Nonparametric weighted estimators for biased data

Fabienne Comte, Tabea Rebafka

COMPARISON OF SOME RECENT METHODS IN ADAPTIVE DENSITY ESTIMATION FOR BIASED DATA

FABIENNE COMTE[(1)], TABEA REBAFKA[(2)]

Abstract. Several adaptive methods to estimate a density from biased data are presented. Risk bounds for the estimators are provided and an empirical study is performed to compare various kernel and projection estimators associated with different adaptation methods, namely Lepski-type bandwidth selection in pointwise and global settings and model selection for projection estimators. A real data example taken from fluorescence lifetime measurements is also studied.

Keywords. Adaptive density estimation. Biased data. Bandwidth selection. Fluorescence lifetimes. January 9, 2015

1. Introduction

In various application settings, functional estimation can be difficult because the observed data are not a sample from the distribution of interest: this may be due to noise, missing data, censored or truncated observations. In this paper biased data models are considered where the observed distribution is the result of a (known) nonlinear distortion of the distribution of interest.

More precisely, we observe a sample $Z_1, \dots, Z_n$ of independent identically distributed (i.i.d.) random variables with probability density function (pdf) g and cumulative distribution function (cdf) G. The observed distribution G is related to the distribution of interest, say F, by some known link function H by the following relation

$$G(z) = H \circ F(z) , \qquad z \in \mathbb{R} . \tag{1}$$

The aim is to recover the pdf f of the distribution of interest F in a nonparametric context using an i.i.d. sample $Z_1, \dots, Z_n$ with distribution G and known link function H.

We have in mind the case where every Z_i is the minimum of a random number N of i.i.d. random variables $Y_1, \dots, Y_N$ with distribution F. An example in physics is the arrival time of the fasted of a random number of emitted photons (Rebafka et al., 2010), or in biostatistics the time to the observation of a tumor originated from a clonogenic cell in the presence of a random number of competing clonogens (Tsodikov, 2001). Various extensions and other examples may be considered. For example, changing the random minimum in a random maximum corresponds in actuarial science to modelling the largest claim received by an insurer in a given time interval (Li and Zuo, 2004), or in transportation theory to the modelling of the maximal accident-free distance of a shipment of, say, explosives, with a random number of defective explosives which may explode and cause an accident during transport (Shaked and Wong, 1997).

(1): MAP5, UMR 8145 CNRS, Sorbonne Paris Cité, Paris Descartes University, France, email: fabienne.comte@parisdescartes.fr
(2): LPMA, University of Paris 6, UPMC, France, email: tabea.rebafka@upmc.fr.
The authors wish to thank PicoQuant GmbH, Berlin, Germany for kindly providing the TCSPC data.

From a methodological point of view, we propose projection and kernel estimators associated with model and bandwidth selection devices. Two properties that hold in the model given by (1) give rise to two different ways of correcting the bias in the data. Hence two different estimation strategies can be used to construct both kernel and projection estimators and they are worth being compared. The mean-square risks of the estimators are studied and oracle-type risk bounds are provided. Adaptive projection estimators correspond to methods originally described by Barron et al. (1999) and applied to survival analysis and biased data by Efromovich (2004a,b) and Brunel et al. (2005); more recently, wavelet projection estimators have been studied by Chesneau (2010), Cutillo et al. (2014). For the bandwidth selection of the kernel estimators the recent approach of Goldenshluger and Lepski (2011) is applied to our model and studied from both a pointwise and a global point of view, and our results on this side, namely finite sample risk bounds for adaptive kernel estimators, are new.

It is worth mentioning that our model can be related to other biased data contexts, which have been studied from other or specific point of view by several authors: strategies for estimating cumulative distribution functions are proposed by Gill et al. (1988), Wu and Mao (1996), Wu (1997), Efromovich (2004b), El Barmi and Simonoff (2000); the specific case of length-biased sampling has been studied in a lot of papers, see Vardi (1982), Jones (1991), de Uña-Álvarez (2004), de Uña-Álvarez and Rodríguez-Casal (2006) Asgharian et al. (2002), among others.

A simulation study is performed to calibrate and compare all those methods. Several questions are in order: What is a good choice of the penalty constants? Can we adapt the existing proposals for model selection to the kernel methods? How do the pointwise and the global strategy compare in specific examples? As each estimator involves a bias correction, is there one that outperforms the others? In our simulation study we focus on the so-called pile-up model, which is used for fluorescence lifetime measurements (O'Connor and Phillips, 1984) and presented in detail in the experimental section of the paper. An application to real fluorescence data is also provided.

The paper is organized as follows. Section 2 presents the model. In Section 3 adaptive kernel and projection estimators are proposed. Section 4 gives theoretical results on risk bounds of the different estimators. In the simulation study (Section 5) different aspects of the estimators are compared. Finally, Section 6 presents the proofs for the theoretical results of the paper.

2. Model and Assumptions

2.1. **Notations.** For two functions u and v, we denote by $u\circ v$ the function $x \mapsto u\circ v(x) := u(v(x))$. If u is a one-to-one map, we denote by u^{-1} the inverse of the function u, that is the function such that $(u^{-1}\circ u)(x) = (u\circ u^{-1})(x) = x$ for all x. The derivative of u is denoted by $\dot{u}$ and the second-order derivative by $\ddot{u}$, provided that they exist. The standard convolution product is given by $u * v(x) = \int u(t)v(x-t)\mathrm{d}t$. Furthermore, we denote by $\| \ . \ \|_p$ the $\mathbb{L}^p$–norm given by $\|u\|_p^p = \int |u(x)|^p \mathrm{d}x$ and by $\| \ . \ \|_\infty$ the $\mathbb{L}^\infty$–norm, $\|u\|_\infty = \sup_{x\in\mathbb{R}} |u(x)|$.

2.2. **Model and assumptions.** The link function $H : [0,1] \to [0,1]$ in relation (1) is necessarily increasing and surjective so that $H \circ F$ is a cdf for any cdf F. Note that even if H is not injective, G given by (1) may still be a cdf. However, our goal is to recover the density of distribution F using a sample from G. Hence, we need H to be bijective to

ensure identifiability of the model. Indeed, when H is a one-to-one map, then

$$F(z) = H^{-1} \circ G(z) \ , \qquad z \in \mathbb{R} \ . \tag{2}$$

Furthermore, we assume that H is differentiable, since in our context both F and G are supposed to be absolutely continuous with pdf f and g respectively. Then, deriving relation (1) implies that the densities f and g are related by

$$g(z) = \dot{H} \circ F(z) \ f(z) \ , \qquad z \in \mathbb{R} \ . \tag{3}$$

Combining (2) and (3) gives $f = g/\dot{H} \circ F = g/\dot{H} \circ H^{-1} \circ G$. Define the weight function w by

$$w(u) = \frac{1}{\dot{H} \circ H^{-1}(u)} \ , \qquad u \in [0,1] \ ,$$

then we obtain

$$f(z) = w \circ G(z) \ g(z) \ , \qquad z \in \mathbb{R} \ . \tag{4}$$

The weight function w is well defined under the assumption that $\dot{H}$ is bounded away from zero. Furthermore, we shall require that w is Lipschitz. This is ensured if there exist finite constants $a, b > 0$ such that

$$\dot{H}(u) \geq a \ , \quad |\ddot{H}(u)| \leq b \ , \qquad u \in [0,1] \ , \tag{5}$$

and a finite constant $d > 0$ such that

$$\dot{H}(u) \leq d \ , \qquad u \in [0,1] \ . \tag{6}$$

Indeed then $1/d \leq w(u) \leq 1/a$ and $\dot{w}(u) = -\ddot{H} \circ H^{-1}(u)/[\dot{H} \circ H^{-1}(u)]^3$, so that the Lipschitz constant of w, say c_w, is such that $c_w \leq b/a^3$. Moreover, we may possibly require that f or g is bounded. Note that f and g are either both bounded or both unbounded, since $a\|f\|_\infty \leq \|g\|_\infty \leq d\|f\|_\infty$.

From relation (4) follows a fundamental property that holds for any measurable bounded function ψ,

$$\mathbb{E}[\psi(Y)] = \mathbb{E}\left[\psi(Z) \ w \circ G(Z)\right] \ , \tag{7}$$

where Y has distribution F and Z is distributed as G. This relation is the basis for the construction of moment estimators of any quantity $\mathbb{E}[\psi(Y)]$ based on i.i.d. observations $Z_1, \ldots, Z_n$ from the distorted distribution G. Replacing the cdf G by its empirical version $\hat{G}_n(z) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{Z_i \leq z\}$ yields a natural estimator of $\mathbb{E}[\psi(Y)]$ given by

$$\hat{L} = \frac{1}{n}\sum_{i=1}^n \psi(Z_i) \ w \circ \hat{G}_n(Z_i) \ . \tag{8}$$

This is a useable estimator since the link function H is supposed to be known and so is the weight function w.

Denote by $Z_{(i)}$ the i-th order statistic associated with $(Z_1, \ldots, Z_n)$ satisfying $Z_{(1)} \leq \cdots \leq Z_{(n)}$. Note that $w \circ \hat{G}_n(Z_{(i)}) = w(i/n)$. Then, we can rewrite $\hat{L}$ as

$$\hat{L} = \frac{1}{n}\sum_{i=1}^n \psi(Z_{(i)}) \ w\left(\frac{i}{n}\right) \ . \tag{9}$$

We see that $\hat{L}$ takes the form of a so-called L-statistic, i.e. a linear combination of order statistics.

3. Adaptive density estimators

3.1. **Estimation strategies.** Here several strategies to estimate f are presented, namely kernel and projection estimators. All estimators make use of relation (4). The first strategy consists in first estimating g from the data, and then multiplying this estimate by $w(\hat{G}_n(z))$ to correct the bias. This estimator is referred to as the *plug-in estimator*. As an estimator of g we use a kernel estimator, for which a bandwidth selection method is provided as well. Concretely, let $\hat{g}_h$ be the standard kernel estimate of g given by

$$\hat{g}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - Z_i) \,, \tag{10}$$

where K is a kernel, that is, an integrable function such that $\int K(u)\mathrm{d}u = 1$, h is a bandwidth parameter and $K_h(u) = h^{-1}K\left(u/h\right)$. Then plugging $\hat{g}_h$ into relation (4) yields the estimator $\hat{f}_h^{(1)}(x)$ of $f(x)$ defined by

$$\hat{f}_h^{(1)}(x) = \hat{g}_h(x) w\left(\hat{G}_n(x)\right) \,. \tag{11}$$

The second method is a kernel estimator as well, however we directly estimate f by using property (7) and an L-statistic of the form of (8). By taking $\psi = K_h(x - \cdot)$ in (8), a kernel estimator of the target density f is obtained by

$$\hat{f}_h^{(2)}(x) = \frac{1}{n}\sum_{i=1}^{n} w \circ \hat{G}_n(Z_i) K_h(x - Z_i) = \frac{1}{n}\sum_{i=1}^{n} w\left(\frac{i}{n}\right) K_h(x - Z_{(i)}) \,. \tag{12}$$

Indeed, the plug-in method as well as property (7) can also be used to construct projection estimators. Here the approach consists in approximating the orthogonal projection of g (or f) onto some function space. More precisely, suppose that the restriction of g (resp. f) on some interval A is square integrable, that is $g\mathbb{1}_A \in \mathbb{L}^2(A)$ (resp. $f\mathbb{1}_A \in \mathbb{L}^2(A)$). Let $(\varphi_j)_{0\le j\le 2m}$ be the trigonometric basis $(\varphi_j)_{0\le j\le 2m}$ on $A = [a, b]$ defined by $\varphi_j = (b - a)^{-1/2}\varphi_j^0((x - a)/(b - a))$ and $\varphi_0^0(x) = \mathbb{1}_{[0,1]}(x)$, $\varphi_{2j+1}^0(x) = \sqrt{2}\cos(2\pi jx)\mathbb{1}_{[0,1]}(x)$ for $j \ge 0$, $\varphi_{2j}(x) = \sqrt{2}\sin(2\pi jx)$ for $j \ge 1$. Then define the subspace $S_m = \mathrm{Span}(\varphi_j, j = 0, 1, \ldots, 2m)$ and $D_m = \dim(S_m) = 2m + 1$. Then, the orthogonal projection g_m (resp. f_m) in the $\mathbb{L}^2$–sense of g (resp. f) on S_m is given by $g_m = \sum_{j=0}^{2m} a_j^{(1)}\varphi_j$ with $a_j^{(1)} = \langle g, \varphi_j\rangle$ (resp. $f_m = \sum_{j=0}^{2m} a_j^{(2)}\varphi_j$ with $a_j^{(2)} = \langle f, \varphi_j\rangle$).

Using the plug-in method, we obtain the projection-type estimate $\hat{f}_m^{(1)}$ defined by

$$\hat{f}_m^{(1)}(x) = \hat{g}_m(x) w\left(\hat{G}_n(x)\right) \,, \quad \text{with} \quad \hat{g}_m = \sum_{j=0}^{2m} \hat{a}_j^{(1)}\varphi_j \quad \text{and} \quad \hat{a}_j^{(1)} = \frac{1}{n}\sum_{i=1}^{n}\varphi(Z_i) \,.$$

By using property (7), a second projection-type estimator of f is given by

$$\hat{f}_m^{(2)} = \sum_{j=0}^{2m} \hat{a}_j^{(2)}\varphi_j \,, \quad \text{with} \quad \hat{a}_j^{(2)} = \frac{1}{n}\sum_{i=1}^{n}\varphi_j(Z_{(i)}) w(i/n) \,. \tag{13}$$

Next, we propose different data-driven bandwidth and model selection methods.

