \frac{n^2}{(\log n)^6} \mathbb{P}(\Omega_\rho^c).
\end{aligned}$$

Finally, we apply Lemma 3 to establish the following bound

$$\mathbb{E}\left[\left(\hat{f}_h(x_0) - f(x_0)\right)^2 \mathbb{1}_{\Omega_\rho^c}\right] \leqslant C_{g,\rho}\left(\frac{16\,\|K\|_\infty^2}{\delta^2}\frac{n^2}{(\log n)^6} + 2\,\|f\|_{\infty,V_n(x_0)}^2\right)\exp\left\{-(\log n)^{3/2}\right\}$$
$$\leqslant \frac{16C_{g,\rho}\,\|K\|_\infty^2\,\|f\|_{\infty,V_n(x_0)}^2}{\delta^2}\frac{1}{n^2},$$

which ends the proof of Lemma 4.

Acknowledgment

We are very grateful to Catherine Matias for interesting discussions on mixture models. The research of the authors is partly supported by the french Agence Nationale de la Recherche (ANR-18-CE40-0014 projet SMILES) and by the french Région Normandie (projet RIN AStERiCs 17B01101GR).

References

[1] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[2] Karine Bertin, Claire Lacour, and Vincent Rivoirard. Adaptive pointwise estimation of conditional density function. *preprint arXiv:1312.7402*, 2013.

[3] Karine Bertin, Claire Lacour, and Vincent Rivoirard. Adaptive pointwise estimation of conditional density function. *Ann. Inst. H. Poincar Probab. Statist.*, 52(2):939–980, 05 2016.

[4] C. Butucea. Two adaptive rates of convergence in pointwise density estimation. *Math. Methods Statist.*, 9(1):39–64, 2000.

[5] Alain Celisse and Stéphane Robin. A cross-validation based estimation of the proportion of true null hypotheses. *Journal of Statistical Planning and Inference*, 140(11):3132–3147, 2010.

[6] Gaëlle Chagny. Penalization versus goldenshluger- lepski strategies in warped bases regression. *ESAIM: Probability and Statistics*, 17:328–358, 2013.

[7] Michaël Chichignoud, Van Ha Hoang, Thanh Mai Pham Ngoc, Vincent Rivoirard, et al. Adaptive wavelet multivariate regression with errors in variables. *Electronic journal of statistics*, 11(1):682–724, 2017.

[8] F. Comte, S. Gaïffas, and A. Guilloux. Adaptive estimation of the conditional intensity of marker-dependent counting processes. *Ann. Inst. Henri Poincaré Probab. Stat.*, 47(4):1171–1196, 2011.

[9] F. Comte and C. Lacour. Anisotropic adaptive kernel deconvolution. *Ann. Inst. Henri Poincaré Probab. Stat.*, 49(2):569–609, 2013.

[10] Fabienne Comte. *Estimation non-paramétrique*. Spartacus-IDH, 2015.

[11] Fabienne Comte, Valentine Genon-Catalot, and Adeline Samson. Nonparametric estimation for stochastic differential equations with random effects. *Stochastic Processes and their Applications*, 123(7):2522–2551, 2013.

[12] Fabienne Comte and Tabea Rebafka. Nonparametric weighted estimators for biased data. *Journal of Statistical Planning and Inference*, 174:104–128, 2016.

[13] Marie Doumic, Marc Hoffmann, Patricia Reynaud-Bouret, and Vincent Rivoirard. Nonparametric estimation of the division rate of a size-structured population. *SIAM Journal on Numerical Analysis*, 50(2):925–950, 2012.

[14] Bradley Efron, Robert Tibshirani, John D Storey, and Virginia Tusher. Empirical bayes analysis of a microarray experiment. *Journal of the American statistical association*, 96(456):1151–1160, 2001.

[15] Christopher Genovese and Larry Wasserman. Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):499–517, 2002.

[16] Evarist Giné and Richard Nickl. An exponential inequality for the distribution function of the kernel density estimator, with applications to adaptive estimation. *Probab. Theory Related Fields*, 143(3-4):569–596, 2009.

[17] Alexander Goldenshluger and Oleg Lepski. Bandwidth selection in kernel density estimation: orcale inequalities and adaptive minimax optimality. *The Annals of Statistics*, 39(3):1608–1632, 2011.

[18] Peter J Huber. A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, pages 1753–1758, 1965.

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

[20] Mette Langaas, Bo Henry Lindqvist, and Egil Ferkingstad. Estimating the proportion of true null hypotheses, with application to dna microarray data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(4):555–572, 2005.

[21] Oleg Lepski et al. Multivariate density estimation under sup-norm loss: oracle approach, adaptation and independence structure. *The Annals of Statistics*, 41(2):1005–1034, 2013.

[22] Haoyang Liu and Chao Gao. Density estimation with contaminated data: Minimax rates and theory of adaptation. *arXiv preprint arXiv:1712.07801*, 2017.

[23] Van Hanh Nguyen and Catherine Matias. Nonparametric estimation of the density of the alternative hypothesis in a multiple testing setup. application to local false discovery rate estimation. *ESAIM: Probability and Statistics*, 18:584612, 2014.

[24] Van Hanh Nguyen and Catherine Matias. On efficient estimators of the proportion of true null hypotheses in a multiple testing setup. *Scandinavian Journal of Statistics*, 41(4):1167–1194, 2014.

[25] Patricia Reynaud-Bouret, Vincent Rivoirard, Franck Grammont, and Christine Tuleau-Malot. Goodness-of-fit tests and nonparametric adaptive estimation for spike train analysis. *The Journal of Mathematical Neuroscience*, 4(1):1, 2014.

[26] Stéphane Robin, Avner Bar-Hen, Jean-Jacques Daudin, and Laurent Pierre. A semi-parametric approach for mixture models: Application to local false discovery rate estimation. *Computational Statistics & Data Analysis*, 51(12):5483–5493, 2007.

[27] Eugene F Schuster. Incorporating support constraints into nonparametric estimators of densities. *Communications in Statistics-Theory and methods*, 14(5):1123–1136, 1985.

[28] John D Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

[29] Korbinian Strimmer. A unified approach to false discovery rate estimation. *BMC bioinformatics*, 9(1):303, 2008.

[30] Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer series in Statistics. Springer, New York, 2009.