NONPARAMETRIC WEIGHTED ESTIMATORS FOR BIASED DATA

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

Abstract. Starting from a real data example in fluorescence, the problem of nonparametric estimation of a density in a biased data model is considered. Bias correction can be done in two ways: either an estimator is computed with the data and in a second time a correction (plug in estimator) is applied, or weights are directly associated with the data so that a direct estimator of the quantity of interest (weighted estimator) is obtained. In both cases, kernel and projection estimation strategies with bandwidth or model selection devices are developed. The bandwidth selection is inspired from a procedure recently proposed by Goldenshluger and Lespki (2011). Risk bounds are proved showing that the final data-driven estimators perform an automatic finite sample bias-variance tradeoff. A simulation study compares the two bias-correction methods and the different model or bandwidth selection methods. Finally real fluorescence data are studied.

Keywords. Adaptive density estimation. Bandwidth selection. Biased data. Fluorescence lifetimes. Model selection. Weighted estimators.

January 22, 2016

1. Introduction

In various application settings, functional estimation can be difficult because the observations 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 cumulative distribution function (cdf) of the observations, denoted by G, is the result of a (known) nonlinear distortion of the distribution of interest, say F. More precisely, the cdf G and F are related by some known link function H by

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

This paper is concerned with the estimation of the probability density function (pdf) f associated with F from a sample $Z_1, \ldots, Z_n$ with distribution G.

A special case of model (1) is the pile-up model, where a random variable Z with distribution G is defined as the minimum of a random number N of independent and identically distributed (i.i.d.) random variables $Y_1, \ldots, Y_N$ with distribution F. This model is our leading example. It is encountered for example in biostatistics when considering the time until the outbreak of a tumor originated from a clonogenic cell in the presence of a random number of competing clonogens (Tsodikov, 2001). Another example in physics is given by the arrival time of the fastest of a random number of emitted photons (O'Connor and Phillips, 1984). The latter is the setting this paper started from and it will be described in more detail below.

Various extensions and other examples may be considered pointing out the relevance of the model given by (1) from an application viewpoint. For example, the maximum of a random

(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.

number of i.i.d. random variables 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).

It is worth mentioning that our model can be related to other biased data contexts, which have been studied from various points 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 many papers, see Vardi (1982), Jones (1991), de Uña-Álvarez (2004), de Uña-Álvarez and Rodríguez-Casal (2006), Asgharian et al. (2002), among others.

The interest and difficulty of the present work lies in the fact that we have three aims.

(1) The primary concern of the paper is the nonparametric estimation of the pdf f of the distribution of interest F in the model given by (1) based on an i.i.d. sample $Z_1, \dots, Z_n$ with distribution G and known link function H.
(2) Our second question is about a methodological point of view. We want to determine which general approach of density estimation should be used. Indeed, we consider hereafter kernel estimators and projection estimators and wonder which are to be preferred. More precisely, projection estimators with model selection devices and kernel estimators with data-driven bandwidth selection are constructed for the model given by (1). Adaptive projection estimators correspond to methods originally described by Barron et al. (1999), see also Lerasle (2012) for developments more specific to density estimation. These methods have been 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 estimator the recent approach of Goldenshluger and Lepski (2011) is applied to our model and considered from both a pointwise and a global point of view. Here "pointwise" refers to the estimation of the density on an interval with point-by-point bandwidth selection, in contrast to a unique global bandwidth in the "global" strategy. One may expect better results for the pointwise method when the function under consideration has inhomogeneous smoothness on the interval.
(3) Thirdly, the more model specific question is how to take into account the distortion H in the estimation procedure. Indeed, different properties of the model (1) give rise to two different strategies to correct the bias in the data. The first way is a sort of global correction of a primary estimator of g, which we call plug-in estimator, as in Navarro et al. (2015). The other way consists in using a standard density estimator of f while associating specific weights with all observations to correct the bias more locally, and we call it weighted estimator. This has been done in a different context in Rebafka et al. (2010). The same type of question arises in density estimation for censored data: the so-called Inverse Probability Correction Weights (IPCW) can be applied to the data, or a final correction can be applied to a functional estimator, see Brunel et al. (2005).

The combination of every adaptive kernel and projection estimator with each bias correction strategy finally gives rise to six different estimation procedures that are worth being compared. In this paper theoretical results on the mean-square risk of the estimators, more precisely, oracle-type risk bounds are provided. Namely the finite sample risk bounds for the adaptive kernel estimators are new. For adaptive projection estimators, part of the proofs follow the line of

Brunel et al. (2005) which makes the novelty of the results less decisive. Furthermore, a simulation study is conducted to calibrate all methods and to find out how the different estimation procedures compare in specific settings. To avoid a huge number of models, simulations are carried out for the pile-up model and an application to real fluorescence data is provided. The questions in order are: Which strategy to correct the bias has a better performance? Do projection or kernel estimators provide better results? How do the pointwise and the global strategy for bandwidth selection compare in specific examples? We are not aware of any other empirical study answering to these questions.

The paper is organized as follows. Section 2 presents the leading example and the general model. In Section 3, kernel and projection estimators are defined, and risk bounds are given in order to show why a data-driven selection of bandwidth or model is required. Section 4 explains how these procedures are performed and provides theoretical results ensuring that these strategies reach their aim and deliver an adequate data-driven squared bias-variance compromise. In the simulation study (Section 5) different aspects of the estimators are compared. Section 6 summarizes our findings. Finally, Section 7 presents the proofs for the theoretical results of the paper.

2. Model and Assumptions

2.1. **Notations.** Let $u : \mathbb{R} \mapsto \mathbb{R}$ and $v : \mathbb{R} \mapsto \mathbb{R}$ be two real functions. If u is a one-to-one map, denote u^{-1} its inverse function, that is the function verifying $u^{-1}(u(x)) = u(u^{-1}(x)) = x$ for all x. The standard convolution product is given by $u * v(x) = \int u(t)v(x-t)\mathrm{d}t$. 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, $\|u\|_\infty = \sup_{x\in\mathbb{R}} |u(x)|$. The inner product $\langle \cdot, \cdot \rangle$ is defined by $\langle u, v \rangle = \int u(t)v(t)\mathrm{d}t$.

2.2. **Our leading example: the pile-up model in time-resolved fluorescence.** Fluorescence is the phenomenon of photon emission by excited molecules. An important feature is the duration that the molecule spends in the excited state before emitting a photon, also called fluorescence lifetime. As the probability distribution of fluorescence lifetimes depends on numerous molecular features, it is a very powerful mean for physicists to observe and understand physical and chemical molecular processes. Fluorescence lifetimes are e.g. used to determine the speed of rotating molecules or to measure molecular distances and they are the heart of the fluorescence lifetime imaging microscopy (FLIM) technology (Lakowicz, 1999; Valeur, 2002).

Measurements of fluorescence lifetimes are obtained by the technique Time-Correlated Single-Photon Counting (TCSPC) (O'Connor and Phillips, 1984). Here, a laser pulse excites a random number of molecules, but for technical reasons, only the arrival time of the very first fluorescence photon striking the detector can be measured, while the other fluorescence lifetimes are unobservable. That is, one excitation produces only a single observation. So the experiment is repeated many times to generate a convenient number of observations. However, the data are not a sample from the distribution of the fluorescence lifetimes, but of the distribution of the *minimum* of the fluorescence lifetimes of a *random* number of molecules. This model is referred to as the pile-up model.

Mathematically, let p be the number of excitations or laser pulses. Denote N_i^* the number of emitted fluorescence photons after the i-th excitation and $Y_{i,j}, \ldots, Y_{i,N_i}$ the associated fluorescence lifetimes. Then the fluorescence lifetime of the *fastest* photon after the i-th excitation is the random variable Z_i^* defined by

$$Z_i^* = \min\{Y_{i,1}, \ldots, Y_{i,N_i^*}\} \, .$$

By convention, if $N_i^* = 0$, set $Z_i^* = 0$. Note that $Z_1^*, \dots, Z_p^*$ are observed, but not the variables $(Y_{i,j})_{i,j\geq 1}$ neither the numbers $N_1^*, \dots, N_p^*$.

From a physical point of view it is reasonable to suppose that $N_1^*, \dots, N_p^*$ are i.i.d random variables with Poisson distribution, $(Y_{i,j})_{i,j\geq 1}$ are positive i.i.d. random variables with pdf f and cdf F, and $(Y_{i,j})_{i,j\geq 1}$ and $(N_1^*, \dots, N_p^*)$ are independent. The parameter of the Poisson distribution of N_1^*, say μ, is known to the physicists, as it is a tuning parameter depending on the laser intensity. By the way, μ is easily estimated by the proportion of excitations where no photon event is detected, i.e. $\hat{\mu} = -\log(\frac{1}{p}\sum_i \mathbb{1}\{Z_i^* = 0\})$.

If $Z_i^* = 0$, then no photon is detected ($N_i^* = 0$) and hence no information on the distribution of the fluorescence lifetimes is obtained. So, typically, these observations are deleted from the sample. This is similar to consider random variables N_i with Poisson distribution restricted to $\mathbb{N}^* := \{1, 2, \dots\}$ and positive observations $Z_i = \min\{Y_{i,1}, \dots, Y_{i,N_i}\}$. Note that $Z_1, \dots, Z_n$ are i.i.d. random variables with common distribution. Denote G the cdf and g the pdf of Z_1. Then the pile-up model is an instance of the biased data model given in (1), since

$$\begin{aligned} G(x) &= 1 - \mathbb{P}(Z_1 > x) = 1 - \sum_{k=1}^{\infty} \mathbb{P}(N_1 = k)\mathbb{P}\left(\min_{1\leq j\leq k} Y_{1,j} > x\right) \\ &= 1 - \sum_{k=1}^{\infty} (N_1 = k)\left[\mathbb{P}(Y_{1,1} > x)\right]^k = 1 - M(1 - F(x)) \,, \end{aligned}$$

where

$$M(u) = \mathbb{E}\left[u^{N_1}\right] = \sum_{k=1}^{\infty} u^k \mathbb{P}(N_1 = k) = \frac{\mathrm{e}^{\mu u} - 1}{\mathrm{e}^{\mu} - 1}, \quad u \in [0, 1] \,,$$

with $\mathbb{P}(N_1 = k) = 1/(\mathrm{e}^{\mu} - 1)(\mu^k/k!)$ for $k = 1, 2, \dots$. Hence, we see that relation (1) is verified with $H(x) = 1 - M(1 - x)$.

2.3. **Model and assumptions.** More generally, let F and G in (1) be absolutely continuous with pdf f and g, respectively. As our goal is to recover density f from a sample from G, we first look for an expression to relate f to the distribution G of the observations and/or g. Clearly, if the link function $H : [0, 1] \to [0, 1]$ in relation (1) is a one-to-one map, then

$$F(x) = H^{-1}(G(x)) \,, \qquad x \in \mathbb{R} \,.$$

Furthermore, if H is differentiable, deriving the last relation yields $f(x) = g(x)/H'(H^{-1}(G(x)))$. Introduce the weight function w defined by

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

Then the following relation between the densities f and g holds,

$$f(x) = w(G(x))g(x) \,, \qquad x \in \mathbb{R} \,. \tag{2}$$

Concerning the assumptions on the model, first note that the link function $H : [0, 1] \to [0, 1]$ in (1) is necessarily increasing and surjective, otherwise $x \mapsto H(F(x))$ is not a cdf. It is also natural to assume that H is a one-to-one map to ensure identifiability in the sense that the aim is the estimation of the density of distribution F using a sample from G.

The weight function w is well defined under the assumption that H' exists and is bounded away from zero. Furthermore, we shall require that w is Lipschitz. If H is twice differentiable,

then $w'(u) = -H''(H^{-1}(u))/[H'(H^{-1}(u))]^3$. If furthermore there exist finite constants $a, b > 0$ such that

$$H'(u) \geq a \,, \quad |H''(u)| \leq b \,, \qquad u \in [0,1] \,, \tag{3}$$

then w is a Lipschitz function with Lipschitz constant c_w such that $c_w \leq b/a^3$.

Moreover, if there is a finite constant $d > 0$ such that

$$H'(u) \leq d \,, \qquad u \in [0,1] \,. \tag{4}$$

then w is bounded on $[0,1]$ by $0 < 1/d \leq w(u) \leq 1/a$.

Note that we may possibly require that f or g is bounded. In fact, f and g are either both bounded or both unbounded, since $a\|f\|_\infty \leq \|g\|_\infty \leq d\|f\|_\infty$.

3. Density estimators: definition and first risk bounds

This section is concerned with the construction of estimators of density f using an i.i.d. sample $Z_1, \ldots, Z_n$ from distribution G in the model given by (1). We indicate two general approaches to correct the model bias and show how different kernel and projection estimators are obtained when using these correction strategies. Finally, we state risk bounds to show the need for data driven choices of bandwidth or model dimension.