3.2. **Bandwidth selection.** For a data-driven choice of bandwidth h, consider a finite collection $\mathcal{H}$ of bandwidths given by

$$\mathcal{H} = \left\{ h_k, k = 1, \dots, H_n, \frac{1}{n} \le h_k \le 1 \right\} , \quad \text{with } H_n \le n . \tag{14}$$

First an adaptive pointwise estimator of $f(x_0)$ for some fixed x_0 is presented. As suggested in Goldenshluger and Lepski (2011), define the estimators $\hat{g}_{h,h'}$ and $\hat{f}^{(2)}_{h,h'}$ depending on two bandwidths by

$$\hat{g}_{h,h'}(x) = K_{h'} * \hat{g}_h(x) \quad \text{and} \quad \hat{f}^{(2)}_{h,h'}(x) = K_{h'} * \hat{f}^{(2)}_h(x) .$$

Notice the symmetry of both estimators in h and h'. Next we denote, for $i = 1, 2$,

$$\begin{aligned} &V_0^{(i)}(h) = \kappa_0^{(i)} C_i \|K\|_1^2 \|K\|_2^2 \frac{\log n}{nh} , \quad C_1 = \|g\|_\infty , \quad C_2 = \|f\|_\infty / a , \\ &A_0^{(i)}(h, x_0) = \sup_{h' \in \mathcal{H}} \left[\left(\hat{\theta}^{(i)}_{h,h'}(x_0) - \hat{\theta}^{(i)}_{h'}(x_0) \right)^2 - V_0^{(i)}(h') \right]_+ , \quad \theta^{(1)} = g , \quad \theta^{(2)} = f^{(2)} , \\ &\hat{h}^{(i)}(x_0) = \arg\min_{h \in \mathcal{H}} \left\{ A_0^{(i)}(h, x_0) + V_0^{(i)}(h) \right\} . \end{aligned} \tag{15}$$

The numerical constants $\kappa_0^{(i)}$ are typically calibrated by simulation. The other constants are known, except $\|g\|_\infty$ or $\|f\|_\infty$ which in practice are replaced by some estimators (see Section 5.2).

Now two adaptive pointwise estimators of $f(x_0)$ for some fixed x_0 are obtained by

$$\hat{f}^{(1)}_{\hat{h}^{(1)}(x_0)}(x_0) = \hat{g}_{\hat{h}^{(1)}(x_0)}(x_0) w(\hat{G}_n(x_0)) \quad \text{and} \quad \hat{f}^{(2)}_{\hat{h}^{(2)}(x_0)}(x_0) . \tag{16}$$

The risk bounds associated with these estimators are given in Theorem 4.1.

In a similar way, a procedure for global bandwidth selection is developed. Denote for $i = 1, 2$,

$$V^{(i)}(h) = \kappa_1^{(i)} D_i \frac{\max(\|K\|_1^2, 1)\|K\|_2^2}{nh} , \quad D_1 = 1 , D_2 = \frac{1}{a} , \tag{17}$$

$$A^{(i)}(h) = \sup_{h' \in \mathcal{H}} \left(\|\hat{\theta}^{(i)}_{h,h'} - \hat{\theta}^{(i)}_{h'}\|^2 - V^{(i)}(h') \right)_+ , \quad \text{with } \theta^{(1)} = g , \quad \theta^{(2)} = f^{(2)} , \tag{18}$$

$$\hat{h}^{(i)} = \arg\min_{h \in \mathcal{H}} \left(A^{(i)}(h) + V^{(i)}(h) \right) , \tag{19}$$

where $\kappa_1^{(i)}$ are numerical constants. Then define the adaptive estimators

$$\hat{f}^{(1)}_{\hat{h}^{(1)}}(x) = \hat{g}_{\hat{h}^{(1)}}(x) w(\hat{G}_n(x)) \quad \text{and} \quad \hat{f}^{(2)}_{\hat{h}^{(2)}} ,$$

for which a risk bound is given in Theorem 4.2.

3.3. **Model selection.** For the projection estimators the classical penalization approach by Barron et al. (1999) can be applied. In Section 4 we show that a bias-variance trade-off is achieved. Define penalty terms $\mathrm{pen}^{(i)}(m)$ by

$$\mathrm{pen}^{(1)}(m) = \kappa_2^{(1)} \frac{D_m}{n} , \quad \mathrm{pen}^{(2)}(m) = \kappa_2^{(2)} \|w\|_2^2 \frac{D_m}{n} , \tag{20}$$

with appropriate constants $\kappa_2^{(i)}$, for $i = 1, 2$. Then select model $\hat{m}^{(i)}$ given by

$$\hat{m}^{(1)} = \arg\min_{m \in \mathcal{M}_n} \left[-\|\hat{g}_m\|^2 + \mathrm{pen}^{(1)}(m) \right] , \; \hat{m}^{(2)} = \arg\min_{m \in \mathcal{M}_n} \left[-\|\hat{f}^{(2)}_m\|^2 + \mathrm{pen}^{(2)}(m) \right] , \tag{21}$$

and consider the density estimates $\hat{f}^{(i)}_{\hat{m}^{(i)}}$, for $i = 1, 2$.

4. Theoretical results on the estimators

In this section we first provide results for the kernel estimators with super-index $^{(2)}$. Then the flavor of results for kernel estimators with super-index $^{(1)}$ are discussed. Lastly, we study the projection estimators. All proofs are relegated to Section 6.

4.1. Study of the pointwise and integrated risk of kernel estimators.

Classically, for f or g we consider Hölder classes $\Sigma(\beta, C)$ defined as

$$\Sigma(\beta, C) = \{f : \mathbb{T} \to \mathbb{T}, f^{(\ell)} \text{exists for } \ell = \lfloor\beta\rfloor \text{and } |f^{(\ell)}(x) - f^{(\ell)}(x')| \leq C|x - x'|^{\beta-\ell}, \forall x, x' \in \mathbb{T}\},$$

where $\mathbb{T} \subset \mathbb{R}$. The set $\mathbb{T}$ may be an interval or the entire set $\mathbb{R}$ depending on the support of the kernel. For instance, if K has compact support $[-1, 1]$, the set $\mathbb{T}$ for a given point x_0 under study and bandwidth less than 1 can be taken equal to $[x_0 - 1, x_0 + 1]$.

The following assumptions are useful to provide risk bounds.

(H1) f belongs to the Hölder class $\Sigma(\beta, C)$.

(H2) K is a kernel of order $\ell = \lfloor\beta\rfloor$ satisfying $\int |u|^\beta |K(u)|\mathrm{d}u < \infty$. Furthermore, $\int u^2 K(u)\mathrm{d}u < \infty$ and $\|K\|_\infty < \infty$.

We recall that a kernel of order ℓ satisfies $\int x^k K(x)\mathrm{d}x = 0$ for $k = 1, \ldots, \ell$. Assumption (H2) is classical.

The following proposition is easily shown, and proved in Section 6.

Proposition 4.1. *If f is bounded and (H1)-(H2), (5) and (6) are fulfilled, then, for any x_0, the estimator $\hat{f}^{(2)}_h$ defined by (12) satisfies*

$$\mathbb{E}[(\hat{f}^{(2)}_h(x_0) - f(x_0))^2] \leq 3C_1^2 h^{2\beta} + \frac{C_2}{nh},$$

where $C_1 = C \int |x|^\beta |K(u)|\mathrm{d}u/\ell!$ and $C_2 = 3\left(1/a + 5db^2/(4a^6)\right)\|f\|_\infty \|K\|_2^2$.

An upper bound of the mean integrated squared error (MISE) of $\hat{f}_h$ can be easily derived for densities f belonging to the Nikol'ski class. Let $\beta > 0$ and $L > 0$. Define the Nikol'ski class $\mathcal{N}(\beta, L)$ as

$$\mathcal{N}(\beta, L) = \left\{ f : \mathbb{R} \to \mathbb{R} : \left[\int \left(f^{(\ell)}(x + t) - f^{(\ell)}(x) \right)^2 \mathrm{d}x \right]^{1/2} \leq L|t|^{\beta-\ell}, \quad \forall t \in \mathbb{R} \right\}.$$

Proposition 4.2. *Let $\beta > 0$ and $L > 0$. Assume that f is a square integrable density function belonging to the Nikol'ski class $\mathcal{N}(\beta, L)$ and that (2) and (5) hold. Then the MISE of $\hat{f}^{(2)}_h$ defined by (12) satisfies*

$$\mathbb{E}\left[\|\hat{f}^{(2)}_h - f\|_2^2\right] \leq 3C_3^2 h^{2\beta} + \frac{C_4}{nh},$$

where $C_3 = [L/(\ell - 1)!] \int |u|^\beta |K(u)|\mathrm{d}u$ and $C_4 = 3(b^2/a^6 + 1/a)\|K\|_2^2$.

Risk bounds for the plug-in kernel estimators are a consequence of usual density estimation results or of the above bounds in the particular case of $w \equiv 1$, and of the following inequality.

$$\begin{aligned} \left[\hat{f}^{(1)}_h(x_0) - f(x_0)\right]^2 &= \left[w \circ \hat{G}_n(x_0)(\hat{g}_h(x_0) - g(x_0)) + (w \circ \hat{G}_n(x_0) - w \circ G(x_0))g(x_0)\right]^2 \\ &\leq \frac{2}{a^2}\left[\hat{g}_h(x_0) - g(x_0)\right]^2 + \frac{2b^2}{a^6} g^2(x_0)\left[\hat{G}_n(x_0) - G(x_0)\right]^2. \end{aligned} \tag{22}$$

Since $\mathbb{E}([\hat{G}_n(x_0) - G(x_0)]^2) \leq 1/n$,

$$\mathbb{E}\left([\hat{f}_h^{(1)}(x_0) - f(x_0)]^2\right) \leq \frac{2}{a^2}\mathbb{E}\left([\hat{g}_h(x_0) - g(x_0)]^2\right) + \frac{2b^2 g^2(x_0)}{a^6}\frac{1}{n} .$$

Thus, Proposition 4.1 holds for $\hat{f}_h^{(1)}$, for different constants C_1, C_2, under assumptions on g. Analogously, in the integrated case, we get

$$\mathbb{E}\left[\|\hat{f}_h^{(1)} - f\|_2^2\right] \leq \frac{2}{a^2}\mathbb{E}\left[\|\hat{g}_h - g\|_2^2\right] + 2\frac{b^2\|g\|_2^2}{a^6}\frac{1}{n} .$$

It is noteworthy that here the regularity parameter involved in the bounds is related to g instead of f.

Clearly, both pointwise and integrated risks admit a bias-variance decomposition with standard terms. If the bandwidth could be chosen of order $n^{-1/(2\beta+1)}$, where β is the Hölder or the Nikol'ski regularity index, then the resulting rate of the estimators would be of order $n^{-2\beta/(2\beta+1)}$. As β is unknown, this compromise cannot be performed in that naive way. For this reason, we propose data-driven methods for bandwidth selection.

4.2. **Adaptive pointwise kernel estimation.** Now we consider bandwidth collection satisfying the following assumption.

(H3) The collection $\mathcal{H}$ of bandwidths is a finite set given by

$$\mathcal{H} = \left\{h_k, k = 1, \ldots, H_n, \frac{1}{n} \leq h_k \leq 1\right\} , \text{ with } H_n \leq n , \tag{23}$$

and there exists some finite constant S (independent of n) such that

$$\frac{1}{n}\sum_{h\in\mathcal{H}}\frac{1}{h} \leq S . \tag{24}$$

The definition of the bandwidth collection $\mathcal{H}$ via (23) is very general, only condition (24) requires some comments and illustration. We give some examples satisfying (24).

- (C1) For $a \in (0, 1)$, the bandwidth collection $\mathcal{H} = \{h_k = (k/n)^a, k = 1, \ldots, n\}$ satisfies (H3) with $H_n = n$ and $S = 1/(1-a)$. Note that the case $a = 1$ does not fulfill (H3), but $S = \log(n)$ would be admissible provided that $\log(n)/n$ is replaced by $\log^2(n)/n$ in inequality (25) below.
- (C2) The collection $\mathcal{H} = \{h_k = 2^{-k}, k = 1, \ldots, [\log_2(n)]\}$ is another example satisfying (H3) with $H_n = [\log_2(n)]$ and $S = 2$.
- (C3) The collection $\mathcal{H} = \{h_k = 1/k, k = 1, \ldots, [\sqrt{n}]\}$ satisfies (H3) with $H_n = [\sqrt{n}]$ and $S = 1$.

Now the following result holds for the estimator $\hat{f}^{(2)}_{\hat{h}^{(2)}(x_0)}(x_0)$.

Theorem 4.1. *Assume that (5), (6) and assumptions (H1)-(H3) hold. Then, there are constants $C^*, \bar{C} > 0$ depending only on $a, b, \|f\|_\infty, S$, $\|K\|_1$, $\|K\|_2$, $\|K\|_\infty$ and on the Hölder class parameters β and C such that*

$$\mathbb{E}\left[\left(\hat{f}^{(2)}_{\hat{h}^{(2)}(x_0)}(x_0) - f(x_0)\right)^2\right] \leq C^* \inf_{h\in\mathcal{H}}\left(h^{2\beta} + V_0^{(2)}(h)\right) + \bar{C}\frac{\log n}{n} . \tag{25}$$

Assumption (H1) requires the regularity of the density f, however for our estimation procedure the order β and the constant C need not to be known.

The right-hand side of (25) is of order $(n/\log(n))^{-2\beta/(2\beta+1)}$ provided that $\mathcal{H}$ contains bandwidths h_k of order $n^{-1/(2\beta+1)}$. This is the case for collections [C1] and [C2], and also

for [C3] if $\beta \geq 1/2$. Then (25) implies an almost optimal compromise between the two terms of the bound given in Proposition 4.1: the bias-variance trade-off is realized, with a loss of order $\log(n)$, which is classical for pointwise adaptive procedures. In density estimation (corresponding to $w \equiv 1$), such a loss is known to be unavoidable and thus adaptive minimax (see Butucea (2000)).

4.3. **Adaptive global kernel estimator.** For the integrated risk the following result holds.

Theorem 4.2. *Assume that (5) and assumptions (H2)-(H3) hold, and that the bandwidth collection $\mathcal{H}$ is such that for any $c > 0$, there exists a finite constant $\Sigma(c)$ (independent of n) such that*

$$\sum_{h\in\mathcal{H}} \mathrm{e}^{-c/h} \leq \Sigma(c) \ . \tag{26}$$

Denote $f_h = K_h * f$. Then there exists a constant $\kappa_1^{(2)}$ such that

$$\mathbb{E}\left[\|\hat{f}^{(2)}_{\hat{h}^{(2)}} - f\|_2^2\right] \leq C \inf_{h\in\mathcal{H}} \left\{\|f - f_h\|_2^2 + V(h)\right\} + C' \frac{\log n}{n} \ ,$$

where C is a numerical constant and C' is a constant depending on a, b, $\|f\|_\infty$, S, $\|K\|_1$, $\|K\|_2$, $\|K\|_\infty$.

It is worth emphasizing that this inequality proves that a bias-variance trade-off is achieved in a nonasymptotic way and without any regularity assumption on f. Moreover, there is no additional $\log(n)$ factor in the definition of $V(h)$, contrary to $V_0(h)$. As a consequence, if $f \in \mathcal{N}(\beta, L)$, where β and L need not to be known, then $\|f - f_h\|_2^2 \leq Ch^{2\beta}$, and thus

$$\inf_{h\in\mathcal{H}} \left\{\|f - f_h\|_2^2 + V(h)\right\} = O\left(n^{-2\beta/(2\beta+1)}\right) \ ,$$

provided that $\mathcal{H}$ contains bandwidths of order $n^{-1/(2\beta+1)}$, which is the case in example (C2) as well as in (C3) if $\beta \geq 1/2$. Therefore, the best bias-variance trade-off is automatically achieved by the procedure.

4.4. **Plug-in kernel estimators.** The results for the plug-in kernel estimators are a consequence of the previous ones. It follows from (22) and the bound $\mathbb{E}([\hat{G}_n(x_0) - G(x_0)]^2) \leq 1/n$ that

$$\mathbb{E}\left[(\hat{f}^{(1)}_{\hat{h}^{(1)}(x_0)}(x_0) - f(x_0))^2\right] \leq \frac{2}{a^2}\mathbb{E}\left[(\hat{g}_{\hat{h}^{(1)}(x_0)}(x_0) - g(x_0))^2\right] + \frac{2b^2 g^2(x_0)}{a^6}\frac{1}{n} \ .$$

Clearly, inequality (25) holds for $w \equiv 1$. Furthermore, if g satisfies the assumptions of Theorem 4.1 and belongs to a Hölder class with regularity parameter β^*, the following bound holds

$$\mathbb{E}\left[\left(\hat{f}^{(1)}_{\hat{h}^{(1)}(x_0)}(x_0) - f(x_0)\right)^2\right] \leq C^* \inf_{h\in\mathcal{H}} \left(h^{2\beta^*} + V_0^{(1)}(h)\right) + \bar{C}' \frac{\log n}{n} \ . \tag{27}$$

Therefore, the risk bound on $\hat{f}^{(1)}_{\hat{h}^{(1)}(x_0)}(x_0)$ is an automatic compromise related to the regularity of g and is optimal if f and g belong to the same Hölder space.

Analogously, in the integrated case, we get

$$\mathbb{E}\left[\|\hat{f}^{(1)}_{\hat{h}^{(1)}} - f\|_2^2\right] \leq \frac{2}{a^2}\mathbb{E}\left[\|\hat{g}_{\hat{h}^{(1)}} - g\|_2^2\right] + 2\frac{b^2\|g\|_2^2}{a^6}\frac{1}{n} \ .$$

The result of Theorem 4.2 can be obtained for $\hat{f}^{(1)}_{\hat{h}^{(1)}}$, but with squared bias term related to g instead of f. Thus, it is optimal if f and g belong to the same Nikol'ski space.

4.5. **Projection estimators.** The following risk bound holds for the projection estimator.

Proposition 4.3. *Consider the estimator (13), then*

$$\mathbb{E}\left[\|\hat{f}^{(2)}_m - f\mathbb{1}_A\|_2^2\right] \leq \|f\mathbb{1}_A - f_m\|_2^2 + C\frac{D_m}{n}\ , \tag{28}$$

where $C = 2(1/a + b^2/a^6)$.

To evaluate the bias of a projection estimator, it is common to consider regularity spaces that are different from those used in kernel estimation. Let $f\mathbb{1}_A = f_A$ belong to a ball of some Besov space $\mathcal{B}_{\alpha,2,\infty}(A)$ with $r+1 \geq \alpha$. Then for $\|f_A\|_{\alpha,2,\infty} \leq L$ we have $\|f_A - f_m\|_2^2 \leq C(\alpha, L)D_m^{-2\alpha}$ (Barron et al., 1999, Lemma 12). Thus, choosing $D_{m^*} = O(n^{1/(2\alpha+1)})$ in inequality (28) yields that the mean square risk satisfies $\mathbb{E}(\|\hat{f}_{m^*} - f_A\|_2^2) \leq O(n^{-2\alpha/(2\alpha+1)})$. This rate is known to be optimal in the minimax sense for density estimation for direct observations (Donoho et al., 1996).

Using this approach, the following result can be shown.

Theorem 4.3. *Assume that* $m_n \leq O(\sqrt{n})$ *and that* f *is bounded on* A, *i.e.* $\|f\|_\infty < \infty$. *Then there exists a numerical constant* $\kappa_2^{(2)}$ *such that we have*

$$\mathbb{E}\left[\|f - \hat{f}^{(2)}_{\hat{m}^{(2)}}\|_2^2\right] \leq C \inf_{m\in\mathcal{M}_n}\left(\|f - f_m\|_2^2 + \|w\|_2^2\frac{D_m}{n}\right) + K\frac{\log^2(n)}{n}, \tag{29}$$

where C *is a numerical constant and* K *depends on* a, b, $\|f\|_\infty$ *and the basis.*

Risk bounds of the form (29) are often called oracle inequality. Note that the last term $c\log^2(n)/n$ is clearly negligible with respect to the order of the infimum (in particular, in all Besov cases described above). This result can easily be generalized to other bases, such as piecewise polynomials or wavelets.

The proof of the theorem relies on Talagrand's inequality and follows the line of the proof of Theorem 4.2 in Brunel and Comte (2005). Therefore, only a sketch of the proof is provided in Section 6.

Lastly, using inequality (22) yields that $\hat{f}^{(1)}_{\hat{m}^{(1)}}$ leads to an optimal bias-variance trade-off with respect to density g, and is optimal if f and g belong to the same Besov space.

5. Experimental study

We have at hand six estimators with (nearly-)optimal rates corresponding to different statistical methods that are intrinsically interesting to compare. To this end this section provides a simulation study and numerical results on a real data example in the specific case of the pile-up model.

5.1. **Pile-up model.** In the pile-up model the minimum of a random number of variables is observed. More precisely, let $\{Y_k, k \geq 1\}$ be a sequence of i.i.d. random variables with pdf f and cdf F. Let N be a random variable taking its values in $\mathbb{N}^* = \{1, 2, \dots\}$ independently of this sequence. Let $Z = \min\{Y_1, \dots, Y_N\}$. The cdf of Z denoted by G is related to F by

$$G(z) = H\circ F(z)\ ,\ z \in \mathbb{R}\ , \qquad \text{with}\ \ H(u) = 1-M(1-u)\ \ \text{and}\ \ M(u) = \mathbb{E}\left[u^N\right]\ ,\ u \in [0,1]\ .$$

The estimation problem consists in extracting density f from an i.i.d. sample $Z_1, \dots Z_n$ from the pile-up distribution G, when the distribution of N is known and of Poisson type.

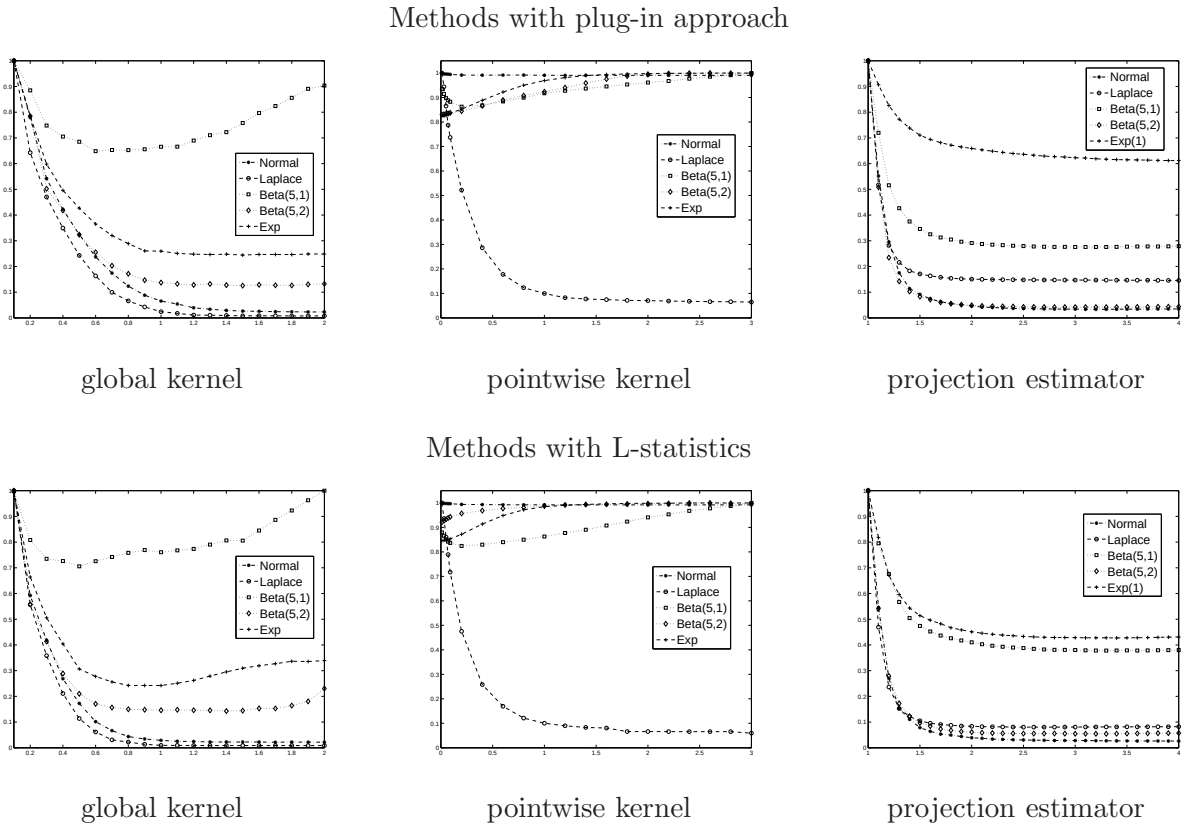

FIGURE 1. (rescaled) MISE for all estimators as a function of κ for different distributions of f.

This problem arises with data in time-resolved fluorescence, where only the shortest arrival time of a group of photons can be observed (O'Connor and Phillips, 1984). Here it is assumed that the size N of a group of photons follows the Poisson distribution restricted to $\mathbb{N}^*$ with known parameter μ. The renormalized probability masses are given by $\mathbb{P}(N = k) = \mu^k/k!/(\mathrm{e}^\mu - 1)$. As μ is known, the link function H is known as well with $M(u) = (\mathrm{e}^{\mu u} - 1)/(\mathrm{e}^\mu - 1)$. We see that the pile-up model is a special case of model (1) and model assumptions (5) and (6) are fulfilled with $a = \mu/(\mathrm{e}^\mu - 1)$, $d = \mu\mathrm{e}^\mu/(\mathrm{e}^\mu - 1)$ and $b = \mu^2/(1 - \mathrm{e}^{-\mu})$. Furthermore, the weight function is given by $w(u) = (1 - \mathrm{e}^{-\mu})/[\mu(u(\mathrm{e}^{-\mu} - 1) + 1)]$.

5.2. **Computational issues and calibration.** We implemented the adaptive pointwise kernel estimators $\hat{f}^{(i)}_{\hat{h}_i(x_0)}(x_0), i = 1, 2$ defined by (16) with $\hat{h}_i(x_0)$ given by (15), the adaptive global kernel estimators $\hat{f}^{(i)}_{\hat{h}^{(i)}}, i = 1, 2$ given by (17)-(19) as well as the adaptive projection estimators $\hat{f}^{(i)}_{\hat{m}^{(i)}}, i = 1, 2$ described in Section 3.3.

In the following simulations the bandwidth collection (C2) is used for the kernel estimators. Indeed, collections (C1) and (C3) are much larger, without leading to proportionally better results. For the kernel estimators we used the gaussian kernel, i.e. $K(u) = \mathrm{e}^{-u^2}/\sqrt{2\pi}$.