3.1. **Estimation strategies.**

The two standard approaches in nonparametric density estimation are kernel estimators and projection estimators. We consider both of them to provide estimators of f in our model.

We start with kernel estimators. Denote $\hat{g}_h^{\ker}$ the standard kernel estimator of g based on observations $Z_1, \ldots, Z_n$ from the distribution G given by

$$\hat{g}_h^{\ker}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - Z_i) \,, \quad x \in \mathbb{R} \,,$$

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 $K_h(u) = K(u/h)/h$. According to formula (2), a *plug-in estimator* of f is given by

$$\hat{f}_h^{\text{ker-P}}(x) = w(\hat{G}_n(x))\hat{g}_h^{\ker}(x) \,, \quad x \in \mathbb{R} \,,$$

where $\hat{G}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{Z_i \leq x\}$ denotes the empirical cdf associated with the observations $Z_1, \ldots, Z_n$.

Next, notice that (2) also implies that, for any measurable bounded function ψ, we have

$$\mathbb{E}_f[\psi(Y_1)] = \mathbb{E}_g\left[w(G(Z_1))\psi(Z_1)\right] \,. \tag{5}$$

This property allows the construction of another kernel type estimator of f. Indeed, if we had at hand a sample $Y_1, \ldots, Y_n$ from distribution F, then the standard kernel estimator of $f(x)$ would be given by $\hat{f}_h^{\ker}(x) = n^{-1}\sum_{i=1}^{n} K_h(x - Y_i)$, Hence, taking $\psi(z) = K_h(x - z)$ in (5), an alternative estimator of $f(x)$ based on a sample $Z_1, \ldots, Z_n$ from distribution G, that shall be close to $\hat{f}_h^{\ker}(x)$, is given by

$$\hat{f}_h^{\text{ker-W}}(x) = \frac{1}{n}\sum_{i=1}^{n} w(\hat{G}_n(Z_i))K_h(x - Z_i) \,. \tag{6}$$

Denote by $Z_{(i)}$ the i-th order statistic associated with $(Z_1, \ldots, Z_n)$ satisfying $Z_{(1)} \leq \cdots \leq Z_{(n)}$. As $w(\hat{G}_n(Z_{(i)})) = w(i/n)$, we have $\hat{f}_h^{\text{ker-W}}(x) = n^{-1}\sum_{i=1}^{n} w(i/n)K_h(x - Z_{(i)})$.

In a similar way, projection estimators can be developed for the biased data model. The general idea is to approximate g (or f) by its orthogonal projection onto some function space. Let A be the interval on which the function f is estimated.

First, suppose that the restriction of g on some interval A is square integrable, that is $g\mathbb{1}_A \in \mathbb{L}^2(A)$. Let $(\varphi_j)_{j\geq 0}$ be an orthonormal basis of $\mathbb{L}^2(A)$. Define the subspaces $S_m = \mathrm{Span}(\varphi_j, j = 0, \ldots, D_m - 1)$ of dimension D_m. The orthogonal projection g_m in the $\mathbb{L}^2$–sense of g on S_m is given by $g_m = \sum_{j=0}^{D_m-1} a_j\varphi_j$ with coefficients $a_j = \langle g, \varphi_j\rangle = \mathbb{E}[\varphi_j(Z_1)]$. A natural estimator of a_j is given by $\hat{a}_j = n^{-1}\sum_{i=1}^n \varphi_j(Z_i)$. Hence, g_m can be estimated by $\hat{g}_m^{\mathrm{proj}}(x) = \sum_{j=0}^{D_m-1} \hat{a}_j\varphi_j(x)$. Applying the plug-in approach for bias correction, an estimator of f is given by

$$\hat{f}_m^{\text{proj-P}}(x) = w(\hat{G}_n(x))\hat{g}_m^{\mathrm{proj}}(x) = w(\hat{G}_n(x)) \sum_{j=0}^{D_m-1} \left(\frac{1}{n}\sum_{i=1}^n \varphi_j(Z_i)\right)\varphi_j(x)\ , \quad x \in \mathbb{R}\ .$$

To apply the second bias-correction method, suppose that f restricted on A is square integrable, that is $f\mathbb{1}_A \in \mathbb{L}^2(A)$. Then the orthogonal projection f_m of f on S_m is given by $f_m = \sum_{j=0}^{Dm-1} b_j\varphi_j$ with coefficients $b_j = \langle f, \varphi_j\rangle = \mathbb{E}[\varphi_j(Y)]$. With $\psi = \varphi_j$ in (5), an estimator of b_j is given by $\hat{b}_j = n^{-1}\sum_{i=1}^n w(\hat{G}_n(Z_i))\varphi_j(Z_i) = n^{-1}\sum_{i=1}^n w(i/n)\varphi_j(Z_{(i)})$. Hence, a second projection-type estimator of f is given by

$$\hat{f}_m^{\text{proj-W}}(x) = \sum_{j=0}^{D_m-1} \hat{b}_j\varphi_j(x) = \sum_{j=0}^{D_m-1} \frac{1}{n}\left(\sum_{i=1}^n w\left(\frac{i}{n}\right)\varphi_j(Z_{(i)})\right)\varphi_j(x)\ , \quad x \in \mathbb{R}\ . \tag{7}$$

In the rest of the paper $(\varphi_j)_{j\geq 0}$ is the trigonometric basis on $A = [l_1, l_2]$ defined by $\varphi_j(x) = (l_2 - l_1)^{-1/2}\varphi_j^0((x - l_1)/(l_2 - l_1))$ 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 $j \geq 0$, $\varphi_{2j}(x) = \sqrt{2}\sin(2\pi jx)$ for $j \geq 1$. We consider subspaces S_m of dimension $D_m = 2m + 1$. This basis has the advantage of simplicity and provides nested models allowing for fast computations: when increasing the dimension from D_m to D_{m+1}, only two more coefficients are taken into account, the previous ones being still valid. In other words, $\hat{f}_{m+1}^{\text{proj-W}}(x) = \hat{f}_m^{\text{proj-W}}(x) + \hat{b}_{2m+1}\varphi_{2m+1}(x) + \hat{b}_{2m+2}\varphi_{2m+2}(x)$.

3.2. **Pointwise and integrated risk of kernel estimators.** The following proposition is easily shown and proved in Section 7.

Proposition 3.1.
(i) If f is bounded, K is square-integrable and conditions (3) and (4) are fulfilled, then, for any x_0, the estimator $\hat{f}_h^{ker\text{-}W}$ defined by (6) satisfies

$$\mathbb{E}\left[(\hat{f}_h^{ker-W}(x_0) - f(x_0))^2\right] \leq 3(K_h * f(x_0) - f(x_0))^2 + \frac{C_1}{nh}\ ,$$

where $C_1 = 3\left(\|w\|_2^2 + 5db^2/(4a^6)\right)\|f\|_\infty\|K\|_2^2$ and $\|w\|_2^2 = \int_0^1 w^2(u)\mathrm{d}u$.
(ii) Assume that f is a square integrable density and that (2) and (3) hold. Then the MISE of $\hat{f}_h^{ker\text{-}W}$ defined by (6) satisfies

$$\mathbb{E}\left[\|\hat{f}_h^{ker\text{-}W} - f\|_2^2\right] \leq 3\|K_h * f - f\|^2 + \frac{C_2}{nh}\ ,$$

where $C_2 = 3(\|w\|_2^2 + b^2/a^6)\|K\|_2^2$.

We see that both bounds are the sum of two terms, that can be interpreted as a squared bias term and a variance term. Indeed, the squared bias terms $3(K_h * f(x_0) - f(x_0))^2$ and $3\|K_h * f - f\|^2$ decrease when h tends to zero, and the variance terms C_1/nh and C_2/nh increase

when h increases. For this reason bandwidth selection devices always aim at achieving a data-driven compromise between these two antagonist terms, in order to minimize the corresponding square risk.

Under additional assumptions on the regularity of the function f and on the kernel, more precise orders for the bias term may be obtained. Note $\lfloor x \rfloor$ the largest integer not greater than x. For pointwise estimation e.g. suppose that

(H1) f belongs to the Hölder class $\Sigma(\beta, L)$ given by

$$\Sigma(\beta, C) = \{f : \mathbb{R} \to \mathbb{R}, f^{(\ell)} \text{exists for } \ell = \lfloor \beta \rfloor \text{and } |f^{(\ell)}(x) - f^{(\ell)}(x')| \leq L|x - x'|^{\beta - \ell}, \forall x, x' \in \mathbb{R}\},$$

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

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 also considered in Tsybakov (2004) and Kerkyacharian et al. (2001), where examples of such kernels are given. Now, under (H1)-(H2), Tsybakov (2004) shows that $(K_h * f(x_0) - f(x_0))^2 \leq \tilde{C}_1^2 h^{2\beta}$ with $\tilde{C}_1 = L \int |u|^{\beta} |K(u)| \mathrm{d}u / \ell!$. Concerning the integrated risk, when it is computed on a compact set, it has the same order under (H1)-(H2). For more general settings, one may consider functions belonging to Nikol'ski classes. We refer the reader to Tsybakov (2004) for details.

Remark 3.1. The bias is minimal when the order of the kernel is larger than the regularity of the function. Since the regularity is unknown, it is natural to think that the higher the kernel order, the better. But we failed to illustrate it on the simulations, and simply used symmetric kernels (which are of order one).

Risk bounds for the plug-in kernel estimator $\hat{f}_h^{\text{ker-P}}$ 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.

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

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

$$\mathbb{E}\left[(\hat{f}_h^{\text{ker-P}}(x_0) - f(x_0))^2\right] \leq \frac{2}{a^2}\mathbb{E}\left[(\hat{g}_h^{\text{ker}}(x_0) - g(x_0))^2\right] + \frac{2b^2 g^2(x_0)}{a^6 n} . \tag{8}$$

Thus, (i) of Proposition 3.1 holds for $\hat{f}_h^{\text{ker-P}}$, under the assumption that g is bounded and C_1 is replaced by $C_1' = (2\|g\|_\infty / a^2)\left(\|K\|_2^2 + db^2\|g\|_\infty / a^4\right)$.

Analogously, in the integrated case, we get

$$\mathbb{E}\left[\|\hat{f}_h^{\text{ker-P}} - f\|_2^2\right] \leq \frac{2}{a^2}\mathbb{E}\left[\|\hat{g}_h^{\text{ker}} - g\|_2^2\right] + 2\frac{b^2\|g\|_2^2}{a^6 n} , \tag{9}$$

instead of (ii) of Proposition 3.1. For bounds on $\mathbb{E}\left[(\hat{g}_h^{\text{ker}}(x_0) - g(x_0))^2\right]$ or $\mathbb{E}\left[\|\hat{g}_h^{\text{ker}} - g\|_2^2\right]$, we refer to Tsybakov (2004).

In all the previous cases, 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 is of order $n^{-2\beta/(2\beta+1)}$. As β is unknown, this choice cannot be done in that naive way and thus data-driven methods for bandwidth selection are required.

3.3. **Risk of the projection estimator.** The following risk bound holds for the projection estimator.

Proposition 3.2. *Consider the estimator* $\hat{f}_m^{proj\text{-}W}$ *defined by (7), then*

$$\mathbb{E}\left[\|\hat{f}_m^{proj\text{-}W} - f\mathbb{1}_A\|_2^2\right] \leq \|f\mathbb{1}_A - f_m\|_2^2 + C_3\frac{D_m}{n}\ , \tag{10}$$

where $C_3 = 2(\|w\|_2^2 + b^2/a^6)$.

Again, the risk bound involves a squared bias term, $\|f\mathbb{1}_A - f_m\|_2^2$, which decreases when D_m increases, and a variance term, $C_3 D_m/n$, which increases with D_m.
To evaluate the order of 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 (10) 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).

4. Bandwidth and model selection

4.1. **Model selection.** We start with the projection estimators for which the classical penalization approach by Barron et al. (1999) can be easily explained and applied. In Section 3.3 it is made clear that the risk is minimized when a bias-variance trade-off is achieved. In practice this is done by looking for the value m in $\mathcal{M}_n := \{1, \ldots, m_n, m_n \in \mathbb{N}, m_n \leq n\}$ which minimizes an estimate of the risk and more precisely an estimate of the bound $\|f\mathbb{1}_A - f_m\|_2^2 + C_3 D_m/n$. The two terms of the previous sum are considered separately. First the bias can be written $\|f\mathbb{1}_A - f_m\|_2^2 = \|f\mathbb{1}_A\|_2^2 - \|f_m\|_2^2$, since f_m is an $\mathbb{L}^2$-orthogonal projection of $f\mathbb{1}_A$ on S_m. Thus, this term is estimated by canceling $\|f\mathbb{1}_A\|_2^2$ which does not depend on m and replacing $\|f_m\|_2^2$ by the natural estimate $\|\hat{f}_m^{\text{proj-W}}\|_2^2$. Next, the variance is estimated by its bound but with a special care on constants. Here the variance bound has order D_m/n. Then, in C_3, only the first part $2\|w\|_2^2$ is due to the variance, the other part can be shown to be negligible. Lastly, the factor 2 is arbitrary, and has to be evaluated more precisely for the complete procedure to work. This is why it is replaced by a numerical constant κ_1^f: the theoretical result will say that the procedure works if κ_1^f is larger than a minimal value. In "simple" theoretical case (e.g. pure white noise models), the minimal value can be obtained from the proof of the results and it can be proved that the risk of the estimate is explosive if the constant is taken too small, see Birgé and Massart (2007). So, the constant κ_1^f has to be calibrated and methods to do it have been developed, see Baudry et al. (2012) for both principles and practical algorithms.

More precisely, we select models $\hat{m}^g$ and $\hat{m}^f$ defined by

$$\hat{m}^g = \arg\min_{m\in\mathcal{M}_n}\left[-\|\hat{g}_m^{\text{proj}}\|_2^2 + \text{pen}^g(m)\right],\ \hat{m}^f = \arg\min_{m\in\mathcal{M}_n}\left[-\|\hat{f}_m^{\text{proj-W}}\|_2^2 + \text{pen}^f(m)\right],$$

where the penalty terms $\text{pen}^g(m)$ and $\text{pen}^f(m)$ are defined by

$$\text{pen}^g(m) = \kappa_1^g\frac{D_m}{n}\ ,\quad \text{pen}^f(m) = \kappa_1^f\|w\|_2^2\frac{D_m}{n}\ ,$$

with constants κ_1^g, κ_2^f calibrated by preliminary simulations. Then we consider the density estimates $\hat{f}_{\hat{m}^g}^{\text{proj-P}}(x) = w(\hat{G}_n(x))\hat{g}_{\hat{m}^g}^{\text{proj}}(x)$ and $\hat{f}_{\hat{m}^f}^{\text{proj-W}}(x)$. The following result can be shown.

Theorem 4.1. *Assume that $m_n \leq O(\sqrt{n})$ and that f is bounded on A, i.e. $\|f_A\|_\infty < \infty$. Then there exists a numerical constant κ_0 such that for any $\kappa_1^f \geq \kappa_0$ we have*

$$\mathbb{E}\left[\|\hat{f}^{proj\text{-}W}_{\hat{m}^f} - f\mathbb{1}_A\|_2^2\right] \leq C \inf_{m\in\mathcal{M}_n}\left(\|f\mathbb{1}_A - f_m\|_2^2 + \|w\|_2^2\frac{D_m}{n}\right) + K\frac{\log^2(n)}{n}\,, \tag{11}$$

where C is a numerical constant and K depends on a, b, $\|f\mathbb{1}_A\|_\infty$ and the basis.

Risk bounds of the form (11) are often called oracle inequality: it means that the data driven estimator $\hat{f}^{\text{proj-W}}_{\hat{m}^f}$ performs the bias-variance compromise, up to the multiplicative constant C and the residual $K\log^2(n)/n$. Note that the last term is clearly negligible with respect to the order of the infimum 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 7.

Lastly, using inequality (9) yields that $\hat{f}^{\text{proj-P}}_{\hat{m}^g}$ 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. More precisely, we have

$$\mathbb{E}\left[\|\hat{f}^{\text{proj-P}}_{\hat{m}^g} - f\mathbb{1}_A\|_2^2\right] \leq C\frac{2}{a^2}\inf_{m\in\mathcal{M}_n}\left(\|g\mathbb{1}_A - g_m\|_2^2 + \frac{D_m}{n}\right) + K'\frac{\log^2(n)}{n}\,,$$

where C is the same numerical constant as in the theorem and K' is a constant depending on a, b, $\|g\mathbb{1}_A\|_\infty$, $\|g\|^2$ and the basis.

4.2. **Pointwise bandwidth selection.** In this section, devices for a data-driven selection of the bandwidth h for the kernel estimators $\hat{g}^{\text{ker}}_h$ and $\hat{f}^{\text{ker-W}}_h$ are considered. Here we follow Goldenshluger and Lepski (2011) for pointwise adaptive and global adaptive estimators: indeed, this recent method allows for convenient and rigorous control of the estimators, relying on empirical processes study and powerful deviation inequalities. Moreover, part of the results are nonasymptotic, contrary to many kernel studies. Applying this method in the case of biased data is a novelty, in theory and in practice.

Denote by $\mathcal{H}$ a finite collection of bandwidths given by

$$\mathcal{H} = \left\{h_k, k = 1,\dots,M_n, \frac{1}{n} \leq h_k \leq 1\right\}\,, \quad \text{with } M_n \text{ an integer, } M_n \leq n\,. \tag{12}$$

First, consider the estimator $\hat{g}^{\text{ker}}_h$ of g (used to define $\hat{f}^{\text{ker-P}}_h$). Let x_0 be fixed and consider the pointwise adaptive bandwidth selection method of Goldenshluger and Lepski (2011), that is a bandwidth $\hat{h}^g(x_0)$ depending on x_0 is selected. Introduce the estimator $\hat{g}^{\text{ker}}_{h,h'}$ depending on two bandwidths h, h' defined by

$$\hat{g}^{\text{ker}}_{h,h'}(x) = K_{h'} * \hat{g}^{\text{ker}}_h(x)\,, \quad x\in\mathbb{R}\,.$$

Notice the symmetry of $\hat{g}^{\text{ker}}_{h,h'}$ in h and h'. As for model selection, the principle is to approximate the pointwise squared bias. The specific idea here is based on the fact that $K_{h'} * (K_h * g - g)$ tends to $K_h * g - g$ when h' tends to zero. Thus for small h', $\left(\hat{g}^{\text{ker}}_{h,h'}(x_0) - \hat{g}^{\text{ker}}_{h'}(x_0)\right)^2$ may be a relevant approximation of the squared bias. Unfortunately, this estimate has a bias that is of

the same order as the variance. To take account of this bias, we define

$$V_0^g(h) = \kappa_2^g \|K\|_1^2 \|K\|_2^2 \|g\|_\infty \frac{\log n}{nh} ,$$
$$A_0^g(h, x_0) = \sup_{h' \in \mathcal{H}} \left[\left(\hat{g}_{h,h'}^{\ker}(x_0) - \hat{g}_{h'}^{\ker}(x_0) \right)^2 - V_0^g(h') \right]_+ .$$

The term $A_0^g(h, x_0)$ is the squared bias estimate, and $V_0^g(h)$ can be interpreted as a variance estimate, augmented by a $\log(n)$ factor. Consequently, one shall use the bandwidth $\hat{h}^g(x_0)$ minimizing the sum of both terms, that is,

$$\hat{h}^g(x_0) = \arg\min_{h \in \mathcal{H}} \{ A_0^g(h, x_0) + V_0^g(h) \} .$$

Hence, the pointwise adaptive version of estimator $\hat{f}_h^{\text{ker-P}}(x_0)$ of $f(x_0)$ is given by

$$\hat{f}_{\hat{h}^g(x_0)}^{\text{ker-P}}(x_0) = w(\hat{G}_n(x_0)) \hat{g}_{\hat{h}^g(x_0)}^{\ker}(x_0) .$$

Here, the existence of a minimal value for the constant κ_2^g in the variance term $V_0^g(h)$ has been studied only very recently in Lacour and Massart (2015), and the calibration procedures are not yet well understood. This is probably due to the fact that the variance estimate $V_0(h)$ plays two different roles (variance estimate and bias correction). This is why the empirical study of Section 5 is of interest.

For the estimator $\hat{f}_h^{\text{ker-W}}(x_0)$ we proceed in the same way. Define

$$\hat{f}_{h,h'}^{\text{ker-W}}(x) = K_{h'} * \hat{f}_h^{\text{ker-W}}(x) , \quad x \in \mathbb{R} ,$$
$$V_0^f(h) = \kappa_2^f \|K\|_1^2 \|K\|_2^2 \frac{\|f\|_\infty}{a} \frac{\log n}{nh} ,$$
$$A_0^f(h, x_0) = \sup_{h' \in \mathcal{H}} \left[\left(\hat{f}_{h,h'}^{\text{ker-W}}(x_0) - \hat{f}_{h'}^{\text{ker-W}}(x_0) \right)^2 - V_0^f(h') \right]_+ .$$

Then, the pointwise adaptive version of estimator $\hat{f}_h^{\text{ker-W}}(x_0)$ of $f(x_0)$ is given by

$$\hat{f}_{\hat{h}^f(x_0)}^{\text{ker-W}}(x_0) , \quad \text{with} \quad \hat{h}^f(x_0) = \arg\min_{h \in \mathcal{H}} \left\{ A_0^f(h, x_0) + V_0^f(h) \right\} .$$

As mentioned above, the numerical constants κ_2^g and κ_2^f are calibrated by simulation. The other constants in $V_0^g(h)$ and $V_0^f(h)$ are known, except $\|g\|_\infty$ or $\|f\|_\infty$ which in practice are replaced with some estimators (see Section 5.2). We refer to Section 3.3 in Comte et al. (2011) for an example of theoretical penalty involving such estimate; this development is omitted here to avoid additional technicalities. Let us mention that it is an open question to determine if it would be clever to look for variance estimates $V_0^g(h)$, $V_0^f(h)$ depending on x_0: different estimates can be considered, which would imply both theoretical and empirical difficulties.

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

Theorem 4.2. *Assume that f is bounded, K is integrable and bounded and that conditions (3), (4) hold. Let $\mathcal{H}$ be given by (12) and assume that for all $h \in \mathcal{H}$, $nh \geq \log^2(n)$ and that there exists some finite constant S (independent of n) such that*

$$\frac{1}{n} \sum_{h \in \mathcal{H}} \frac{1}{h} \leq S . \tag{13}$$

Then, there exist a numerical constant $\tilde{\kappa}_0$ such that for $\kappa_2^f \geq \tilde{\kappa}_0$,

$$\mathbb{E}\left[\left(\hat{f}^{ker\text{-}W}_{\hat{h}^f(x_0)}(x_0) - f(x_0)\right)^2\right] \leq C^* \inf_{h\in\mathcal{H}}\left(\|K_h * f - f\|_\infty^2 + V_0^f(h)\right) + \bar{C}\frac{\log n}{n}\,, \tag{14}$$

where C^* is a constant depending on $\|K\|_1$ only, and $\bar{C} > 0$ a constant depending on $a, b, \|f\|_\infty, S$, $\|K\|_1$, $\|K\|_2$ and $\|K\|_\infty$.

The definition of the bandwidth collection $\mathcal{H}$ via (12) is very general, only condition (13) requires some comments and illustration. We give two examples satisfying (13).

(C1) The collection $\mathcal{H} = \{h_k = 1/k, k = 1, \ldots, [\sqrt{n}]\}$ satisfies (13) with $M_n = [\sqrt{n}]$ and $S = 1$.

(C2) The collection $\mathcal{H} = \{h_k = 2^{-k}, k = 1, \ldots, [\log_2(n)]\}$ is another example satisfying (13) with $M_n = [\log_2(n)]$ and $S = 2$.

Note that under Assumption (H1)-(H2), $\|K_h * f - f\|_\infty^2 \leq C_2^2 h^{2\beta}$ since the bound given for $(K_h * f(x_0) - f(x_0))^2$ does not depend on x_0. In Theorem 4.2 we keep the general form for the squared bias since it does not require any regularity condition on f. The order of the bias term follows without requiring to know any unreachable constant.

Inequality (14) implies that the procedure almost performs the best possible compromise between the two terms of the bound given in Proposition 3.1: the bias-variance trade-off is achieved, with a loss of order $\log(n)$, which is classical for pointwise adaptive procedures. If Assumptions (H1)-(H2) are fulfilled, then the right-hand side of (14) 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 collection [C2], and also for [C1] if $\beta \geq 1/2$. Moreover, in density estimation (corresponding to $w \equiv 1$), the $\log(n)$-loss is known to be unavoidable and thus adaptive minimax (see Butucea (2000)).

The results for the plug-in kernel estimators are a consequence of the previous ones. Clearly, inequality (14) holds for $w \equiv 1$, and thus by (8) and under the assumptions of Theorem 4.2, there exists a constant κ_0^* such that for $\kappa_2^g \geq \kappa_0^*$, we have

$$\mathbb{E}\left[\left(\hat{f}^{\text{ker-P}}_{\hat{h}^g(x_0)}(x_0) - f(x_0)\right)^2\right] \leq C^* \frac{2}{a^2} \inf_{h\in\mathcal{H}}\left(\|K_h * g - g\|_\infty + V_0^g(h)\right) + \bar{C}''\frac{\log n}{n}\,,$$

where C^* is the same constant as in the theorem and $\bar{C}''$ is a constant depending on $a, b, \|g\|_\infty, S$, $\|K\|_1$, $\|K\|_2$ and $\|K\|_\infty$. If in addition g belongs to a Hölder class with regularity parameter β^*, then $\|K_h * g - g\|_\infty \leq h^{2\beta^*}$. Therefore, the risk bound on $\hat{f}^{\text{ker-P}}_{\hat{h}^g(x_0)}(x_0)$ is an automatic compromise related to the regularity of g and provides the best possible rate if f and g belong to the same Hölder space (*i.e.* $\beta = \beta^*$).