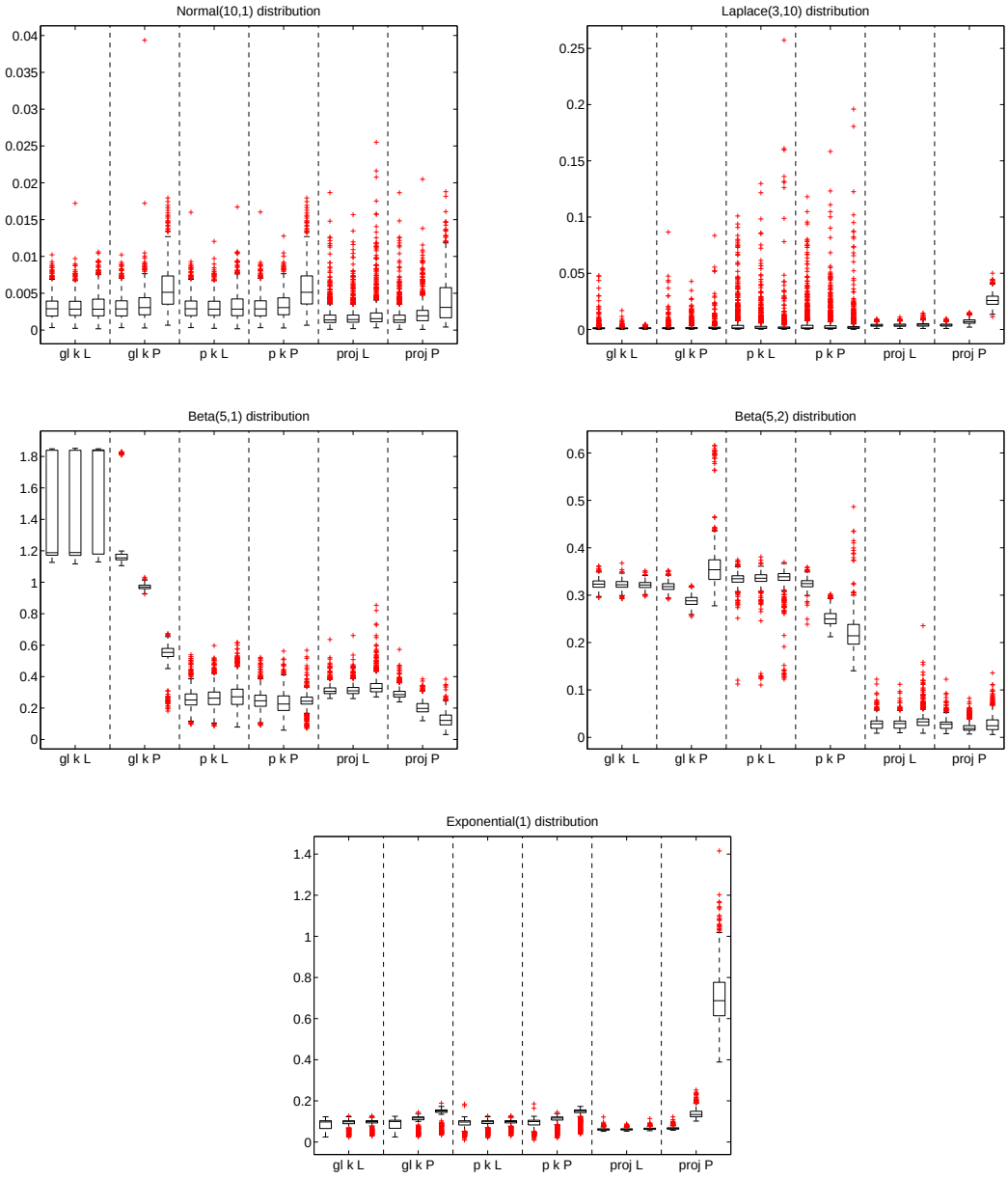

FIGURE 2. Each boxplot represents the values of MISE*1000 of the six estimators computed on 1000 datasets for 5 different distributions with $n = 500$ and for $\mu = 0.1, 0.8, 2$ (from left to right in the figure).

In the quantities $V_0^{(i)}(h), i = 1, 2$ the value of $\|f\|_\infty$ resp. $\|g\|_\infty$ is replaced by estimators. More precisely, $\|g\|_\infty$ is approximated by the 95th percentile of $\{\max_{x_0} \hat{g}_h(x_0), h \in \mathcal{H}\}$. Likewise, $\|f\|_\infty$ is approximated by the 95th percentile of $\{\max_{x_0} \hat{f}_h^{(2)}(x_0), h \in \mathcal{H}\}$.

The terms $\|\hat{\theta}_{h,h'}^{(i)} - \hat{\theta}_{h'}^{(i)}\|^2, i = 1, 2$ in (18) are approximated by Riemann-type discretization.

	Normal $\mathcal{N}(10,1)$ distribution						Laplace $\mathcal{L}(10,3)$ distribution					
μ	0.1		0.8		2		0.1		0.8		2	
n	500	2000	500	2000	500	2000	500	2000	500	2000	500	2000
gl. ker L	3.08	2.76	3.09	2.59	3.23	2.67	**1.76**	**0.940**	**1.14**	**0.579**	**1.22**	**0.416**
gl. ker P	3.09	2.77	3.37	2.80	5.60	4.66	1.93	1.07	2.00	1.44	2.28	1.42
p. ker L	3.09	2.68	3.09	2.65	3.26	2.70	5.50	3.33	4.16	2.66	3.81	1.88
p. ker P	3.10	2.68	3.32	2.83	5.62	4.70	5.70	3.43	5.07	3.51	4.41	2.77
proj L	**1.82**	**0.421**	**1.83**	**0.433**	**2.26**	**0.617**	3.82	2.72	3.92	2.77	4.24	2.87
proj P	1.83	0.426	2.25	0.482	4.03	1.18	3.98	2.89	7.21	6.20	26.2	27.4
	Beta $\mathcal{B}(5,1)$ distribution						Beta $\mathcal{B}(5,2)$ distribution					
μ	0.1		0.8		2		0.1		0.8		2	
n	500	2000	500	2000	500	2000	500	2000	500	2000	500	2000
gl. ker L	1431	1350	1444	1367	1589	1363	324	323	323	323	322	323
gl. ker P	1302	1227	970	971	547	512	318	318	288	288	358	350
p. ker L	258	232	266	236	279	246	334	332	334	334	334	336
p. ker P	**251**	**225**	236	210	247	268	324	321	252	243	219	201
proj L	314	272	317	274	339	279	28.7	11.2	28.5	11.5	34.9	13.8
proj P	292	250	**205**	**166**	**128**	**123**	**27.5**	**10.9**	**22.5**	**10.1**	**28.9**	**11.0**

	Exponential $\mathcal{E}(1)$ distribution					
μ	0.1		0.8		2	
n	500	2000	500	2000	500	2000
gl. ker L	86.1	77.4	89.0	79.3	93.6	84.7
gl. ker P	88.1	79.0	104.6	92.1	140	124
p. ker L	89.0	83.7	91.4	86.1	95.7	90.8
p. ker P	90.7	85.2	106	98.7	142	131
proj L	**62.0**	**53.8**	**62.1**	**53.9**	**64.9**	**54.5**
proj P	67.1	59.0	138	136	704	755

TABLE 1. Mean MISE*1000 values for the six different estimators in 30 different settings.

We noted that the projection estimators are much improved by normalizing $\hat{f}^{(i)}_{\hat{m}^{(i)}}, i = 1, 2$ such that its integral equals one. However, normalization is only appropriate when the interval where the density is estimated covers the main support of the density. For the kernel estimators normalization does not seem to be necessary. In fact, the property is almost automatic if K is a density because then $\int \hat{f}_h(x)\mathrm{d}x = (1/n)\sum_{i=1}^n w(i/n) \simeq \int_0^1 w(u)\mathrm{d}u = 1$.

The different penalty constants $\kappa_j^{(i)}$ are calibrated via simulation. For every estimator the mean integrated squared error $\|\hat{f} - f\|^2$ is approximated on a grid of κ-values for different distributions. The aim is to choose κ such that the MISE is minimized simultaneously for all distributions.

Figure 1 represents the results for the following five distributions for f: normal $\mathcal{N}(10,1)$, Laplace $\mathcal{L}(10,3)$, Beta $\mathcal{B}(5,1)$, Beta $\mathcal{B}(5,2)$ and exponential $\mathcal{E}(1)$ distribution. The Poisson parameter is set to $\mu = 0.8$. For every point of the grid of κ-values, 1000 datasets of sample size 500 are generated, and the 6 estimators and the associated MISE are evaluated. The mean values of the MISE are represented in Figure 1. For the sake of readability, the MISE curves are rescaled such that they are contained in the interval $[0, 1]$. Recall that only the point matters where the MISE attains the minimum, and not the value of the minimum.

One observes that the MISE curves for the methods based on L-statistics (first row) and the plug-in method (second row) are always quite similar. Hence, the same κ may be used for both methods, i.e. $\kappa_j^{(1)} = \kappa_j^{(2)}$ for $j = 1, 2, 3$. However, the MISE curves are quite different from one estimation strategy to another.

The global kernel estimators as well as the projection estimators are rather robust with regard to the choice of κ, as there exists an interval of κ-values where all MISE curves are quite flat and achieve there minimum. In the following we set $\kappa_1^{(1)} = \kappa_1^{(2)} = 1.1$ for the global kernel estimators and $\kappa_2^{(1)} = \kappa_2^{(2)} = 3$ for the projection estimators.

Concerning the pointwise kernel estimators, the MISE is rather sensitive to the value of $\kappa_0^{(i)}$. Slight changes may have a strong impact on the result. Furthermore, the estimators have quite the opposite behavior for the Beta $\mathcal{B}(5, 1)$ and the Laplace distribution. For the former the minimum of the MISE is attained at $\kappa_0^{(i)} = 0.09$, for the latter at $\kappa_0^{(i)} = 5$ (the largest value of $\kappa_0^{(i)}$ considered here). It is not clear which value of $\kappa_0^{(i)}$ achieves the best compromise among all distributions. In the following we set $\kappa_0^{(1)} = \kappa_0^{(2)} = 0.4$.

5.3. **Comparison of all six estimators.** There are several factors potentially influencing the performance of the different estimators. In our simulation study we consider

- two different sample sizes ($n = 500$ and $n = 2000$),
- three levels of the Poisson parameter ($\mu = 0.1, \mu = 0.8$ and $\mu = 2$),
- five different distributions: normal $\mathcal{N}(10, 1)$, Laplace $\mathcal{L}(10, 3)$, Beta $\mathcal{B}(5, 1)$, Beta $\mathcal{B}(5, 2)$, exponential $\mathcal{E}(1)$,

giving a total of 30 settings.

To evaluate the performance of the estimators we proceed exactly as in the calibration study. For each setting the estimators and their MISE are evaluated on 1000 datasets. The boxplots in Figure 2 represent the corresponding results when the sample size is 500. Table 1 shows all means of the MISE in the different settings.

We now analyze the impact of the different factors on the performance of the estimators. Therefore, we study how the mean MISE evolves when one of the factors changes.

Impact of the sample size. As usual, increasing the sample size results in a decrease of the MISE. Interestingly, depending on the estimation strategy and on the type of distribution, this decrease can be more or less pronounced. The increase of the sample size is more beneficial for the projection estimators than for the kernel estimators. The improvement is by far more important for the normal distribution than for the Beta $\mathcal{B}(5, 1)$ or the exponential distribution.

Impact of the Poisson parameter. In the pile-up model the Poisson parameter μ is related to the degree of distortion. In other words, it represents the amount of bias in the model. The three levels of μ considered here correspond to a very low ($\mu = 0.1$), medium ($\mu = 0.8$) and high ($\mu = 2$) degree of distortion. As increasing μ results in a more difficult estimation problem, we observe increasing MISE values in almost all settings, see Figure 2. However, there are some exceptions where the MISE decreases when μ increases. We do not have any explanation for this phenomenon.

Comparison of estimation strategies: Global kernel, pointwise kernel or projection strategy? Depending on the underlying target distribution there are significant differences in the performance of the different estimation strategies.

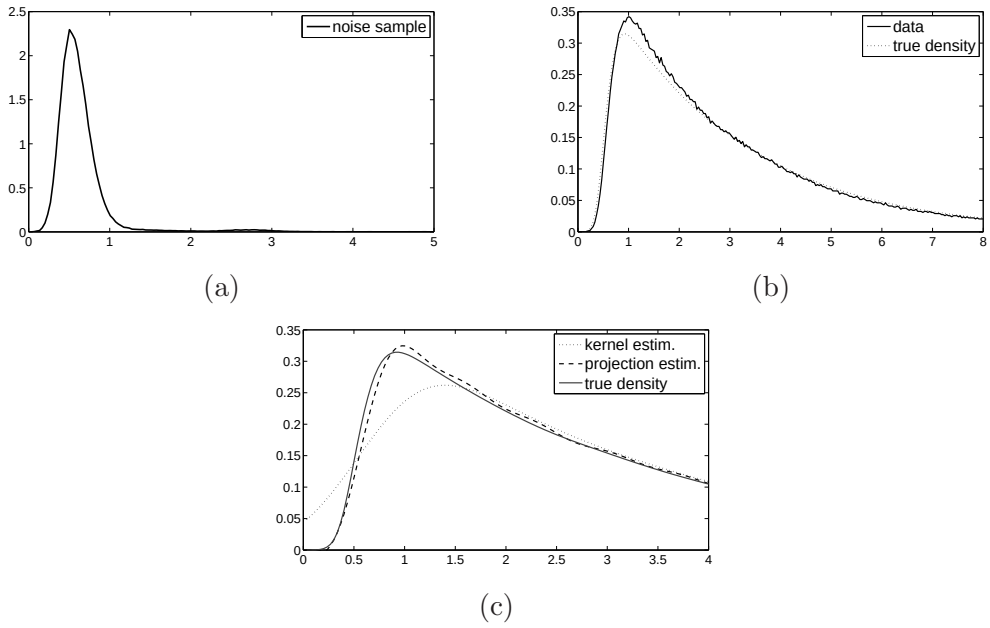

FIGURE 3. (a) Noise sample. (b) Fluorescence data and target density f. (c) Estimators for fluorescence data and target density f.