4.3. **Global bandwidth selection.** In a similar way, a procedure for global bandwidth selection is developed, that is a bandwidth $\hat{h}$ is selected that is valid for all x in $\mathbb{R}$. Denote

$$\begin{aligned}
V^g(h) &= \kappa_3^g \|K\|_1^2 \|K\|_2^2 \frac{1}{nh}\,, \quad V^f(h) = \kappa_3^f \max(\|K\|_1^2, 1)\|K\|_2^2 \frac{\|w\|_2^2}{nh}\,,\\
A^g(h) &= \sup_{h'\in\mathcal{H}}\left(\|\hat{g}^{\text{ker}}_{h,h'} - \hat{g}^{\text{ker}}_{h'}\|^2 - V^g(h')\right)_+\,,\\
A^f(h) &= \sup_{h'\in\mathcal{H}}\left(\|\hat{f}^{\text{ker-W}}_{h,h'} - \hat{f}^{\text{ker-W}}_{h'}\|^2 - V^f(h')\right)_+\,,\\
\hat{h}^g &= \arg\min_{h\in\mathcal{H}}\left(A^g(h) + V^g(h)\right)\,, \quad \hat{h}^f = \arg\min_{h\in\mathcal{H}}\left(A^f(h) + V^f(h)\right)\,,
\end{aligned}$$

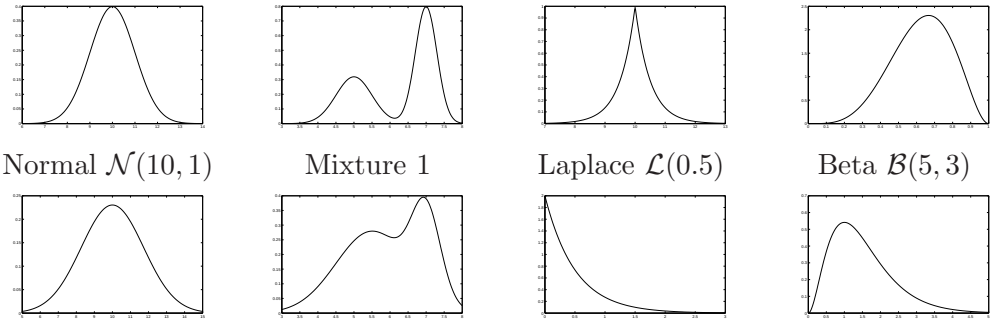


FIGURE 1. Densities for calibration and simulations.

where κ_3^g and κ_3^f are numerical constants calibrated by simulations. Then global adaptive estimators of f are given by

$$\hat{f}^{\text{ker-P}}_{\hat{h}^g}(x) = w(\hat{G}_n(x))\hat{g}^{\text{ker}}_{\hat{h}^g}(x) \quad \text{and} \quad \hat{f}^{\text{ker-W}}_{\hat{h}^f}(x)\,, \quad x \in \mathbb{R}\,,$$

for which a risk bound is given in Theorem 4.3 hereafter.

Theorem 4.3. *Assume that f is bounded and square-integrable, K is integrable and bounded and that conditions (3), (4) hold. Let $\mathcal{H}$ given by (12) and assume that for any $c > 0$, there exists a finite constant $B(c)$ (independent of n) such that*

$$\sum_{h\in\mathcal{H}} \mathrm{e}^{-c/h} \leq B(c)\,. \tag{15}$$

Denote $f_h = K_h * f$. Then there exists a constant $\tilde{\tilde{\kappa}}_0$ for any $\kappa_1^f \geq \tilde{\tilde{\kappa}}_0$, we have

$$\mathbb{E}\left[\|\hat{f}^{ker\text{-}W}_{\hat{h}^f} - f\|_2^2\right] \leq C \inf_{h\in\mathcal{H}} \left\{\|K\|_1^2\|f - f_h\|_2^2 + V^f(h)\right\} + C'\frac{\log n}{n}\,,$$

where C is a numerical constant and C' is a constant depending on a, b, $\|f\|_\infty$, $B(\|K\|_2^2/(12\|f\|_\infty))$, $\|K\|_1$, $\|K\|_2$, $\|K\|_\infty$.

It is worth emphasizing that, as previously, this inequality proves that a bias-variance tradeoff 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 $V^f(h)$, contrary to $V_0^f(h)$. Then, under regularity conditions on g and an order condition on K, we may obtain standard convergence rates.

As previously, for the plug in estimator in the integrated case, we get by (9) and by Theorem 4.3 that

$$\mathbb{E}\left[\|\hat{f}^{\text{ker-P}}_{\hat{h}^g} - f\|_2^2\right] \leq C\frac{2}{a^2} \inf_{h\in\mathcal{H}} \left\{\|K\|_1^2\|g - g_h\|_2^2 + V^g(h)\right\} + C''\frac{\log n}{n}\,,$$

where C is the same numerical constant as in the theorem and C'' is a constant depending on a, b, $\|g\|_\infty$, $\|g\|_2$, $B(\|K\|_2^2/(12\|g\|_\infty))$, $\|K\|_1$, $\|K\|_2$, $\|K\|_\infty$.

5. Experimental study

Now we have at hand six estimators with (nearly-)optimal rates corresponding to different statistical methods that are intrinsically interesting to compare. From a theoretic point of view, all the procedures are proved to deliver the best possible tradeoff when selecting the model or the bandwidth. Now we study their practical performances and try to answer the questions on

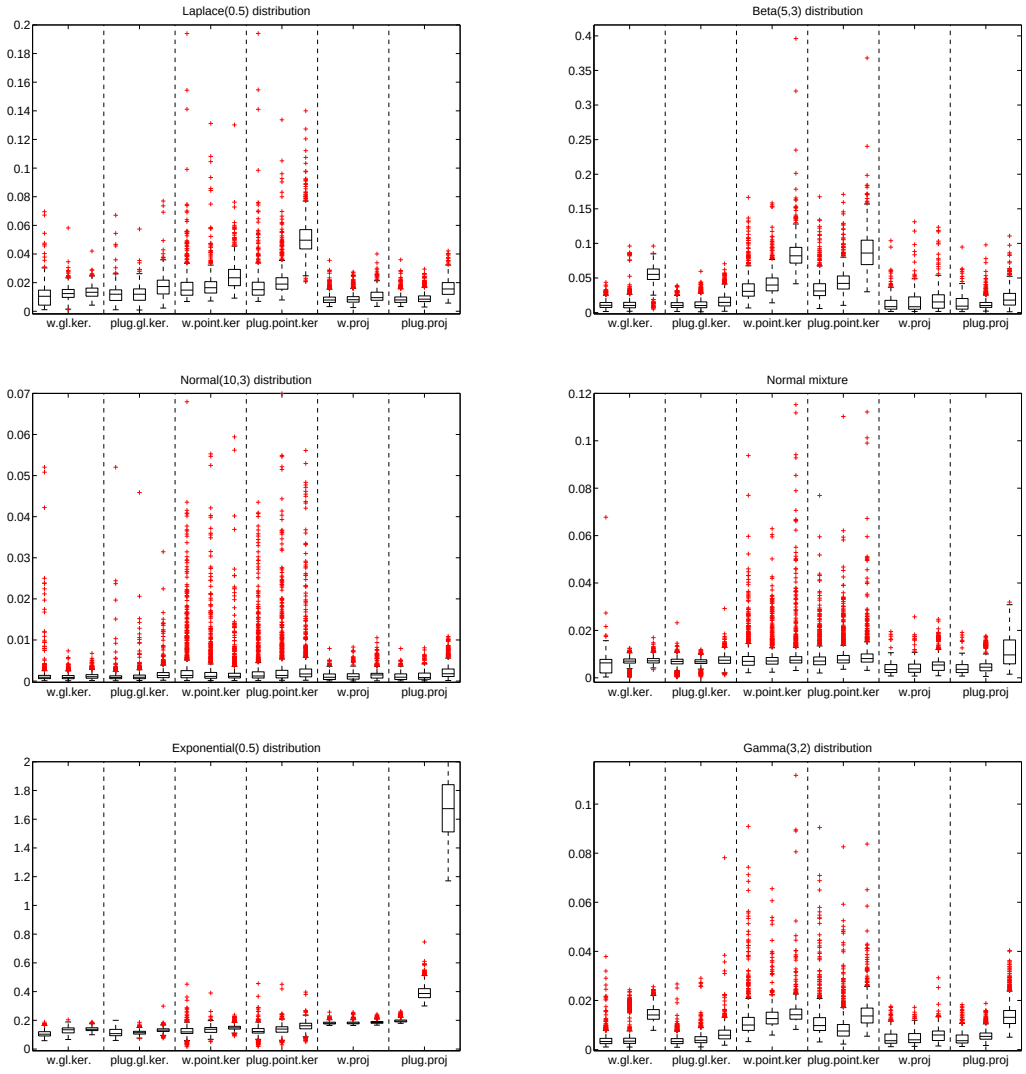


FIGURE 2. Each boxplot represents the values of MISE*1000 of the six estimators computed on 1000 datasets for 6 different distributions with $n = 500$ and for $\mu = 0.1, 0.8, 2$ (from left to right in the figure).

the best bias-correction method, the best estimation strategy and the best data-driven selection approach from a practical viewpoint. To this end this section provides a simulation study and numerical results on a real data example.

5.1. **Pile-up model.** All simulations are carried out in the pile-up model which is motivated by an application in fluorescence. The model is described in detail in Subsection 2.2. We just recall that observations are of the form

$$Z_i = \min\{Y_{i,1}, \dots, Y_{i,N_i}\}\ , \quad i = 1, \dots, n\ , \tag{16}$$

where N_i follows the (restricted) Poisson distribution with parameter μ and the fluorescence lifetimes $Y_{i,j}$ are i.i.d. with density f. We have $H(u) = (1 - \mathrm{e}^{-u\mu})/(1 - \mathrm{e}^{-\mu})$, so that the model assumptions (3) and (4) are fulfilled with $a = \mu/(\mathrm{e}^{\mu} - 1)$, $d = \mu\mathrm{e}^{\mu}/(\mathrm{e}^{\mu} - 1)$ and $b = \mu^2/(1 - \mathrm{e}^{-\mu})$.

TABLE 1. Mean MISE*1000 values for the six different estimators in 36 different settings. Each MISE value is based on 1000 simulated datasets.

	Laplace distribution $\mathcal{L}(0.5)$						Beta distribution $\mathcal{B}(5,3)$					
μ	0.1		0.8		2		0.1		0.8		2	
n	500	2000	500	2000	500	2000	500	2000	500	2000	500	2000
$\hat{f}^{\text{ker-W}}_{\hat{h}^f}$	10.5	4.03	12.4	**2.74**	13.6	6.94	11.7	7.38	12.4	7.37	55.1	7.55
$\hat{f}^{\text{ker-P}}_{\hat{h}^g}$	11.6	3.08	11.7	3.06	17.2	4.05	**11.5**	7.38	**12.0**	7.14	**17.2**	10.1
$\hat{f}^{\text{ker-W}}_{\hat{h}^f(x)}$	17.5	9.18	18.2	10.6	24.6	12.9	34.8	18.7	45.5	18.1	84.8	23.7
$\hat{f}^{\text{ker-P}}_{\hat{h}^g(x)}$	17.5	9.47	21.7	9.64	51.0	16.6	35.0	17.5	45.5	16.0	89.5	23.5
$\hat{f}^{\text{proj-W}}_{\hat{m}^f}$	8.61	**2.93**	**8.79**	3.03	**10.6**	**3.74**	12.5	4.03	13.5	4.11	17.7	**4.74**
$\hat{f}^{\text{proj-P}}_{\hat{m}^g}$	**8.60**	2.94	9.28	3.17	16.5	5.74	12.6	**3.88**	12.1	**3.59**	20.6	5.57
	Normal distribution $\mathcal{N}(10,3)$						Mixture $\frac{7}{10}\mathcal{N}(5.5,1)+\frac{3}{10}\mathcal{N}(7,0.16)$					
μ	0.1		0.8		2		0.1		0.8		2	
n	500	2000	500	2000	500	2000	500	2000	500	2000	500	2000
$\hat{f}^{\text{ker-W}}_{\hat{h}^f}$	1.46	1.14	**1.07**	0.434	**1.24**	0.463	5.31	1.81	6.81	1.61	7.31	4.02
$\hat{f}^{\text{ker-P}}_{\hat{h}^g}$	1.15	0.584	1.29	0.691	1.80	0.880	6.31	1.58	6.75	1.61	7.62	1.82
$\hat{f}^{\text{ker-W}}_{\hat{h}^f(x)}$	3.61	2.67	2.94	1.62	2.51	1.95	8.57	5.02	8.39	5.39	9.69	7.17
$\hat{f}^{\text{ker-P}}_{\hat{h}^g(x)}$	3.61	2.20	2.94	2.36	3.47	1.88	8.37	5.00	9.28	6.76	9.69	8.02
$\hat{f}^{\text{proj-W}}_{\hat{m}^f}$	**1.12**	**0.220**	1.22	**0.223**	1.52	**0.285**	**3.98**	1.18	**4.21**	**1.24**	**5.57**	**1.50**
$\hat{f}^{\text{proj-P}}_{\hat{m}^g}$	**1.12**	0.222	1.25	0.282	2.27	1.05	3.99	**1.17**	4.94	1.26	11.2	4.77
	Exponential distribution $\mathcal{E}(0.5)$						Gamma distribution $\Gamma(3,0.5)$					
μ	0.1		0.8		2		0.1		0.8		2	
n	500	2000	500	2000	500	2000	500	2000	500	2000	500	2000
$\hat{f}^{\text{ker-W}}_{\hat{h}^f}$	**109**	**96.4**	129	**97.2**	138	**98.3**	4.02	2.86	4.88	2.57	14.5	2.67
$\hat{f}^{\text{ker-P}}_{\hat{h}^g}$	114	97.6	**115**	98.6	**132**	105	**3.76**	2.66	**4.36**	1.77	6.61	2.32
$\hat{f}^{\text{ker-W}}_{\hat{h}^f(x)}$	125	105	135	109	150	120	11.9	4.71	13.4	4.83	15.2	9.17
$\hat{f}^{\text{ker-P}}_{\hat{h}^g(x)}$	126	105	136	105	160	104	11.7	4.60	8.99	5.42	14.6	10.7
$\hat{f}^{\text{proj-W}}_{\hat{m}^f}$	182	169	182	169	186	171	4.52	**1.48**	4.72	**1.52**	**5.95**	**1.87**
$\hat{f}^{\text{proj-P}}_{\hat{m}^g}$	197	184	395	398	1712	1810	4.55	1.49	5.76	1.83	14.1	4.35

5.2. **Computational issues.** We implemented the adaptive pointwise kernel estimators $\hat{f}^{\text{ker-P}}_{\hat{h}^g(x_0)}$ and $\hat{f}^{\text{ker-W}}_{\hat{h}^f(x_0)}$ defined in Subsection 4.2 for different x_0 in an interval, the adaptive global kernel estimators $\hat{f}^{\text{ker-P}}_{\hat{h}^g}$ and $\hat{f}^{\text{ker-W}}_{\hat{h}^f}$ given in in Subsection 4.3 as well as the adaptive projection estimators $\hat{f}^{\text{proj-P}}_{\hat{m}^g}$ and $\hat{f}^{\text{proj-W}}_{\hat{m}^f}$ described in Section 4.1.

For the kernel estimators the bandwidth collection (C2) is used. Indeed, collection (C1) is much larger without leading to proportionally better results. Moreover, we used the gaussian kernel of order 1, i.e. $K(u) = \mathrm{e}^{-u^2}/\sqrt{2\pi}$. In the quantities $V_0^g(h)$ and $V_0^f(h)$ the value of $\|g\|_\infty$ resp. $\|f\|_\infty$ is replaced by some estimator. More precisely, $\|g\|_\infty$ is approximated by the 95th percentile of $\{\max_{x_0} \hat{g}^{\text{ker}}_h(x_0), h \in \mathcal{H}\}$. Likewise, $\|f\|_\infty$ is approximated by the 95th percentile of $\{\max_{x_0} \hat{f}^{\text{ker}}_h(x_0), h \in \mathcal{H}\}$. Terms involving the L_2-norm as e.g. $\|\hat{g}^{\text{ker}}_{h,h'} - \hat{g}^{\text{ker}}_{h'}\|_2^2$ in $A^g(h)$ are approximated by Riemann-type discretization.

We observed that the projection estimators are much improved by normalizing $\hat{f}^{\text{proj-P}}_{\hat{m}^g}$ and $\hat{f}^{\text{proj-W}}_{\hat{m}^f}$ such that their integrals equal 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 $\int \hat{f}^{\text{ker}}_h(x)\mathrm{d}x = n^{-1}\sum_{i=1}^n w(i/n) \approx \int_0^1 w(u)\mathrm{d}u = 1$.

5.3. **Calibration.** All our estimators involve constants, namely $\kappa^g_j, \kappa^f_j, j = 1, 2, 3$, that have to be calibrated. Here, simulations are carried out to determine the best values of these constants, that means, that we are looking for the value such that the associated MISE is minimal.

More precisely, consider any of our adaptive estimators, that we denote for a moment by $\hat{f}_\kappa$ to stress the dependence of the estimator on some constant κ that has to be calibrated. We fix some fine grid $\mathcal{K} = [\kappa_1, \ldots, \kappa_K]$ of candidate values for κ. Furthermore, we choose some sample size n, some Poisson parameter μ and some density f. Then a dataset of the corresponding pile-up model is generated and, for every $\kappa \in \mathcal{K}$, the estimator $\hat{f}_\kappa$ and the associated $\text{MISE}_\kappa = \|\hat{f}_\kappa - f\|^2$ are evaluated on this dataset. This is repeated for 1000 datasets and the mean values of the MISE_κ for every $\kappa \in \mathcal{K}$ are computed. The latter are represented in Figure 4 for all of our six estimators and for different choices of the sample size ($n = 500$ and $n = 2000$), different Poisson parameters ($\lambda \in \{0.1, 0.8, 2\}$) and different densities f, namely

- normal distribution $\mathcal{N}(10, 1)$,
- a mixture of two normal distributions, $\frac{2}{5}\mathcal{N}(5, 0.25) + \frac{3}{5}\mathcal{N}(7, 0.09)$,
- Laplace distribution $\mathcal{L}(0.5)$ (with $f(x) = \mathrm{e}^{-2|x|}$),
- Beta $\mathcal{B}(5, 3)$.

These densities are of quite different form, see Figure 1 for illustration.

Finally, the constant κ of an estimator $\hat{f}_\kappa$ is chosen as the value in $\mathcal{K}$ such that the mean MISE_κ values are minimized (or at least small) in all here considered setups.

Our first observation is that the MISE_κ-curves in Figure 4 are rather similar for both bias correction methods. In other words, the weighted version of an estimator (i.e. $\hat{f}^{\text{ker-W}}_{\hat{h}^f}$, $\hat{f}^{\text{ker-W}}_{\hat{h}^f(x_0)}$, $\hat{f}^{\text{proj-W}}_{\hat{m}^f}$ resp.) produces very similar MISE_κ-curves as its plug-in counterpart ($\hat{f}^{\text{ker-P}}_{\hat{h}^g}$, $\hat{f}^{\text{ker-P}}_{\hat{h}^g(x_0)}$, $\hat{f}^{\text{proj-P}}_{\hat{m}^g}$ resp.). As a consequence, similar constants may be used for both versions of an estimator. However, the MISE_κ-curves are quite different from one estimation strategy to another.

The projection estimators seem to be quite robust with regard to the choice of κ as the MISE_κ-curves are rather flat on a large interval in all settings. In the following we set $\kappa^g_1 = \kappa^f_1 = 3.5$. Concerning the global kernel estimators MISE_κ-curves are less flat, but any value around 1.5 seems to be fine in all set-ups. We will choose $\kappa^f_3 = 1.4$ and $\kappa^g_3 = 1.7$. Finally, a good choice for the constants of the pointwise kernel estimators is $\kappa^f_2 = 0.8$ and $\kappa^g_2 = 0.9$.

Let us make a remark on the choice of the densities f considered in these simulations. It is well known that kernel estimators may suffer from boundary effects. Indeed, when there are observations too close to the boundaries of the interval on which the density shall be estimated, then the kernel estimator puts mass beyond the interval boundaries (if no correction is performed). As a consequence, the density is systematically underestimated on the boundaries. To avoid that those boundary effects influence on the calibration of the κ constants, in this section only densities have been considered that vanish (or at least get close to 0) on the estimation interval.

5.4. **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 $\mu = 2$),
- six different distributions: two distributions from calibration study (Laplace $\mathcal{L}(0.5)$ and Beta $\mathcal{B}(5,3)$), two other distributions from the calibration study but with different parameters (normal $\mathcal{N}(10,3)$ and a mixture $\frac{7}{10}\mathcal{N}(5.5,1)+\frac{3}{10}\mathcal{N}(7,0.16)$) and two completely new distributions (exponential $\mathcal{E}(0.5)$ and Gamma $\Gamma(3,0.5)$), see Figure 1 for illustration.

To evaluate the performance of the estimators we proceed 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 and varying μ. 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.

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 pointwise kernel estimators. The global kernel estimators are somewhere in between. The improvement is by far more important for the Laplace distribution than for the exponential distribution, for example.

Impact of the Poisson parameter. In the pile-up model the Poisson parameter μ can be viewed as an indicator for the amount of bias in the model. Indeed, when $N_i = 1$ in (16), there is a single fluorescence lifetime $Y_{i,1}$ and no minimum is taken in (16). So, if $\mathbb{P}(N_i = 1)$ is close to 1, then distribution G of $Z_1, \dots, Z_n$ is approximately F and the amount of bias in the model is very low. Intuitively, in this case the estimation problem should be much easier than in the case where most of the observations Z_i are the minimum of several values implying that G is very different from F, that is, the model contains a lot of bias. The three levels of μ considered here correspond to a very low ($\mu = 0.1$), medium ($\mu = 0.8$) and high ($\mu = 2$) amount of bias. More precisely, the corresponding probabilities that an observation Z_i is the minimum of several random variables (i.e. that the observation Z_i is biased) are given by

$$\mathbb{P}_{\lambda=0.1}(N_i \geq 2) \approx 0.05 \ , \quad \mathbb{P}_{\lambda=0.8}(N_i \geq 2) \approx 0.35 \ , \quad \mathbb{P}_{\lambda=2}(N_i \geq 2) \approx 0.69 \ .$$

As increasing μ results in a more difficult estimation problem, it is natural that the MISE values increase with μ. This effect can be easily observed in Figure 2. However, there are some exceptions where the MISE decreases when μ increases, see for example results for the Beta distribution for $n = 2000$ in Table 1. This phenomenon occurs essentially for kernel estimators and when μ passes from 0.1 to 0.8. In Subsection 5.5 we will come back to this phenomenon.

Comparison of bias-correction approaches: weighted estimators or plug-in strategy? When $\mu = 0.1$ there is almost no bias in the model, as only 5% of the observations are biased data. Consequently, almost no bias correction is necessary and thus no significant difference between the weighted and the plug-in versions of any estimator is observed. However, for larger μ, differences in the MISE appear for the different bias-correction methods, see for instance the projection estimators in the exponential case or pointwise kernel estimators in the Laplace case. In Table 1 weighted estimators outperform their plug-in counterparts 60 times, plug-in versions are better in 39 cases and in the remaining 9 cases they achieve equal MISE values. Moreover, in some cases the plug-in version is really bad, while the weighted estimators seem to produce more robust results. Consequently, preference should be given to weighted estimators. In particular the weighted projection estimator almost always yields better results than its plug-in version.

Comparison of estimation strategies: Global kernel, pointwise kernel or projection strategy? Both pointwise kernel estimators are mostly far behind all other estimators. It seems that the pointwise bandwidth selection fails. This is surprising as the pointwise method conceptually outplays the global one, since it is conceived to capture peaks like in the exponential or Laplace distribution. In our simulation study, only for the exponential distribution the pointwise method achieves similar results to the global one, but still is doing worse. Indeed, in Figure 2 we see that the boxplots associated with the pointwise kernel estimators are much more dispersed than the other ones. Apparently, the method does not succeed to make good local choices. It is possible that much larger sample sizes are necessary for the pointwise method to outperform the global one.

Concerning the difference of the global kernel and the projection estimators, we note that the performance depends on the underlying distribution and that there are significant differences. Clearly, the projection estimator outperforms all other estimators for the Laplace distribution and the normal mixture, while the global kernel estimator is better in the exponential case. In the other cases, both strategies achieve comparable results.

5.5. **Comparison to the oracle.** In the previous simulations we also evaluated the oracles for the different estimators. Here, by oracle we mean the MISE of the best estimator that could have been chosen. More precisely, for the projection estimator $\hat{f}_m^{\text{proj-W}}$, for instance, the oracle for a given dataset is given by $\min_{m\in\mathcal{M}} \|\hat{f}_m^{\text{proj-W}} - f\|^2$. The mean values of these oracles are reported in Table 2. Analyzing the oracles may give a hint on the quality of the (simple) estimators. Furthermore, a comparison with the corresponding MISE values allows us to evaluate the quality of the different data-driven bandwidth and model selection devices.