Both projection methods are best for the normal distribution and the Beta $\mathcal{B}(5,2)$-distribution. For the Beta $\mathcal{B}(5,1)$-distribution they perform similarly good as the pointwise kernel methods. However, in the exponential case there is a stark difference between the projection estimator based on L-statistics, which is best in all settings, and the projection plug-in estimator being by far the worst method for large μ.

The global kernel estimators are best in the Laplace case, and do generally well, except for the Beta distributions.

Pointwise kernel estimators are traditionally known to be well adapted to capture peaks like in the exponential or Laplace distribution. Here, we cannot observe this property, since in almost all settings global strategies (kernel and projection) are doing better.

Consequently, there is no estimation strategy that outperforms the others in all set-ups, but a preference should be given to projection strategies and global kernel estimators.

Comparison of bias correction approaches: L-statistics or plug-in strategy? When $\mu = 0.1$ there is almost no bias in the model. Consequently, no significant difference between the methods is observed. However, for larger μ, differences in the MISE appear for the different bias correction methods. Increasing n amplifies the difference between the methods.

Depending on the type of distribution, one or the other correction approach is preferable. In the case of the normal, Laplace and the exponential distribution, all estimators using L-statistics always lead to better results than the corresponding plug-in estimators. However, in the case of the Beta distributions, the plug-in versions mostly give better results.

5.4. **Application to real data.** We now apply our statistical methods to real fluorescence lifetime measurements. The data are supposed to be generated by the pile-up model with Poisson parameter $\mu = 0.166$. The target density f is known to be the convolution of an exponential density $f_{\mathcal{E}}$ with mean 2.54 ns and some positive noise f_η coming from the measuring instrument, i.e. $f = f_\eta \otimes f_{\mathcal{E}}$. The noise density f_η can be observed separately, so

that we assume that f_η is a known function. Figure 3(a) gives a histogram of an independent noise sample of size 259,386. Note that the same dataset has already been analyzed in **?**, but from a deconvolution point of view, that is the aim was the recovery of the exponential density $f_{\mathcal{E}}$. Here we are interested in the estimation of $f = f_\eta \otimes f_{\mathcal{E}}$.

Figure 3(b) shows the data in form of a histogram with very fine bins and target density f. The sample size is 17,402. Here we clearly observe the pile-up effect, that is the histogram is biased in the sense that mass is shifted to the origin compared to the original density f.

On this dataset the pointwise and global kernel estimators coincide for both strategies (L-statistics and plug-in). The two projection estimators differ only slightly. For illustration, Figure 3(c) displays the plug-in projection estimator and the global kernel estimator based on L-statistics compared to density f. We can see that the projection estimator gives a very good recovery of the target density f, whereas the kernel estimator seems to do too much bias correction. Indeed, the corresponding squared errors $\|f - \hat{f}\|^2$ are given by

kernel estimator using L-statistics (pointwise and global)	$6.056\ 10^{-3}$
kernel estimator by plug-in (pointwise and global)	$6.214\ 10^{-3}$
projection estimator using L-statistics	$0.433\ 10^{-3}$
projection estimator by plug-in	$0.431\ 10^{-3}$.

Consequently the plug-in projection estimator gives the best approximation.

6. Appendix

6.1. **Proof of Proposition 4.1.** Let x_0 be a fixed point. Denote by $\check{f}_h$ the pseudo-estimator of f given by $\check{f}_h(x) = \frac{1}{n}\sum_{i=1}^n w \circ G(Z_i) K_h(x - Z_i)$. We write

$$\hat{f}_h^{(2)}(x_0) - f(x_0) = \left(\hat{f}_h^{(2)} - \check{f}_h\right)(x_0) + \left(\check{f}_h - \mathbb{E}[\check{f}_h]\right)(x_0) + \left(\mathbb{E}[\check{f}_h] - f\right)(x_0) \,. \tag{30}$$

First, we state that by property (7) we have

$$\mathbb{E}[\check{f}_h(x_0)] = \mathbb{E}[K_h(x_0 - Y_i)] = K_h * f(x_0) =: f_h \,. \tag{31}$$

The last term in (30) is a standard bias term in kernel density estimation (Tsybakov, 2004)). Denote $b(x) = \left(\mathbb{E}[\check{f}_h] - f\right)(x)$. As $\int K(u)\mathrm{d}u = 1$,

$$b(x_0) = \int K_h(x_0 - y) f(y)\mathrm{d}y - f(x_0) = \int K(u)\left[f(x_0 - uh) - f(x_0)\right]\mathrm{d}u \,.$$

By a standard Taylor expansion of f, we get (see e.g. Tsybakov (2004))

$$|b(x_0)| \le \frac{C\int |x|^\beta |K(x)|\mathrm{d}x}{\ell!} h^\beta = C_1 h^\beta \,. \tag{32}$$

To study the second term of (30), we successively apply property (7), $0 \le w \le 1/a$ and the fact that K is square-integrable to obtain

$$\begin{aligned}
\mathbb{E}\left[\left(\check{f}_h - \mathbb{E}(\check{f}_h)\right)^2(x_0)\right] &= \frac{1}{n}\mathrm{Var}\left(w \circ G(Z_1) K_h(x_0 - Z_1)\right) \le \frac{1}{n}\mathbb{E}\left[\left\{w \circ G(Z_1) K_h\left(x_0 - Z_1\right)\right\}^2\right] \\
&\le \frac{1}{an}\mathbb{E}\left[K_h\left(x_0 - Y_1\right)^2\right] = \frac{1}{anh^2}\int K^2\left(\frac{x_0 - y}{h}\right) f(y)\mathrm{d}y \\
&\le \frac{1}{anh}\|f\|_\infty \|K\|_2^2 \,. \qquad (33)
\end{aligned}$$

For the first term in decomposition (30), the Lipschitz property of w implies that

$$\mathbb{E}\left[\left(\hat{f}_h^{(2)}-\check{f}_h\right)^2(x_0)\right] \le \frac{c_w^2}{n}\sum_{i=1}^n \mathbb{E}\left[(\hat{G}_n-G)^2(Z_i)K_h^2(x_0-Z_i)\right]$$
$$= \frac{c_w^2}{n}\left(\sum_{i=1}^n \mathbb{E}\left[\left(\hat{G}_{n,i}-\frac{n-1}{n}G\right)^2(Z_i)K_h^2(x_0-Z_i)\right]+\mathbb{E}\left[\left(\frac{1}{n}(1-G(Z_i))\right)^2 K_h^2(x_0-Z_i)\right]\right),$$

since the cross product term is centered, where $\hat{G}_{n,i}(x)=n^{-1}\sum_{j=1,j\neq i}^n \mathbb{1}_{Z_j\le x}$. Then

$$\mathbb{E}\left[\left(\hat{G}_{n,i}-\frac{n-1}{n}G\right)^2(Z_i)K_h^2(x_0-Z_i)\right] = \mathbb{E}\left[\mathbb{E}\left[\left(\hat{G}_{n,i}-\frac{n-1}{n}G\right)^2(Z_i)K_h^2(x_0-Z_i)\middle| Z_i\right]\right]$$
$$= \mathbb{E}\left[\frac{n-1}{n^2}G(Z_i)(1-G(Z_i))K_h^2(x_0-Z_i)\right] \le \frac{1}{4n}\mathbb{E}\left[K_h^2(x_0-Z_i)\right] \le \frac{\|g\|_\infty\|K\|_2^2}{4nh}.$$

Therefore, as $\|g\|_\infty \le d\|f\|_\infty$ and $c_w \le b/a^3$, we obtain that

$$\mathbb{E}\left[\left(\hat{f}_h^{(2)}-\check{f}_h\right)^2(x_0)\right] \le \frac{5db^2}{4nha^6}\|f\|_\infty\|K\|_2^2. \tag{34}$$

Gathering (32), (33) and (34) yields the result. □

6.2. **Proof of Proposition 4.2.** By (30) it follows that

$$\mathbb{E}\left[\|\hat{f}_h^{(2)}-f\|_2^2\right] \le 3\left(\mathbb{E}\left[\|\hat{f}_h^{(2)}-\check{f}_h\|_2^2\right]+\mathbb{E}\left[\|\check{f}_h-\mathbb{E}[\check{f}_h]\|_2^2\right]+\|\mathbb{E}[\check{f}_h]-f\|_2^2\right). \tag{35}$$

Concerning the last term of (35), proceeding as in Tsybakov (2004), we get

$$\|\mathbb{E}[\check{f}_h]-f\|_2^2 \le \left[\frac{Lh^\beta}{(\ell-1)!}\int |K(u)||u|^\beta \mathrm{d}u\right]^2. \tag{36}$$

For the first right-hand-side term of (35) we obtain

$$\begin{aligned}
\mathbb{E}\left[\|\hat{f}_h^{(2)}-\check{f}_h\|_2^2\right] &= \int \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n\left(w\circ\hat{G}_n(Z_i)-w\circ G(Z_i)\right)K_h(x-Z_i)\right)^2\right]\mathrm{d}x\\
&\le \int \mathbb{E}\left[\left(w\circ\hat{G}_n(Z_1)-w\circ G(Z_1)\right)^2 K_h^2(x-Z_1)\right]\mathrm{d}x\\
&\le \frac{c_w^2}{h}\|K\|_2^2\,\mathbb{E}\left[\left(\hat{G}_n(Z_1)-G(Z_1)\right)^2\right]\\
&\le \frac{b^2}{nha^6}\|K\|_2^2, \qquad (37)
\end{aligned}$$

where we used that $\mathbb{E}[(\hat{G}_n(Z_1)-G(Z_1))^2]\le 1/n$. This property can be shown by proceeding as in the pointwise case and using that $G(Z_1)$ has uniform distribution.

For the second term of (35) we obtain

$$\begin{aligned}
\mathbb{E}\left[\|\check{f}_h-\mathbb{E}[\check{f}_h]\|_2^2\right] &= \frac{1}{n}\int \mathbb{E}\left[w\circ G(Y_1)\left(K_h\left(x-Y_1\right)\right)^2\right]\mathrm{d}x\\
&\le \frac{1}{an}\int\int \left(K_h(x-y)\right)^2\mathrm{d}x f(y)\mathrm{d}y\\
&= \frac{1}{anh}\|K\|_2^2. \qquad (38)
\end{aligned}$$

Combining (36), (37) and (38) completes the proof. □

6.3. **Proof of Theorem 4.1.** For the sake of readability, super-indices $^{(2)}$ are omitted in the whole proof. For any $h \in \mathcal{H}$,

$$\begin{aligned}\left(\hat{f}_{\hat{h}(x_0)} - f\right)^2(x_0) &\le 3\left\{\left(\hat{f}_{\hat{h}(x_0)} - \hat{f}_{h,\hat{h}(x_0)}\right)^2(x_0) + \left(\hat{f}_{h,\hat{h}(x_0)} - \hat{f}_h\right)^2(x_0) + \left(\hat{f}_h - f\right)^2(x_0)\right\} \\ &\le 3\left\{\left(A_0(h,x_0) + V_0(\hat{h}(x_0))\right) + \left(A_0(\hat{h}(x_0),x_0) + V_0(h)\right) + \left(\hat{f}_h - f\right)^2(x_0)\right\} \\ &\le 6A_0(h,x_0) + 6V_0(h) + 3\left(\hat{f}_h(x_0) - f(x_0)\right)^2 , \qquad (39)\end{aligned}$$

where the second inequality holds by the definition of A_0, i.e. for all $h, h' \in \mathcal{H}$ we have $A_0(h,x_0) + V_0(h') \ge \left(\hat{f}_{h,h'}(x_0) - \hat{f}_{h'}(x_0)\right)^2$. The last inequality holds by the definition of $\hat{h}(x_0)$, that is $A_0(\hat{h}(x_0),x_0) + V_0(\hat{h}(x_0)) \le A_0(h,x_0) + V_0(h)$ for all $h \in \mathcal{H}$. The term $\mathbb{E}[(\hat{f}_h(x_0) - f(x_0))^2]$ is controlled by Proposition 4.1. Hence, it is sufficient to study the term $\mathbb{E}[A_0(h,x_0)]$. We state that

(40)
$$A_0(h,x_0) = \sup_{h' \in \mathcal{H}} \left[\left(\hat{f}_{h,h'}(x_0) - \hat{f}_{h'}(x_0)\right)^2 - V_0(h')\right]_+ \le 5\left(D_1 + D_2 + D_3 + D_4 + D_5\right) ,$$

where

$$D_1 = \sup_{h' \in \mathcal{H}} \left(\hat{f}_{h,h'}(x_0) - \check{f}_{h,h'}(x_0)\right)^2 , \quad D_2 = \sup_{h' \in \mathcal{H}} \left[\left(\check{f}_{h,h'}(x_0) - \mathbb{E}\left[\check{f}_{h,h'}(x_0)\right]\right)^2 - \frac{V_0(h')}{10}\right]_+ ,$$

$$D_3 = \sup_{h' \in \mathcal{H}} \left(\mathbb{E}\left[\check{f}_{h,h'}(x_0)\right] - \mathbb{E}\left[\check{f}_{h'}(x_0)\right]\right)^2 , \quad D_4 = \sup_{h' \in \mathcal{H}} \left[\left(\mathbb{E}\left[\check{f}_{h'}(x_0)\right] - \check{f}_{h'}(x_0)\right)^2 - \frac{V_0(h')}{10}\right]_+ ,$$

and $D_5 = \sup_{h' \in \mathcal{H}} \left(\check{f}_{h'}(x_0) - \hat{f}_{h'}(x_0)\right)^2$, with $\check{f}_{h,h'} = K_{h'} * \check{f}_h$.