It is clear that the oracles of the pointwise kernel methods must be better than their global counterparts, as for the pointwise method the best bandwidth is chosen at every point, while the global method selects a single bandwidth that is used for the entire estimation interval. We may conclude that the pointwise bandwidth selection method yields a $\log(n)$-loss in the variance and in the rate, which makes at the end the method less reliable. It is interesting to see that the difference between the oracles of the pointwise and the global kernel methods essentially depends on the type of distribution. While there is a factor 4 between the pointwise and global oracles in the Laplace, mixture and Gamma case, there is only a factor 1.5 in the exponential case.

As the projection estimators also rely on a global selection method for the entire interval, it is natural that their oracles are much worse than the one of the pointwise kernel methods. Among the global selection methods, the weighted global kernel estimator $\hat{f}_{\hat{h}^f}^{\text{ker-W}}$ outperforms the others in the Laplace and the exponential case, whereas the weighted projection estimator $\hat{f}_{\hat{m}^f}^{\text{proj-W}}$ has the best oracles in the Gamma and the normal distribution. In the mixture model, all global estimators achieve competing results, and in the Beta setting both projection estimators outmatch the global kernel methods. This illustrates that the performance of different estimation strategies depends on the density to estimate.

As for the MISE values it sometimes occurs that the oracles diminish when the Poisson parameter passes from 0.1 to 0.8, see e.g. the Laplace or the Beta distribution. This is counter-intuitive as a larger μ value means more data bias. Thus, it is not the bandwidth or model selection device that causes the phenomenon, but it is likely that the problem is inherent to the estimation strategies. As the problem occurs only for small values of μ, it is possible that the bias correction fails in some settings where there is only little amount of bias in the data.

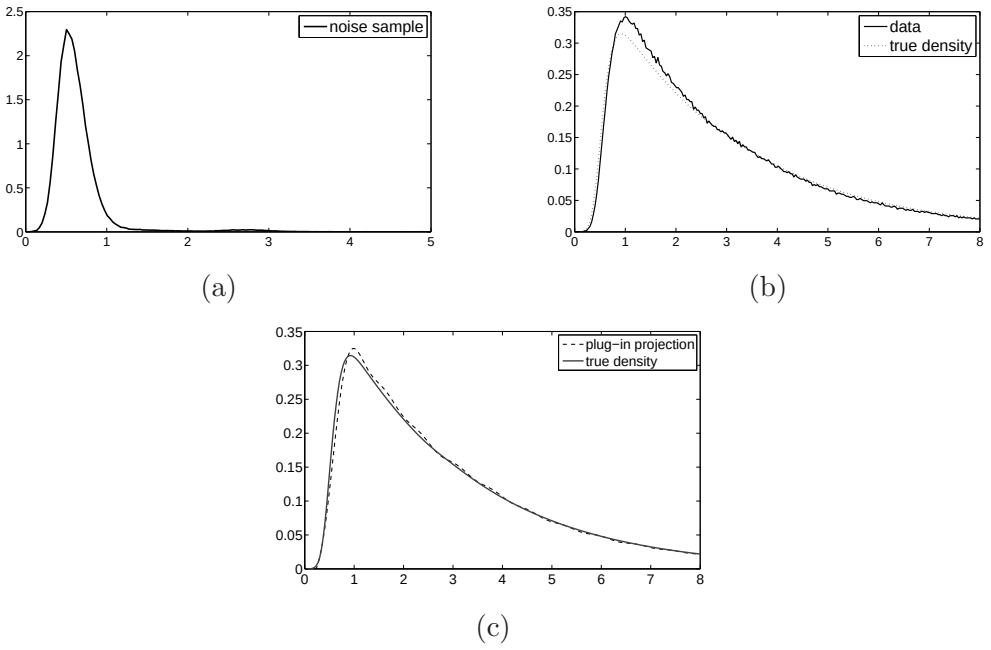
Now it is interesting to analyze the difference of the oracles in Table 2 with the actually achieved MISE values by the adaptive procedures given in Table 1. Obviously, the kernel

TABLE 2. Mean oracles*1000 values for the six different estimators corresponding to the simulation results in Table 1.

	Laplace distribution $\mathcal{L}(0.5)$						Beta distribution $\mathcal{B}(5,3)$					
μ	0.1		0.8		2		0.1		0.8		2	
n	500	2000	500	2000	500	2000	500	2000	500	2000	500	2000
$\hat{f}^{\text{ker-W}}_{\hat{h}^f}$	5.40	1.95	5.29	2.04	6.62	2.53	10.9	4.13	11.2	4.23	12.5	4.90
$\hat{f}^{\text{ker-P}}_{\hat{h}^g}$	5.40	1.95	5.37	2.06	7.55	2.81	10.9	4.13	11.3	4.18	15.4	5.44
$\hat{f}^{\text{ker-W}}_{\hat{h}^f(x)}$	1.50	0.443	**1.43**	**0.477**	**1.73**	**0.569**	**3.92**	**1.29**	4.10	1.34	**4.68**	**1.53**
$\hat{f}^{\text{ker-P}}_{\hat{h}^g(x)}$	**1.49**	**0.442**	1.52	0.506	2.31	0.755	**3.92**	**1.29**	**4.02**	**1.25**	4.97	1.58
$\hat{f}^{\text{proj-W}}_{\hat{m}^f}$	5.98	2.11	5.99	2.15	7.19	2.66	8.03	2.85	7.97	2.93	9.29	3.26
$\hat{f}^{\text{proj-P}}_{\hat{m}^g}$	5.98	2.12	6.11	2.19	8.44	3.11	7.80	2.77	6.22	2.48	11.0	3.48
	Normal distribution $\mathcal{N}(10,3)$						Mixture $\frac{7}{10}\mathcal{N}(5.5,1)+\frac{3}{10}\mathcal{N}(7,0.16)$					
μ	0.1		0.8		2		0.1		0.8		2	
n	500	2000	500	2000	500	2000	500	2000	500	2000	500	2000
$\hat{f}^{\text{ker-W}}_{\hat{h}^f}$	0.985	0.373	1.06	0.374	1.24	0.447	2.70	1.06	2.73	1.09	3.20	1.28
$\hat{f}^{\text{ker-P}}_{\hat{h}^g}$	0.985	0.373	1.11	0.383	1.54	0.555	2.70	1.06	2.71	1.09	3.38	1.30
$\hat{f}^{\text{ker-W}}_{\hat{h}^f(x)}$	**0.444**	**0.127**	**0.481**	**0.127**	**0.598**	**0.162**	**0.806**	**0.273**	**0.835**	**0.292**	**1.05**	**0.360**
$\hat{f}^{\text{ker-P}}_{\hat{h}^g(x)}$	**0.444**	**0.127**	0.490	0.129	0.660	0.182	**0.806**	**0.273**	0.840	0.293	1.06	0.366
$\hat{f}^{\text{proj-W}}_{\hat{m}^f}$	0.655	0.195	0.717	0.191	0.829	0.247	2.78	0.933	2.86	0.979	3.44	1.15
$\hat{f}^{\text{proj-P}}_{\hat{m}^g}$	0.656	0.196	0.822	0.242	1.65	0.656	2.75	0.917	2.77	0.967	4.35	2.37
	Exponential distribution $\mathcal{E}(0.5)$						Gamma distribution $\Gamma(3,0.5)$					
μ	0.1		0.8		2		0.1		0.8		2	
n	500	2000	500	2000	500	2000	500	2000	500	2000	500	2000
$\hat{f}^{\text{ker-W}}_{\hat{h}^f}$	94.5	83.7	97.5	84.8	103	89.0	3.27	1.17	3.46	1.22	4.18	1.51
$\hat{f}^{\text{ker-P}}_{\hat{h}^g}$	95.4	84.2	104	88.6	120	98.8	3.29	1.17	3.64	1.26	5.26	1.69
$\hat{f}^{\text{ker-W}}_{\hat{h}^f(x)}$	**65.4**	**60.5**	**67.9**	**62.1**	**70.2**	**65.1**	**0.849**	**0.291**	**0.854**	**0.290**	**0.938**	**0.315**
$\hat{f}^{\text{ker-P}}_{\hat{h}^g(x)}$	65.8	60.7	71.2	63.7	77.5	68.4	0.857	0.293	0.953	0.315	1.44	0.441
$\hat{f}^{\text{proj-W}}_{\hat{m}^f}$	174	164	175	164	179	166	3.10	1.06	3.25	1.11	4.06	1.36
$\hat{f}^{\text{proj-P}}_{\hat{m}^g}$	188	178	357	350	1054	1050	3.19	1.08	4.08	1.33	7.72	2.43

pointwise estimator is not able to take advantage of its very small oracles. It is evident that the pointwise bandwidth selection fails completely. There is a factor 10 to 20 between the oracles and the corresponding MISE values, increasing with μ. The exponential case is the only exception with a factor 2. For the global kernel estimators the loss is at most a factor 2, and in the exponential, normal and Gamma case the situation is even better. Finally, concerning projection estimators, they even do a bit better than the global kernel methods.

5.6. **Application to real data.** We now apply our statistical methods to real fluorescence lifetime measurements. Here the variables $Y_{i,j}$ are the sum of two quantities, say $Y_{i,j} = F_{i,j} + I_{i,j}$, where $F_{i,j}$ denotes the fluorescence lifetime of a molecule and $I_{i,j}$ is a sort of (random) process time of the signal due to the measuring instrument. Denote f_F and f_I the densities of $F_{i,j}$ and $I_{i,j}$, resp. Then density f, which is to be estimated from the data, is the convolution $f = f_F \otimes f_I$.



(a) (b)

(c)

FIGURE 3. (a) Histogram of instrumental function f_I. (b) Fluorescence data and density f. (c) Best estimator and density f.

The principal advantage of the dataset at hand is that density f is exactly known, such that we are able to evaluate the performance of the different estimators. According to the physicists, the fluorescence lifetimes of the here analyzed specimen are known to be exponentially distributed with mean 2.54 ns. The instrumental function f_I can be observed separately, that is, we dispose of an independent noise sample of size $259,386$ (see Figure 3(a)). That is, f_I is considered to be a known function. The Poisson parameter, which is directly related to the laser intensity, is known to be $\mu = 0.166$. Note that the same dataset has already been analyzed in Comte and Rebafka (2012), but from a deconvolution point of view, that is the aim was the recovery of the exponential density f_F. Here we are interested in the estimation of $f = f_F \otimes f_I$.

Figure 3(b) shows the data in form of a histogram with very fine bins and density f. The sample size is $17,402$. Here we clearly observe the model bias, 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 estimators achieve the following squared errors $\|f - \hat{f}\|^2$:

	$\hat{f}^{\text{ker-W}}_{\hat{h}^f(x_0)}$	$\hat{f}^{\text{ker-P}}_{\hat{h}^g(x_0)}$	$\hat{f}^{\text{ker-W}}_{\hat{h}^f}$	$\hat{f}^{\text{ker-P}}_{\hat{h}^g}$	$\hat{f}^{\text{proj-W}}_{\hat{m}^f}$	$\hat{f}^{\text{proj-P}}_{\hat{m}^g}$
$\|f - \hat{f}\|^2$	$5.18\ 10^{-5}$	$5.30\ 10^{-5}$	$6.18\ 10^{-5}$	$6.74\ 10^{-5}$	$4.48\ 10^{-5}$	$4.43\ 10^{-5}$

We note that the squared error terms are all of the same order and the plug-in projection estimators $\hat{f}^{\text{proj-P}}_{\hat{m}^g}$ yields the best approximation. For illustration, Figure 3(c) displays the plug-in projection estimator $\hat{f}^{\text{proj-P}}_{\hat{m}^g}$and the true density f. We can see that the projection estimator gives a very good recovery of the target density f.

6. Conclusion

We resume the three aims of our work enunciated in the introduction (Section 1).

First of all, we managed to construct nonparametric estimators for the biased data model given by (1), namely projection and kernel estimators both with data-driven model or bandwidth

selection. On a real data example of fluorescence lifetime measurements all estimators achieve very satisfying results.

Second, from a theoretical point of view it is shown that all these estimators are (nearly-)rate optimal. In other words, all procedures minimize the MISE when automatically selecting the model or the bandwidth. However, an extensive simulation study for the pile-up model reveals that the pointwise adaptive kernel estimators fail in practice and should not be used in general. Nevertheless, projection estimators as well as global adaptive kernel estimators achieve very good results in various settings. Furthermore, according to our numerical results, the loss of the adaptation step in comparison to the oracles is rather small for these estimators.

Third, for all projection and kernel estimators the correction of the model bias can be done in two ways: global correction (plug-in estimators) or correction of every datapoint (weighted estimators). The theoretical results hold for both bias-correction methods. Numerical results show that both methods work very well in practice. Weighted estimators slightly tend to do better and seem to be more robust than the corresponding plug-in methods.