We start with term D_3. Recall that $\mathbb{E}[\check{f}_h(x_0)] = K_h * f(x_0)$ by (31). Likewise, by property (7), $\mathbb{E}\left[\check{f}_{h,h'}(x_0)\right] = K_{h'} * K_h * f(x_0)$. In general we have $\|s * r\|_\infty \le \|s\|_\infty \|r\|_1$ and $\|K_h\|_1 = \|K\|_1$, yielding

$$\left|\mathbb{E}\left[\check{f}_{h,h'}(x_0)\right] - \mathbb{E}\left[\check{f}_{h'}(x_0)\right]\right| = \left|K_{h'} * (K_h * f - f)(x_0)\right| \le \|K_h * f - f\|_\infty \|K\|_1 .$$

Now remark that $(K_h * f - f)(x) = b(x)$ is the pointwise bias term considered in the proof of Proposition 4.1 Hence, (32) yields

$$D_3 \le C_1^2 \|K\|_1^2 h^{2\beta} . \qquad (41)$$

Concerning term D_4 we note that

$$\begin{aligned}\mathbb{E}[D_4] &\le \sum_{h \in \mathcal{H}} \mathbb{E}\left[\left(\left\{\check{f}_h(x_0) - \mathbb{E}\left[\check{f}_h(x_0)\right]\right\}^2 - \frac{V_0(h)}{10}\right)_+\right] \\ &= \sum_{h \in \mathcal{H}} \int_0^\infty \mathbb{P}\left(\left[\left\{\check{f}_h(x_0) - \mathbb{E}\left[\check{f}_h(x_0)\right]\right\}^2 - \frac{V_0(h)}{10}\right]_+ > x\right) \mathrm{d}x \\ &= \sum_{h \in \mathcal{H}} \int_0^\infty \mathbb{P}\left(\left|\check{f}_h(x_0) - \mathbb{E}\left[\check{f}_h(x_0)\right]\right| > \sqrt{\frac{V_0(h)}{10} + x}\right) \mathrm{d}x .\end{aligned}$$

The probability in the last term can be bounded by the Bernstein inequality. To this end we introduce the random variables $S_i = w \circ G(Z_i) K_h(x_0 - Z_i)$. Obviously, $|S_i| \leq \|K\|_\infty/(ah) =: M$ almost surely and by property (7)

$$\mathrm{Var}\,(S_i) \leq \mathbb{E}\left[w^2 \circ G(Z_1) K_h^2(x_0 - Z_1)\right] = \mathbb{E}\left[w \circ G(Y_1) K_h^2(x_0 - Y_1)\right] \leq \frac{\|K\|_2^2 \|f\|_\infty}{ah} =: v\ .$$

Hence, the Bernstein inequality implies for any $x > 0$

$$\begin{aligned}\mathbb{P}\left(\left|\check{f}_h(x_0) - \mathbb{E}\left[\check{f}_h(x_0)\right]\right| \geq \sqrt{\frac{V_0(h)}{10} + x}\right) &= \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n (S_i - \mathbb{E}[S_i])\right| \geq \sqrt{\frac{V_0(h)}{10} + x}\right) \\ &\leq 2\max\left\{\exp\left(-\frac{n}{4v}\left(\frac{V_0(h)}{10} + x\right)\right), \exp\left(-\frac{n\alpha}{4M}\sqrt{\frac{V_0(h)}{10}}\right)\exp\left(-\frac{n(1-\alpha)}{4M}\sqrt{x}\right)\right\}\ ,\end{aligned}$$

for any $\alpha \in [0,1]$ as $\sqrt{\cdot}$ is a concave function. By the definition of $V_0(h)$

$$\frac{n}{4v}\frac{V_0(h)}{10} = \frac{\kappa_0 \|K\|_1^2 \log n}{40} \geq p \log n\ ,$$

for $\kappa_0 \geq 40p$, since $\|K\|_1^2 \geq 1$. Furthermore,

$$\frac{n\alpha}{4M}\sqrt{\frac{V_0(h)}{10}} = \frac{\alpha \|K\|_2 \|K\|_1}{4\|K\|_\infty}\sqrt{\frac{\kappa_0 a \|f\|_\infty h n \log n}{10}} = \rho_\alpha \sqrt{hn \log n} \geq p \log n\ ,$$

for $nh \geq (p^2/\rho_\alpha^2)\log n$. We can choose $\alpha \in [0,1]$ sufficiently close to 0 such that $p/\rho_\alpha > 1$. Then the inequality in the last display holds under (H3), yielding

$$\begin{aligned}\mathbb{E}[D_4] &\leq \sum_{h \in \mathcal{H}} \int_0^\infty 2n^{-p} \max\left\{\exp\left(-\frac{nhax}{4\|K\|_2^2\|f\|_\infty}\right), \exp\left(-\frac{(1-\alpha)nha\sqrt{x}}{4\|K\|_\infty}\right)\right\} \mathrm{d}x \\ &\leq 2n^{-p} \sum_{h \in \mathcal{H}} \int_0^\infty \max\left\{\mathrm{e}^{-\tau_1 nhx}, \mathrm{e}^{-\tau_2 nh\sqrt{x}}\right\} \mathrm{d}x \leq 2n^{-p} \sum_{h \in \mathcal{H}} \max\left\{\frac{1}{\tau_1}, \frac{2}{\tau_2^2}\right\} \leq C' n^{-p+1}\ ,\end{aligned}$$

as $h \geq 1/n$ and the cardinality of $\mathcal{H}$ verifies $\#\mathcal{H} \leq n$. Finally, we choose $p = 2$ to get

$$\mathbb{E}[D_4] \leq \frac{C'}{n}\ . \tag{42}$$

Term D_2 can be treated in exactly the same way as D_4. More precisely, instead of S_i use $T_i = w \circ G(Z_i) K_h * K_{h'}(Z_i - x_0)$ verifying

$$\check{f}_{h,h'}(x_0) - \mathbb{E}\left[\check{f}_{h,h'}(x_0)\right] = \frac{1}{n}\sum_{i=1}^n T_i - \mathbb{E}[T_i]\ ,$$

and $|T_i| \leq \|K\|_\infty \|K\|_1/(ah') =: \bar{M}$ and $\mathrm{Var}(T_1) \leq \|f\|_\infty \|K\|_1^2 \|K\|_2^2/(ah') =: \bar{v}$. Hence, the Bernstein inequality yields

$$\mathbb{E}[D_2] \leq \frac{C''}{n}\ . \tag{43}$$

To study the terms D_5 and D_1 we first prove the following property.

Lemma 6.1. *Under the assumptions (H2) and (5), for any set Ω and for all $t \in \mathbb{R}$,*

$$\mathbb{E}\left[\left(\hat{f}_h(t) - \check{f}_h(t)\right)^2 \mathbb{1}_{\Omega^c}\right] \leq c_w^2 \|K\|_\infty^2 n^2 \mathbb{P}(\Omega^c) \quad \text{and}$$
$$\mathbb{E}\left[\left(\hat{f}_{h',h}(t) - \check{f}_{h',h}(t)\right)^2 \mathbb{1}_{\Omega^c}\right] \leq c_w^2 \|K\|_\infty^2 \|K\|_1^2 n^2 \mathbb{P}(\Omega^c) \ .$$

Proof. By using $\|\hat{G}_n - G\|_\infty \leq 1$, we have

$$\begin{aligned}
\mathbb{E}\left[\left(\hat{f}_h(t) - \check{f}_h(t)\right)^2 \mathbb{1}_{\Omega^c}\right] &= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n (w \circ \hat{G}_n(Z_i) - w \circ G(Z_i)) K_h(t - Z_i)\right)^2 \mathbb{1}_{\Omega^c}\right] \\
&\leq \frac{c_w^2}{n^2}\mathbb{E}\left[\left(\sum_{i=1}^n |K_h(t - Z_i)|\right)^2 \mathbb{1}_{\Omega^c}\right] \leq c_w^2 \mathbb{E}\left[K_h^2(t - Z_1)\mathbb{1}_{\Omega^c}\right] \\
&\leq c_w^2 \|K_h\|_\infty^2 \mathbb{E}\left[\mathbb{1}_{\Omega^c}\right] = \frac{c_w^2}{h^2}\|K\|_\infty^2 \mathbb{P}(\Omega^c) \leq c_w^2 \|K\|_\infty^2 n^2 \mathbb{P}(\Omega^c) \ ,
\end{aligned}$$

as $1/h \leq n$. In the same way, we show the second statement of the Lemma, by using $\|K_{h'} * K_h\|_\infty \leq \|K_{h'}\|_\infty \|K_h\|_1 \leq n\|K\|_\infty \|K\|_1$. □

Now let $\Omega = \{\omega : \|\hat{G}_n - G\|_\infty \leq s\}$ for some constant $s > 0$. Then (see Massart (1990)),

$$\mathbb{P}(\Omega^c) = \mathbb{P}(\|\hat{G}_n - G\|_\infty > s) \leq \mathrm{e}^{-2ns^2} \ , \tag{44}$$

by the Dvoretzky-Kiefer-Wolfowitz inequality. This implies that

$$\begin{aligned}
\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\hat{f}_h(x_0) - \check{f}_h(x_0)\right)^2 \mathbb{1}_{\Omega^c}\right] &\leq \sum_{h\in\mathcal{H}} \mathbb{E}\left[\left(\hat{f}_h(x_0) - \check{f}_h(x_0)\right)^2 \mathbb{1}_{\Omega^c}\right] \leq \sum_{h\in\mathcal{H}} c_w^2 \|K\|_\infty^2 n^2 \mathrm{e}^{-2ns^2} \\
&= c_w^2 \|K\|_\infty^2 n^3 \mathrm{e}^{-2ns^2} < \infty \ ,
\end{aligned}$$

as $\#\mathcal{H} \leq n$. Furthermore,

$$\begin{aligned}
\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\hat{f}_h(x_0) - \check{f}_h(x_0)\right)^2 \mathbb{1}_{\Omega}\right] &\leq c_w^2 \mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^n |\hat{G}_n(Z_i) - G(Z_i)||K_h(x_0 - Z_i)|\right)^2 \mathbb{1}_{\Omega}\right] \\
&\leq s^2 c_w^2 \mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^n |K_h(x_0 - Z_i)|\right)^2\right] \\
&\leq 2s^2 c_w^2 \left\{\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^n (|K_h(x_0 - Z_i)| - \mathbb{E}[|K_h(x_0 - Z_i)|])\right)^2\right] + \sup_{h\in\mathcal{H}}\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n |K_h(x_0 - Z_i)|\right]\right)^2\right\} \\
&\leq 2s^2 c_w^2 \left\{\frac{1}{n}\sum_{h\in\mathcal{H}} \mathrm{Var}\left(|K_h(x_0 - Z_1)|\right) + \sup_{h\in\mathcal{H}}\left[\mathbb{E}(|K_h(x_0 - Z_1)|)\right]^2\right\} \ .
\end{aligned}$$

On the one hand,

$$\frac{1}{n}\sum_{h\in\mathcal{H}} \mathrm{Var}\left(|K_h(x_0 - Z_1)|\right) \leq \frac{1}{n}\sum_{h\in\mathcal{H}} \mathbb{E}\left[K_h^2(x_0 - Z_1)\right] = \frac{1}{n}\sum_{h\in\mathcal{H}} \frac{1}{h}\|K\|_2^2 \|g\|_\infty \leq S\|K\|_2^2 d\|f\|_\infty \ ,$$

where S is defined in (H3) and d in (6). On the other hand,

$$\sup_{h\in\mathcal{H}}\left[\mathbb{E}(|K_h(x_0-Z_1)|)\right]^2=\sup_{h\in\mathcal{H}}\left(\int|K(z)|g(x_0-zh)\mathrm{d}z\right)^2\leq d^2\|f\|_\infty^2\|K\|_1^2\,.$$

It follows that $\mathbb{E}[D_5]\leq\mu_1 n^3\mathrm{e}^{-2ns^2}+\mu_2 s^2$, with constants $\mu_1=c_w^2\|K\|_\infty^2$ and $\mu_2=2c_w^2 d\|f\|_\infty(S\|K\|_2^2+d\|f\|_\infty\|K\|_1^2)$. Choosing $s^2=2\log n/n$ gives

$$\mathbb{E}[D_5]\leq\frac{\mu_1}{n}+2\mu_2\frac{\log n}{n}\,. \tag{45}$$

Finally, the study of D_1 follows the same line as the study of D_5. That is, on the one hand, we have for the same set Ω

$$\mathbb{E}\left[D_1\mathbb{1}_{\Omega^c}\right]\leq c_w^2\|K\|_\infty^2\|K\|_1^2 n^3\mathrm{e}^{-2ns^2}\,.$$

On the other hand,

$$\mathbb{E}\left[D_1\mathbb{1}_{\Omega}\right]\leq 2s^2c_w^2\left\{\frac{1}{n}\sum_{h\in\mathcal{H}}\mathbb{E}\left[(K_{h'}*K_h(x_0-Z_1))^2\right]+\sup_{h\in\mathcal{H}}\left(\mathbb{E}\left[|K_{h'}*K_h(x_0-Z_1)|\right]\right)^2\right\}\,.$$

By the generalized Minkowski inequality, we obtain

$$\begin{aligned}\mathbb{E}\left[(K_{h'}*K_h(x_0-Z_1))^2\right]&\leq\left[\int|K_{h'}(u)|\left(\int K_h^2(x_0-z-u)g(z)\mathrm{d}z\right)^{1/2}\mathrm{d}u\right]^2\\&\leq\|g\|_\infty\|K_h\|_2^2\|K_{h'}\|_1^2\leq d\|f\|_\infty\|K\|_2^2\|K\|_1^2/h\,.\end{aligned}$$

Furthermore,

$$\begin{aligned}\sup_{h\in\mathcal{H}}\left(\mathbb{E}\left[|K_{h'}*K_h(x_0-Z_1)|\right]\right)^2&\leq\sup_{h\in\mathcal{H}}\left(\int\left|\int K_{h'}(u)\right|\int|K(v)|\,g(x_0-vh-u)\mathrm{d}v\mathrm{d}u\right)^2\\&\leq(d\|f\|_\infty\|K\|_1^2)^2\,.\end{aligned}$$

It follows with $\tilde{\mu}_1=\mu_1\|K\|_1^2$ and $\tilde{\mu}_2=\mu_2\|K\|_1^2$ that $\mathbb{E}[D_1]\leq\tilde{\mu}_1 n^3\mathrm{e}^{-2ns^2}+\tilde{\mu}_2 s^2$. Hence,

$$\mathbb{E}[D_1]\leq\frac{\tilde{\mu}_1}{n}+2\tilde{\mu}_2\frac{\log n}{n}\,, \tag{46}$$

with $s^2=2\log n/n$. Now, if we plug (41), (42), (43), (45) and (46) into (40), we get

$$\mathbb{E}[A_0(h,x_0)]\leq\tilde{C}_1h^{2\beta}+\tilde{C}_2\frac{\log n}{n},$$

which, associated with Proposition 4.1, can be inserted in (39) to end the proof of Theorem 4.1. □

6.4. **Proof of Theorem 4.2.** In all the proof below, super-indices $^{(2)}$ are omitted. Similar to the pointwise case, we have for any $h\in\mathcal{H}$

$$\|\hat{f}_{\hat{h}}-f\|_2^2\leq 6A(h)+6V(h)+3\|\hat{f}_h-f\|_2^2\,. \tag{47}$$

By the proof of Proposition 4.2,

$$\mathbb{E}\left[\|\hat{f}_h-f\|_2^2\right]\leq 3\|f_h-f\|_2^2+\frac{C_4}{nh}\,. \tag{48}$$

Hence, only term $\mathbb{E}[A(h)]$ needs to be studied. By analogy to the proof of Theorem 4.1,

$$A(h)\leq 5(F_1+F_2+F_3+F_4+F_5)\,, \tag{49}$$

where

$$F_1 = \sup_{h' \in \mathcal{H}} \|\hat{f}_{h,h'} - \check{f}_{h,h'}\|_2^2 \,, \quad F_2 = \sup_{h' \in \mathcal{H}} \left(\|\check{f}_{h,h'} - \mathbb{E}[\check{f}_{h,h'}]\|_2^2 - \frac{V(h')}{10} \right)_+ \,,$$
$$F_3 = \sup_{h' \in \mathcal{H}} \|\mathbb{E}[\check{f}_{h,h'}] - \mathbb{E}[\check{f}_{h'}]\|_2^2 \,, \quad F_4 = \sup_{h' \in \mathcal{H}} \left(\|\mathbb{E}[\check{f}_{h'}] - \check{f}_{h'}\|_2^2 - \frac{V(h')}{10} \right)_+ \,,$$
$$F_5 = \sup_{h' \in \mathcal{H}} \|\check{f}_{h'} - \hat{f}_{h'}\|_2^2 \,.$$

First, we study term F_3. The inequality $\|u * v\|_2 \leq \|u\|_1 \|v\|_2$ yields

$$F_3 = \sup_{h' \in \mathcal{H}} \|K_{h'} * K_h * f - K_{h'} * f\|_2^2 \leq \sup_{h' \in \mathcal{H}} \|K_{h'}\|_1^2 \|K_h * f - f\|_2^2 = \|K\|_1^2 \|f - f_h\|_2^2 \,. \tag{50}$$

To study term F_4 we introduce the centered empirical process $\nu_{n,h}$ defined by

$$\begin{aligned} \nu_{n,h}(\psi) &= \langle \check{f}_h - \mathbb{E}[\check{f}_h], \psi \rangle \\ &= \frac{1}{n} \sum_{i=1}^n \int \left(w \circ G(Z_i) K_h(u - Z_i) - \mathbb{E}[w \circ G(Z_i) K_h(u - Z_i)] \right) \psi(u) \mathrm{d}u \,. \end{aligned}$$

As $\psi \mapsto \nu_{n,h}(\psi)$ is continuous, the supremum can be taken over a countable dense subset of $\{\psi \in \mathbb{L}_2, \|\psi\| = 1\}$, which we denote by $\mathcal{B}(1)$. Then, $\|\check{f}_h - \mathbb{E}[\check{f}_h]\|_2^2 = \sup_{\psi \in \mathcal{B}(1)} \langle \check{f}_h - \mathbb{E}[\check{f}_h], \psi \rangle^2 = \sup_{\psi \in \mathcal{B}(1)} \nu_{n,h}(\psi)$. Therefore we obtain

$$\mathbb{E}[F_4] \leq \sum_{h \in \mathcal{H}} \mathbb{E}\left[\left(\|\check{f}_h - \mathbb{E}[\check{f}_h]\|_2^2 - \frac{V(h)}{10} \right)_+ \right] = \sum_{h \in \mathcal{H}} \mathbb{E}\left[\left(\sup_{\psi \in \mathcal{B}(1)} \nu_{n,h}^2(\psi) - \frac{V(h)}{10} \right)_+ \right] .$$

The expectation in the last term can be bounded by Talagrand's inequality (see Subsection 6.7). More precisely, to apply this result, we have to determine the values of the constants H, M and v. Denote $f_\psi(z) = w \circ G(z) K_h * \psi(z)$, so that $\nu_{n,h}(\psi) = \frac{1}{n} \sum_{i=1}^n (f_\psi(Z_i) - \mathbb{E}[f_\psi(Z_i)])$. First, for any $\psi \in \mathcal{B}(1)$ the Cauchy-Schwarz inequality gives

$$\|f_\psi\|_\infty \leq \frac{1}{a} \|K_h * \psi\|_\infty = \frac{1}{a} \sup_z |\langle |K_h(\cdot - z)|, |\psi| \rangle| \leq \frac{\|K_h\|_2 \|\psi\|_2}{a} \leq \frac{\|K\|_2}{a\sqrt{h}} =: M \,.$$

Next, we see that

$$\left(\mathbb{E}\left[\sup_{\psi \in \mathcal{B}(1)} |\nu_{n,h}(\psi)| \right] \right)^2 \leq \mathbb{E}\left[\sup_{\psi \in \mathcal{B}(1)} \nu_{n,h}^2(\psi) \right] \leq \mathbb{E}\left[\|\check{f}_h - \mathbb{E}[\check{f}_h]\|_2^2 \right] \leq \frac{V(h)}{\kappa_1} =: H^2 \,,$$

by (38). Furthermore, let $\varepsilon^2 = 1/2$. To obtain $4H^2 = V(h)/10$, we set $H = \sqrt{V(h)/40}$.

Lastly, for any $\psi \in \mathcal{B}(1)$ we show that by (7)

$$\begin{aligned} \mathrm{Var}\left(f_\psi(Z)\right) &\leq \mathbb{E}\left[(w \circ G(Z) K_h * \psi(Z))^2 \right] \leq \frac{1}{a} \int (K_h * \psi(y))^2 f(y) \mathrm{d}y \\ &\leq \frac{1}{a} \|f\|_\infty \|K_h * \psi\|_2^2 \leq \frac{1}{a} \|f\|_\infty \|K_h\|_1^2 \|\psi\|_2^2 \leq \frac{1}{a} \|f\|_\infty \|K\|_1^2 =: v \,. \end{aligned}$$

Finally, Talagrand's inequality yields

$$\mathbb{E}\left[\left(\sup_{\psi \in \mathcal{B}(1)} \nu_{n,h}^2(\psi) - \frac{V(h)}{10} \right)_+ \right] \leq \frac{\tilde{C}_1}{n} \left(\mathrm{e}^{-\tilde{C}_2/h} + \frac{1}{nh} \mathrm{e}^{-\tilde{C}_3 \sqrt{n}} \right) \leq \frac{\tilde{C}_1}{n} \left(\mathrm{e}^{-\tilde{C}_2/h} + \frac{\tilde{C}_4}{n} \right) ,$$

where $\tilde{C}_k > 0, k = 1, \dots, 4$ are constants depending on K, $\|f\|_\infty$ and a. Consequently,

$$\mathbb{E}[F_4] \leq \frac{\tilde{C}_1}{n} \sum_{h \in \mathcal{H}} \left(\mathrm{e}^{-\tilde{C}_2/h} + \frac{\tilde{C}_4}{n} \right) \leq \frac{\tilde{C}_5}{n} , \tag{51}$$

as $\#\mathcal{H} \leq n$ and $\sum_{h \in \mathcal{H}} \mathrm{e}^{-\tilde{C}_2/h} \leq \Sigma(C_2')$ under condition (26).

In the same way we obtain for F_2

$$\mathbb{E}[F_2] \leq \frac{\bar{C}}{n} . \tag{52}$$

Now let us turn to F_5. We note that

$$\|\hat{f}_h - \check{f}_h\|_2^2 \leq \frac{4}{a^2 n^2} \int \left(\sum_{i=1}^n |K_h(u - Z_i)| \right)^2 \mathrm{d}u \leq \frac{4}{a^2 n} \sum_{i=1}^n \|K_h(\cdot - Z_i)\|_2^2 = \frac{4}{a^2 h} \|K\|_2^2 \leq \frac{4n}{a^2} \|K\|_2^2 .$$

Therefore,

$$\mathbb{E}\left[\sup_{h \in \mathcal{H}} \|\hat{f}_h - \check{f}_h\|_2^2 \mathbb{1}_{\Omega^c} \right] \leq \frac{4n}{a^2} \|K\|_2^2 \mathbb{P}(\Omega^c) ,$$

where $\Omega = \{\omega : \|G - \hat{G}_n\|_\infty \leq s\}$ as previously, and we recall that $\mathbb{P}(\Omega^c) \leq \mathrm{e}^{-2ns^2}$.

Following the same line as for D_5 in the pointwise case and by choosing $s = \sqrt{\log n / n}$, we conclude that

$$\mathbb{E}[F_5] \leq \frac{C_1'}{n} + C_2' \frac{\log n}{n} . \tag{53}$$

For F_1, we follow the same line as for F_5 to obtain

$$\mathbb{E}[F_1] \leq \|K\|_1^2 \left(\frac{C_1'}{n} + C_2' \frac{\log n}{n} \right) . \tag{54}$$

Consequently, plugging (50), (51), (52), (53) and (54) into (49) gives a bound of $\mathbb{E}[A(h)]$. Combining this with (48), (47) and the definition of $V(h)$ yields Theorem 4.2. □

6.5. **Proof of Proposition 4.3.** Pythagoras formula yields $\|f - \hat{f}_m\|_2^2 = \|f - f_m\|_2^2 + \|f_m - \hat{f}_m\|_2^2$. By definition of the orthogonal projection $f_m = \sum_{j=0}^{2m} a_j \varphi_j$ and by using equality (7), we have $a_j = \langle \varphi_j, f \rangle = \mathbb{E}(\varphi_j(Y)) = \mathbb{E}(\varphi_j(Z_1) w \circ G(Z_1))$. This, together with formula (13) implies that $\|f_m - \hat{f}_m\|_2^2 = \sum_{j=0}^{2m} (a_j - \hat{a}_j)^2$. If we define

$$\nu_n(h) = \frac{1}{n} \sum_{i=1}^n [h(Z_i) w \circ G(Z_i) - \mathbb{E}(h(Z_i)\ w \circ G(Z_i))], \tag{55}$$

$$R_n(h) = \frac{1}{n} \sum_{i=1}^n h(Z_i) [w \circ \hat{G}_n(Z_i) - w \circ G(Z_i)] , \tag{56}$$

then we get $\|f_m - \hat{f}_m\|_2^2 \leq 2 \sum_{j=0}^{2m} (\nu_n(\varphi_j)^2 + R_n(\varphi_j)^2)$. We have, on the one hand,

$$\begin{aligned} \sum_{j=0}^{2m} \mathbb{E}(\nu_n^2(\varphi_j)) &= \sum_{j=0}^{2m} \frac{1}{n} \mathrm{Var}(\varphi_j(Z_i) w \circ G(Z_i)) \leq \sum_{j=0}^{2m} \frac{1}{n} \mathbb{E}\left[\varphi_j^2(Z_1) (w \circ G(Z_1))^2 \right] \\ &\leq \frac{1}{n} \mathbb{E}\left[\| \sum_{j=0}^{2m} \varphi_j^2 \|_\infty (w \circ G(Z_1))^2 \right] \leq \frac{D_m}{n} \mathbb{E}[(w \circ G(Z_1))^2] \leq \frac{1}{a^2} \frac{D_m}{n} , \end{aligned} \tag{57}$$

because the basis satisfies $\sum_{j=0}^{2m}\varphi_j^2 = 2m+1 = D_m$. On the other hand, we have

$$\begin{aligned}\sum_{j=0}^{2m}\mathbb{E}(R_n^2(\varphi_j)) &\le \sum_{j=0}^{2m}\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\varphi_j(Z_i)[w\circ\hat{G}_n(Z_i)-w\circ G(Z_i)]\right)^2\right]\\ &\le \frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{2m}\mathbb{E}\left(\varphi_j^2(Z_i)[w\circ\hat{G}_n(Z_i)-w\circ G(Z_i)]^2\right)\\ &\le c_w^2\sum_{j=0}^{2m}\mathbb{E}\left(\|G-\hat{G}_n\|_\infty^2\varphi_j^2(Z_i)\right) \le c_w^2 D_m\mathbb{E}\left(\|G-\hat{G}_n\|_\infty^2\right) \le c_w^2\frac{D_m}{n}\,, \qquad (58)\end{aligned}$$

with (5) and because of $\mathbb{E}\left(\|G-\hat{G}_n\|_\infty^2\right) \le 1/n$ (see e.g. Brunel and Comte, 2005, p. 462). By gathering all terms, we obtain the risk bound stated in Proposition 4.3. □