The final conclusion of the simulation study is that although the performance of the global kernel estimators is the best in some settings, the weighted projection estimator has an excellent overall performance and should be the method of choice.

7. Appendix

7.1. **Proof of Proposition 3.1.**

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

$$\hat{f}_h^{\text{ker-W}}(x_0) - f(x_0) = \left(\hat{f}_h^{\text{ker-W}}(x_0) - \check{f}_h(x_0)\right) + \left(\check{f}_h(x_0) - \mathbb{E}[\check{f}_h(x_0)]\right) + \left(\mathbb{E}[\check{f}_h(x_0)] - f(x_0)\right) \ . \tag{17}$$

First, we state that by property (5) and with the notation $f_h(x) = K_h * f(x)$, we have

$$\mathbb{E}[\check{f}_h(x_0)] - f(x_0) = \mathbb{E}[K_h(x_0 - Y_i)] - f(x_0) = K_h * f(x_0) - f(x_0) = (f_h(x_0) - f(x_0)) \ . \tag{18}$$

Therefore, the last term in (17) is a standard bias term in kernel density estimation (Tsybakov, 2004). To study the second term of (17), we successively apply property (5), $0 \le w \le 1/a$ and the fact that K is square-integrable to obtain

$$\begin{aligned}
\mathbb{E}\left[\left(\check{f}_h(x_0) - \mathbb{E}(\check{f}_h(x_0))\right)^2\right] &= \frac{1}{n}\mathrm{Var}\left(w(G(Z_1))K_h(x_0 - Z_1)\right) \\
&\le \frac{1}{n}\mathbb{E}\left[\{w(G(Z_1))K_h\left(x_0 - Z_1\right)\}^2\right] \\
&\le \frac{1}{an}\mathbb{E}\left[K_h^2\left(x_0 - Y_1\right)\right] = \frac{1}{anh^2}\int K^2\left(\frac{x_0 - y}{h}\right) f(y)\mathrm{d}y \\
&\le \frac{1}{anh}\|f\|_\infty \|K\|_2^2 \ . \qquad (19)
\end{aligned}$$

For the first term in decomposition (17), the Lipschitz property of w implies that

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

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

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

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

$$\mathbb{E}\left[\left(\hat{f}_h^{\text{ker-W}}(x_0) - \check{f}_h(x_0)\right)^2\right] \leq \frac{5db^2}{4nha^6}\|f\|_\infty \|K\|_2^2\,. \tag{20}$$

Gathering (18), (19) and (20) yields the result. □

Proof of (ii). By (17) it follows that

$$\mathbb{E}\left[\|\hat{f}_h^{\text{ker-W}} - f\|_2^2\right] \leq 3\left(\mathbb{E}\left[\|\hat{f}_h^{\text{ker-W}} - \check{f}_h\|_2^2\right] + \mathbb{E}\left[\|\check{f}_h - \mathbb{E}[\check{f}_h]\|_2^2\right] + \|\mathbb{E}[\check{f}_h] - f\|_2^2\right)\,. \tag{21}$$

Concerning the last term of (21), we get

$$\|\mathbb{E}[\check{f}_h] - f\|_2^2 = \|K_h * f - f\|^2 \tag{22}$$

For the first right-hand-side term of (21) we obtain

$$\begin{aligned}
\mathbb{E}\left[\|\hat{f}_h^{\text{ker-W}} - \check{f}_h\|_2^2\right] &= \int \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\left(w(\hat{G}_n(Z_i)) - w(G(Z_i))\right)K_h(x - Z_i)\right)^2\right]\mathrm{d}x \\
&\leq \int \mathbb{E}\left[\left(w(\hat{G}_n(Z_1)) - w(G(Z_1))\right)^2 K_h^2(x - Z_1)\right]\mathrm{d}x \\
&\leq \frac{c_w^2}{h}\|K\|_2^2\,\mathbb{E}\left[\left(\hat{G}_n(Z_1) - G(Z_1)\right)^2\right] \\
&\leq \frac{b^2}{nha^6}\|K\|_2^2\,, \qquad (23)
\end{aligned}$$

where we used that $\mathbb{E}[(\hat{G}_n(Z_1) - G(Z_1))^2] \leq 1/n$. 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 (21) we obtain

$$\begin{aligned}
\mathbb{E}\left[\|\check{f}_h - \mathbb{E}[\check{f}_h]\|_2^2\right] &= \frac{1}{n}\int \mathbb{E}\left[(w(G(Z_1)))^2 K_h^2\left(x - Z_1\right)\right]\mathrm{d}x \\
&\leq \frac{1}{an}\int\int K_h^2(x-z)\mathrm{d}x(w(G(z)))^2 g(z)\mathrm{d}z = \frac{\|w\|_2^2}{nh}\|K\|_2^2\,. \qquad (24)
\end{aligned}$$

Combining (22), (23) and (24) completes the proof. □

7.2. **Proof of Proposition 3.2.** Pythagoras formula yields $\|f-\hat{f}_m^{\text{proj-W}}\|_2^2 = \|f-f_m\|_2^2+\|f_m-\hat{f}_m^{\text{proj-W}}\|_2^2$. By definition of the orthogonal projection $f_m=\sum_{j=0}^{2m} a_j\varphi_j$ and by using equality (5), we have $a_j=\langle\varphi_j,f\rangle=\mathbb{E}(\varphi_j(Y))=\mathbb{E}(\varphi_j(Z_1)w(G(Z_1)))$. This, together with formula (7) implies that $\|f_m-\hat{f}_m^{\text{proj-W}}\|_2^2=\sum_{j=0}^{2m}(a_j-\hat{a}_j)^2$. If we define

$$\nu_n(h)=\frac{1}{n}\sum_{i=1}^{n}[h(Z_i)w(G(Z_i))-\mathbb{E}(h(Z_i)w(G(Z_i)))]\,, \tag{25}$$

$$R_n(h)=\frac{1}{n}\sum_{i=1}^{n}h(Z_i)[w(\hat{G}_n(Z_i))-w(G(Z_i))]\,, \tag{26}$$

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

$$\begin{aligned}\sum_{j=0}^{2m}\mathbb{E}(\nu_n^2(\varphi_j)) &= \sum_{j=0}^{2m}\frac{1}{n}\mathrm{Var}(\varphi_j(Z_i)w(G(Z_i)))\le\sum_{j=0}^{2m}\frac{1}{n}\mathbb{E}\left[\varphi_j^2(Z_1)(w(G(Z_1)))^2\right]\\ &\le\frac{1}{n}\mathbb{E}\left[\|\sum_{j=0}^{2m}\varphi_j^2\|_\infty(w(G(Z_1)))^2\right]\le\frac{D_m}{n}\mathbb{E}[(w(G(Z_1)))^2]=\|w\|_2^2\frac{D_m}{n}\,,\end{aligned}$$

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

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

with (3) 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 3.2. □

7.3. **Sketch of proof of Theorem 4.1.** In the following, we omit the super index $^{\text{proj-W}}$. It is easy to see that $\hat{f}_m=\arg\min_{t\in S_m}\gamma_n(t)$ for $\gamma_n(t)=\|t\|^2-2n^{-1}\sum_{i=1}^{n}w(\hat{G}_n(Z_i))t(Z_i)$. 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 (25) and (26). By definition of $\hat{f}_{\hat{m}}$ we have for all $m\in\mathcal{M}_n$, $\gamma_n(\hat{f}_{\hat{m}})+\mathrm{pen}^f(\hat{m})\le\gamma_n(f_m)+\mathrm{pen}^f(m)$. This can be rewritten as $\|\hat{f}_{\hat{m}}-f\|_2^2\le\|f_m-f\|_2^2+\mathrm{pen}^f(m)+2\nu_n(\hat{f}_{\hat{m}}-f_m)-\mathrm{pen}^f(\hat{m})+2R_n(\hat{f}_{\hat{m}}-f_m)$. Using this and and that $2xy\le x^2/\theta+\theta y^2$ for all nonnegative x,y,θ, we obtain

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

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

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

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

Then the term $\mathbb{E}\left(\sup_{t\in S_{\hat{m}}+S_m,\|t\|_2=1}[\nu_n(t)]^2 - (\mathrm{pen}^f(m)+\mathrm{pen}^f(\hat{m}))/4\right)_+$ is bounded by C/n by using Talagrand Inequality in a standard way (see e.g. Brunel et al., 2005), as soon as $\kappa_1^f/4 \ge 4$ ($\epsilon = 1/2$ in Lemma (7.2)). 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

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

As in (32), we use Massart (1990) and get

$$\mathbb{P}(\sqrt{n}\|\hat{G}_n - G\|_\infty \ge \lambda) \le 2e^{-2\lambda^2} .$$

This implies that $\mathbb{P}(\Omega_G^c) \le 2/n^2$. 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

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

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

$$\mathbb{E}\left(\sup_{t\in S_{m_n},\|t\|_2=1}|\nu_n'(t^2)|\right) \le C\log(n)$$

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 $m \in \mathcal{M}_n$, $\mathbb{E}[\|f - \hat{f}_{\hat{m}}\|_2^2] \leq 7\|f - f_m\|_2^2 + 8\mathrm{pen}^f(m) + K\log^2(n)/n$, which ends the proof. □

7.4. **Proof of Theorem 4.2.** For the sake of readability, super-indices $^{\text{ker-W}}$ and ^{f} are omitted in the whole proof. For any $h \in \mathcal{H}$,

$$\begin{aligned}
&\left(\hat{f}_{\hat{h}(x_0)}(x_0) - f(x_0)\right)^2 \\
&\qquad \leq 3\left\{\left(\hat{f}_{\hat{h}(x_0)}(x_0) - \hat{f}_{h,\hat{h}(x_0)}(x_0)\right)^2 + \left(\hat{f}_{h,\hat{h}(x_0)}(x_0) - \hat{f}_h(x_0)\right)^2 + \left(\hat{f}_h(x_0) - f(x_0)\right)^2\right\} \\
&\qquad \leq 3\left\{\left(A_0(h,x_0) + V_0(\hat{h}(x_0))\right) + \left(A_0(\hat{h}(x_0),x_0) + V_0(h)\right) + \left(\hat{f}_h(x_0) - f(x_0)\right)^2\right\} \\
&\qquad \leq 6A_0(h,x_0) + 6V_0(h) + 3\left(\hat{f}_h(x_0) - f(x_0)\right)^2 ,
\end{aligned} \tag{27}$$

where the second inequality holds by the definition of A_0, i.e. for all $h, h' \in \mathcal{H}$ we have $A_0(h,x_0) + V_0(h') \geq \left(\hat{f}_{h,h'}(x_0) - \hat{f}_{h'}(x_0)\right)^2$. The last inequality holds by the definition of $\hat{h}(x_0)$, that is $A_0(\hat{h}(x_0),x_0) + V_0(\hat{h}(x_0)) \leq A_0(h,x_0) + V_0(h)$ for all $h \in \mathcal{H}$. The term $\mathbb{E}[(\hat{f}_h(x_0) - f(x_0))^2]$ is controlled by Proposition 3.1. Hence, it is sufficient to study the term $\mathbb{E}[A_0(h,x_0)]$. We state that

$$A_0(h,x_0) = \sup_{h'\in\mathcal{H}}\left[\left(\hat{f}_{h,h'}(x_0) - \hat{f}_{h'}(x_0)\right)^2 - V_0(h')\right]_+ \leq 5\left(D_1 + D_2 + D_3 + D_4 + D_5\right) , \tag{28}$$

where

$$\begin{aligned}
D_1 &= \sup_{h'\in\mathcal{H}}\left(\hat{f}_{h,h'}(x_0) - \check{f}_{h,h'}(x_0)\right)^2 , \quad D_2 = \sup_{h'\in\mathcal{H}}\left[\left(\check{f}_{h,h'}(x_0) - \mathbb{E}\left[\check{f}_{h,h'}(x_0)\right]\right)^2 - \frac{V_0(h')}{10}\right]_+ , \\
D_3 &= \sup_{h'\in\mathcal{H}}\left(\mathbb{E}\left[\check{f}_{h,h'}(x_0)\right] - \mathbb{E}\left[\check{f}_{h'}(x_0)\right]\right)^2 , \quad D_4 = \sup_{h'\in\mathcal{H}}\left[\left(\mathbb{E}\left[\check{f}_{h'}(x_0)\right] - \check{f}_{h'}(x_0)\right)^2 - \frac{V_0(h')}{10}\right]_+ ,
\end{aligned}$$

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 (18). Likewise, by property (5), $\mathbb{E}\left[\check{f}_{h,h'}(x_0)\right] = K_{h'} * K_h * f(x_0)$. In general we have $\|s * r\|_\infty \leq \|s\|_\infty \|r\|_1$ and $\|K_h\|_1 = \|K\|_1$, yielding