6.6. **Sketch of proof of Theorem 4.3.** In the following, we omit the super index $^{(2)}$. It is easy to see that $\hat{f}_m = \arg\min_{t\in S_m}\gamma_n(t)$ for $\gamma_n(t) = \|t\|^2 - 2n^{-1}\sum_{i=1}^n w\circ\hat{G}_n(Z_i)t(Z_i)$. Thus, we can write $\gamma_n(t)-\gamma_n(s) = \|t-f\|_2^2 - \|s-f\|_2^2 - 2\nu_n(t-s) - 2R_n(t-s)$, where ν_n and R_n are defined by (55) and (56). By definition of $\hat{f}_{\hat{m}}$ we have for all $m\in\mathcal{M}_n$, $\gamma_n(\hat{f}_{\hat{m}})+\mathrm{pen}(\hat{m}) \le \gamma_n(f_m)+\mathrm{pen}(m)$. This can be rewritten as $\|\hat{f}_{\hat{m}}-f\|_2^2 \le \|f_m-f\|_2^2 + \mathrm{pen}(m)+2\nu_n(\hat{f}_{\hat{m}}-f_m)-\mathrm{pen}(\hat{m})+2R_n(\hat{f}_{\hat{m}}-f_m)$. Using this and and that $2xy \le x^2/\theta+\theta y^2$ for all nonnegative x,y,θ, we obtain

$$\|f-\hat{f}_{\hat{m}}\|_2^2 \le \|f-f_m\|_2^2 + \mathrm{pen}(m) + 2\nu_n(\hat{f}_{\hat{m}}-f_m) - \mathrm{pen}(\hat{m}) + 2R_n(\hat{f}_{\hat{m}}-f_m)$$

$$\begin{aligned}\|f-\hat{f}_{\hat{m}}\|_2^2 &\le \|f-f_m\|_2^2 + \mathrm{pen}(m) + 2\|\hat{f}_{\hat{m}}-f_m\|_2\sup_{t\in S_{\hat{m}}+S_m,\|t\|_2=1}|\nu_n(t)| - \mathrm{pen}(\hat{m})\\ &\quad + 2\|\hat{f}_{\hat{m}}-f_m\|_2\sup_{t\in S_{\hat{m}}+S_m,\|t\|_2=1}|R_n(t)|\\ &\le \|f-f_m\|_2^2 + \mathrm{pen}(m) + \frac{1}{4}\|\hat{f}_{\hat{m}}-f_m\|_2^2 + 2\sup_{t\in S_{\hat{m}}+S_m,\|t\|_2=1}[\nu_n(t)]^2\\ &\quad - \mathrm{pen}(\hat{m}) + \frac{1}{8}\|\hat{f}_{\hat{m}}-f_m\|_2^2 + 8\sup_{t\in S_{\hat{m}}+S_m,\|t\|_2=1}[R_n(t)]^2\,.\end{aligned}$$

As $\|\hat{f}_{\hat{m}}-f_m\|_2^2 \le 2(\|\hat{f}_{\hat{m}}-f\|_2^2 + \|f_m-f\|_2^2)$, this yields

$$\begin{aligned}\frac{1}{4}\mathbb{E}[\|f-\hat{f}_{\hat{m}}\|_2^2] &\le \frac{7}{4}\|f-f_m\|_2^2 + 2\mathrm{pen}(m) + 8\mathbb{E}\left(\sup_{t\in S_{m_n},\|t\|_2=1}[R_n(t)]^2\right)\\ &\quad + 4\mathbb{E}\left(\sup_{t\in S_{\hat{m}}+S_m,\|t\|_2=1}[\nu_n(t)]^2 - (\mathrm{pen}(m)+\mathrm{pen}(\hat{m}))/4\right)_+\,.\end{aligned}$$

Then the term $\mathbb{E}\left(\sup_{t\in S_{\hat{m}}+S_m,\|t\|_2=1}[\nu_n(t)]^2 - (\mathrm{pen}(m)+\mathrm{pen}(\hat{m}))/4\right)_+$ is bounded by C/n by using Talagrand Inequality in a standard way (see e.g. Brunel et al., 2005). For the last term $\mathbb{E}\left(\sup_{t\in S_{m_n},\|t\|_2=1}[R_n(t)]^2\right)$, we define Ω_G by

$$\Omega_G = \{\sqrt{n}\|\hat{G}_n - G\|_\infty \le \sqrt{\log(n)}\}. \qquad (59)$$

As in (44), we use Massart (1990) and get

$$\mathbb{P}(\sqrt{n}\|\hat{G}_n - G\|_\infty \geq \lambda) \leq 2e^{-2\lambda^2}. \tag{60}$$

This implies that $\mathbb{P}(\Omega_G^c) \leq 2/n^2$. Then we write that $\mathbb{E}\left(\sup_{t\in S_{m_n},\|t\|_2=1}[R_n(t)]^2\right)$ is less than

$$\mathbb{E}\left(\sup_{t\in S_{m_n},\|t\|_2=1}[R_n(t)\mathbb{1}_{\Omega_G}]^2\right) + \mathbb{E}\left(\sup_{t\in S_{m_n},\|t\|_2=1}[R_n(t)\mathbb{1}_{\Omega_G^c}]^2\right) \qquad := \mathcal{R}_1 + \mathcal{R}_2.$$

For the first term, we have

$$\begin{aligned}
\mathcal{R}_1 &\leq c_w^2\mathbb{E}\left[\|\hat{G}_n - G\|_\infty^2\mathbb{1}_{\Omega_G}\mathbb{E}\left(\sup_{t\in S_{m_n},\|t\|_2=1}(\frac{1}{n}\sum_{i=1}^n|t(Z_i)|)^2\right)\right] \\
&\leq c_w^2\frac{\log(n)}{n}\mathbb{E}\left(\sup_{t\in S_{m_n},\|t\|_2=1}(\frac{1}{n}\sum_{i=1}^n t^2(Z_i))\right) \\
&\leq 2c_w^2\frac{\log(n)}{n}\left[\mathbb{E}\left(\sup_{t\in S_{m_n},\|t\|_2=1}|\nu_n'(t^2)|\right) + \sup_{t\in S_{m_n},\|t\|_2=1}\mathbb{E}(t^2(Z_1))\right]
\end{aligned}$$

where $\nu_n'(t) = \frac{1}{n}\sum_{i=1}^n(t(Z_i) - \mathbb{E}(t(Z_1))$. It is proved in Brunel and Comte (2005) that $\mathbb{E}\left(\sup_{t\in S_{m_n},\|t\|_2=1}|\nu_n'(t^2)|\right) \leq C\log(n)$ if the density of Z_1 is bounded and $N_n \leq O(\sqrt{n})$ for the trigonometric basis. Moreover $\mathbb{E}(t^2(Z_1)) \leq \|t\|_2^2\|f\|_\infty/w_0$. We obtain $\mathcal{R}_1 \leq C\log^2(n)/n$. On the other hand, we have

$$\mathcal{R}_2 \leq \sum_j \mathbb{E}(R_n^2(\varphi_j)\mathbb{1}_{\Omega^c}) \leq c_w^2 n\mathbb{E}^{1/2}(\|\hat{G}_n - G\|_\infty^4)\mathbb{P}^{1/2}(\Omega_G^c) \leq \frac{C}{n}.$$

This yields $\mathbb{E}\left(\sup_{t\in S_{m_n},\|t\|_2=1}[R_n(t)]^2\right) \leq C\log^2(n)/n$. Finally we obtain that, for all $m \in \mathcal{M}_n$, $\mathbb{E}[\|f - \hat{f}_{\hat{m}}\|_2^2] \leq 7\|f - f_m\|_2^2 + 8\mathrm{pen}(m) + K\log^2(n)/n$, which ends the proof. □

6.7. **The Talagrand inequality.** The following result follows from the Talagrand concentration inequality given in (Klein and Rio, 2005) and arguments in (Birgé and Massart, 1998) (see the proof of their Corollary 2 page 354).

Lemma 6.2. *(Talagrand's inequality) Let $Y_1, \ldots, Y_n$ be independent random variables, let $\nu_{n,Y}(f) = (1/n)\sum_{i=1}^n[f(Y_i) - \mathbb{E}(f(Y_i))]$ and let $\mathcal{F}$ be a countable class of uniformly bounded measurable functions. Then for $\epsilon^2 > 0$*

$$\mathbb{E}\Big[\sup_{f\in\mathcal{F}}|\nu_{n,Y}(f)|^2 - 2(1+2\epsilon^2)H^2\Big]_+ \leq \frac{4}{K_1}\left(\frac{v}{n}\mathrm{e}^{-K_1\epsilon^2\frac{nH^2}{v}} + \frac{98M^2}{K_1n^2C^2(\epsilon^2)}\mathrm{e}^{-\frac{2K_1C(\epsilon^2)\epsilon}{7\sqrt{2}}\frac{nH}{M}}\right),$$

with $C(\epsilon^2) = \sqrt{1+\epsilon^2} - 1$, $K_1 = 1/6$, and

$$\sup_{f\in\mathcal{F}}\|f\|_\infty \leq M, \quad \mathbb{E}\Big[\sup_{f\in\mathcal{F}}|\nu_{n,Y}(f)|\Big] \leq H, \quad \sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{k=1}^n\mathrm{Var}(f(Y_k)) \leq v.$$

By standard denseness arguments, this result can be extended to the case where $\mathcal{F}$ is a unit ball of a linear normed space, after checking that $f \mapsto \nu_n(f)$ is continuous and $\mathcal{F}$ contains a countable dense family.

References

Asgharian, M., M'Lan, C. E., and Wolfson, D. B. (2002). Length-biased sampling with right censoring: an unconditional approach. *J. Amer. Statist. Assoc.*, 97(457):201–209.

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

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

Brunel, E. and Comte, F. (2005). Penalized contrast estimation of density and hazard rate with censored data. *Sankhya*, 67:441–475.

Brunel, E., Comte, F., and Guilloux, A. (2005). Nonparametric density estimation in presence of bias and censoring. *Test*, 18:166–194.

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

Chesneau, C. (2010). Wavelet block thresholding for density estimation in the presence of bias. *J. Korean Statist. Soc.*, 39(1):43–53.

Cutillo, L., De Feis, I., Nikolaidou, C., and Sapatinas, T. (2014). Wavelet density estimation for weighted data. *J. Statist. Plann. Inference*, 146:1–19.

de Uña-Álvarez, J. (2004). Nonparametric estimation under length-biased sampling and type I censoring: a moment based approach. *Ann. Inst. Statist. Math.*, 56(4):667–681.

de Uña-Álvarez, J. and Rodríguez-Casal, A. (2006). Comparing nonparametric estimators for length-biased data. *Comm. Statist. Theory Methods*, 35(4-6):905–919.

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

Efromovich, S. (2004a). Density estimation for biased data. *Ann. Statist.*, 32(3):1137–1161.

Efromovich, S. (2004b). Distribution estimation for biased data. *J. Statist. Plann. Inference*, 124(1):1–43.

El Barmi, H. and Simonoff, J. S. (2000). Transformation-based density estimation for weighted distributions. *J. Nonparametr. Statist.*, 12(6):861–878.

Gill, R. D., Vardi, Y., and Wellner, J. A. (1988). Large sample theory of empirical distributions in biased sampling models. *Ann. Statist.*, 16(3):1069–1112.

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

Jones, M. C. (1991). Kernel density estimation for length biased data. *Biometrika*, 78(3):511–519.

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

Li, X. and Zuo, M. J. (2004). Preservation of stochastic orders for random minima and maxima, with applications. *Naval Research Logistics*, 51(3):332–344.

Massart, P. (1990). The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *Annals of Probability*, 18:1269–1283.

O'Connor, D. V. and Phillips, D. (1984). *Time-correlated single photon counting*. Academic Press, London.

Rebafka, T., Roueff, F., and Souloumiac, A. (2010). A corrected likelihood approach for the pile-up model with application to fluorescence lifetime measurements using exponential mixtures. *The International Journal of Biostatistics*, 6(1).

Shaked, M. and Wong, T. (1997). Stochastic comparisons of random minima and maxima. *Journal of Applied Probability*, 34(2):420–425.

Tsodikov, A. (2001). Estimation of survival based on proportional hazards when cure is a possibility. *Mathematical and Computer Modelling*, 33(12–13):1227–1236.

Tsybakov, A. B. (2004). *Introduction à l'estimation non-paramétrique*, volume 41 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Berlin.

Vardi, Y. (1982). Nonparametric estimation in the presence of length bias. *Ann. Statist.*, 10(2):616–620.

Wu, C. O. (1997). A cross-validation bandwidth choice for kernel density estimates with selection biased data. *J. Multivariate Anal.*, 61(1):38–60.

Wu, C. O. and Mao, A. Q. (1996). Minimax kernels for density estimation with biased data. *Ann. Inst. Statist. Math.*, 48(3):451–467.