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

Hence

$$D_3 \leq \|K_h * f - f\|_\infty^2 \|K\|_1^2 . \tag{29}$$

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 $S_i = w(G(Z_i))K_h(x_0 - Z_i)$. Obviously, $|S_i| \le \|K\|_\infty/(ah) =: M$ almost surely and by property (5)

$$\mathrm{Var}\,(S_i) \le \mathbb{E}\left[w^2(G(Z_1))K_h^2(x_0 - Z_1)\right] \le \frac{1}{h}\|K\|_2^2\|w\|_\infty\|f\|_\infty \le \frac{1}{ah}\|K\|_2^2\|f\|_\infty =: v\,.$$

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

$$\begin{aligned}&\mathbb{P}\left(\left|\check{f}_h(x_0) - \mathbb{E}\left[\check{f}_h(x_0)\right]\right| \ge \sqrt{\frac{V_0(h)}{10} + x}\right) = \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n (S_i - \mathbb{E}[S_i])\right| \ge \sqrt{\frac{V_0(h)}{10} + x}\right) \\ &\le 2\max\left\{\exp\left(-\frac{n}{4v}\left(\frac{V_0(h)}{10} + x\right)\right), \exp\left(-\frac{n}{8M}\sqrt{\frac{V_0(h)}{10}}\right)\exp\left(-\frac{n}{8M}\sqrt{x}\right)\right\}\,.\end{aligned}$$

By the definition of $V_0(h)$

$$\frac{n}{4v}\frac{V_0(h)}{10} = \frac{\kappa_2\|K\|_1^2\log n}{40} \ge p\log n\,,$$

for $\kappa_2 \ge 40p$, since $\|K\|_1^2 \ge 1$. Furthermore,

$$\frac{n}{8M}\sqrt{\frac{V_0(h)}{10}} = \frac{\|K\|_2\|K\|_1}{8\|K\|_\infty}\sqrt{\frac{\kappa_2 a\|f\|_\infty hn\log n}{10}} := \rho\sqrt{\kappa_2 hn\log n}.$$

As $nh \ge \log^2(n)$, we get $(n/8M)\sqrt{V_0(h)/10} \ge \log n$ if $\rho^2\kappa_2 log(n) \ge p$, which holds automatically for n large enough, and thus for a well chosen κ_2. Then we get

$$\begin{aligned}\mathbb{E}[D_4] &\le \sum_{h\in\mathcal{H}} \int_0^\infty 2n^{-p}\max\left\{\exp\left(-\frac{nhax}{4\|K\|_2^2\|f\|_\infty}\right), \exp\left(-\frac{nha\sqrt{x}}{8\|K\|_\infty}\right)\right\}\mathrm{d}x \\ &\le 2n^{-p}\sum_{h\in\mathcal{H}}\int_0^\infty \max\left\{\mathrm{e}^{-\tau_1 nhx}, \mathrm{e}^{-\tau_2 nh\sqrt{x}}\right\}\mathrm{d}x \le 2n^{-p}\sum_{h\in\mathcal{H}}\max\left\{\frac{1}{\tau_1}, \frac{2}{\tau_2^2}\right\} \le C' n^{-p+1}\,,\end{aligned}$$

as $h \ge 1/n$ and the cardinality of $\mathcal{H}$ verifies $\#\mathcal{H} \le n$. Finally, we choose $p = 2$ (and thus $\kappa_2 \ge 80$) to get

$$\mathbb{E}[D_4] \le \frac{C'}{n}\,. \tag{30}$$

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

$$\mathbb{E}[D_2] \leq \frac{C''}{n} \,. \tag{31}$$

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

Lemma 7.1. *Under the assumptions of Theorem 4.2, for any set Ω and for all $t \in \mathbb{R}$,*

$$\mathbb{E}\left[\left(\hat{f}_h(t) - \check{f}_h(t)\right)^2 \mathbb{1}_{\Omega^c}\right] \leq c_w^2 \|K\|_\infty^2 n^2 \mathbb{P}(\Omega^c) \quad \textit{and}$$

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

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

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

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)),

$$\mathbb{P}(\Omega^c) = \mathbb{P}(\|\hat{G}_n - G\|_\infty > s) \leq \mathrm{e}^{-2ns^2} \,, \tag{32}$$

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,

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

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 (13) and d in (4). 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

$$\mathbb{E}[D_5]\leq\frac{\mu_1}{n}+2\mu_2\frac{\log n}{n}\ . \tag{33}$$

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,

$$\mathbb{E}[D_1]\leq\frac{\tilde{\mu}_1}{n}+2\tilde{\mu}_2\frac{\log n}{n}\ , \tag{34}$$

with $s^2=2\log n/n$. Now, if we plug (29), (30), (31), (33) and (34) into (28), 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 3.1, can be inserted in (27) to end the proof of Theorem 4.2. □

7.5. **Proof of Theorem 4.3.** In all the proof below, super-indices $^{\text{ker-W}}$ are omitted. Similar to the pointwise case, we have for any $h\in\mathcal{H}$

$$\|\hat{f}_{\hat{h}}-f\|_2^2\leq 6A(h)+6V(h)+3\|\hat{f}_h-f\|_2^2\ . \tag{35}$$

By the proof of (ii) of Proposition 3.1,

$$\mathbb{E}\left[\|\hat{f}_h-f\|_2^2\right]\leq 3\|f_h-f\|_2^2+\frac{C_4}{nh}\ . \tag{36}$$

Hence, only term $\mathbb{E}[A(h)]$ needs to be studied. By analogy to the proof of Theorem 4.2,

$$A(h)\leq 5(F_1+F_2+F_3+F_4+F_5)\ , \tag{37}$$

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

$$F_3 = \sup_{h'\in\mathcal{H}} \|K_{h'} * K_h * f - K_{h'} * f\|_2^2 \leq \sup_{h'\in\mathcal{H}} \|K_{h'}\|_1^2 \|K_h * f - f\|_2^2 = \|K\|_1^2 \|f - f_h\|_2^2 \,. \tag{38}$$

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

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

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 7.6). More precisely, to apply this result, we have to determine the values of the constants H, M and v. Denote $f_\psi(z) = w(G(z)) K_h * \psi(z)$, 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 (24). 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 (5)

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

Finally, Talagrand's inequality yields, for $\kappa_3^f/10 \geq 4$,

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

where $\tilde{C}_k > 0, k = 1, \ldots, 4$ are constants depending on K, $\|f\|_\infty$ and a, and in particular $\tilde{C}_2 = \|K\|_2^2/(12\|f\|_\infty)$. Consequently, for $\kappa_3^f \geq \kappa_0$ (with here $\kappa_0 = 40$),

$$\mathbb{E}[F_4] \leq \frac{\tilde{C}_1}{n} \sum_{h \in \mathcal{H}} \left(\mathrm{e}^{-\tilde{C}_2/h} + \frac{\tilde{C}_4}{n} \right) \leq \frac{\tilde{C}_5}{n} , \tag{39}$$

as $\#\mathcal{H} \leq n$ and $\sum_{h \in \mathcal{H}} \mathrm{e}^{-\tilde{C}_2/h} \leq B(C_2')$ under condition (15).

In the same way we obtain for F_2

$$\mathbb{E}[F_2] \leq \frac{\bar{C}}{n} . \tag{40}$$

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

$$\mathbb{E}[F_5] \leq \frac{C_1'}{n} + C_2' \frac{\log n}{n} . \tag{41}$$

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

$$\mathbb{E}[F_1] \leq \|K\|_1^2 \left(\frac{C_1'}{n} + C_2' \frac{\log n}{n} \right) . \tag{42}$$

Consequently, plugging (38), (39), (40), (41) and (42) into (37) gives a bound of $\mathbb{E}[A(h)]$. Combining this with (36), (35) and the definition of $V(h)$ yields Theorem 4.3. □

7.6. **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 7.2. *(Talagrand's inequality) Let $Y_1, \ldots, Y_n$ be independent random variables, let $\nu_{n,Y}(f) = n^{-1} \sum_{i=1}^n [f(Y_i) - \mathbb{E}(f(Y_i))]$ and let $\mathcal{F}$ be a countable class of uniformly bounded measurable functions. Then for $\epsilon^2 > 0$*

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

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.

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

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

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

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.

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

Comte, F. and Rebafka, T. (2012). Adaptive density estimation in the pile-up model involving measurement errors. *Electronic Journal of Statistics*, 6:2002–2037.

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.

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

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

Lacour, C. and Massart, P. (2015). Minimal penalty for goldenshluger-lepski method. *arXiv:1503:00946*, page 17.

Lakowicz, J. R. (1999). *Principles of Fluorescence Spectroscopy*. Academic/Plenum, New York.

Lerasle, M. (2012). Optimal model selection in density estimation. *Ann. Inst. Henri Poincaré Probab. Stat.*, 48(3):884–908.

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.

Navarro, F., Chesneau, C., and Fadili, J. (2015). On adaptive wavelet estimation of a class of weighted densities. *Communications in Statistics - Simulation and Computation*, 44(8):2137–2150.

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.

Valeur, B. (2002). *Molecular Fluorescence*. Wiley-VCH, Weinheim.

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.

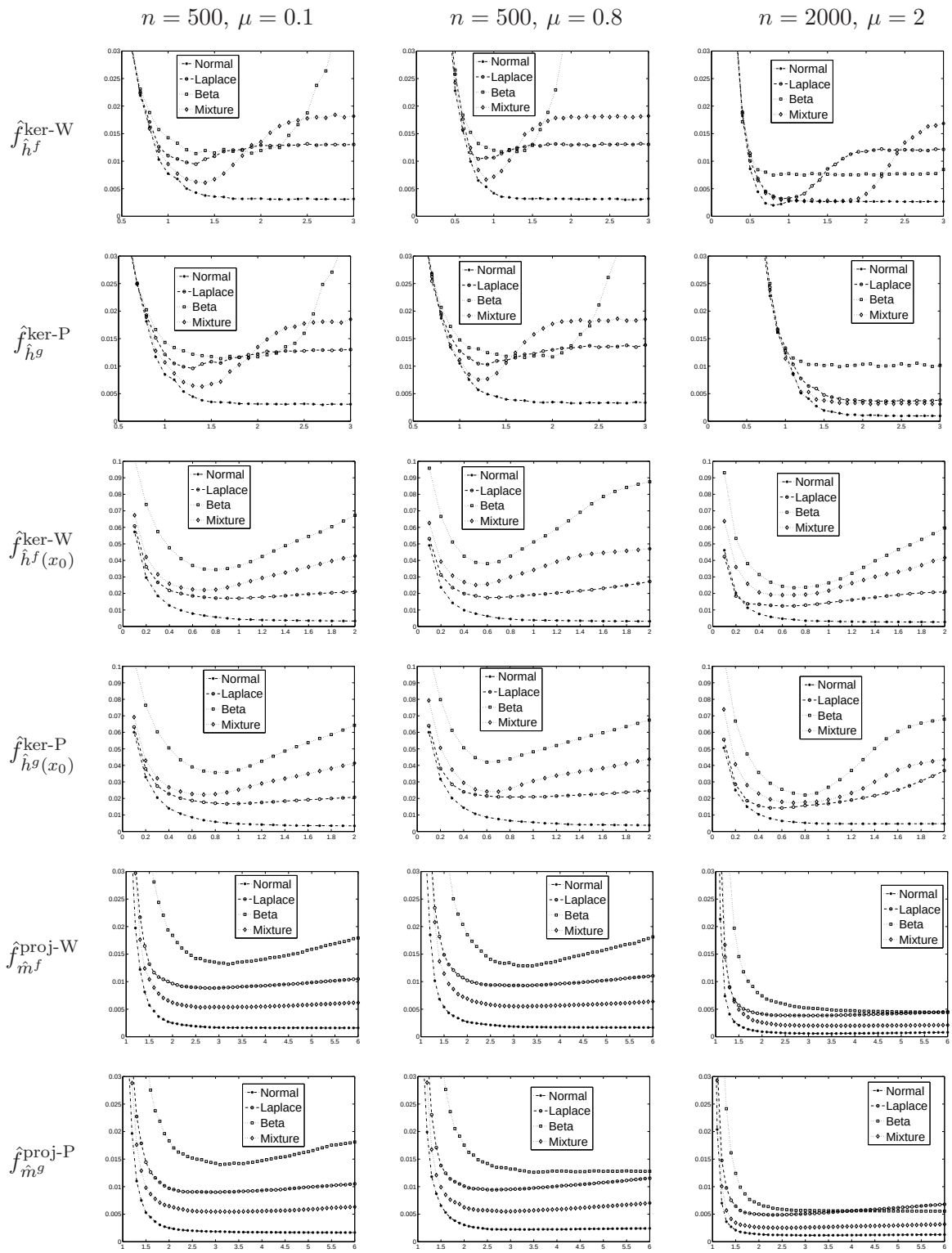


FIGURE 4. Mean MISE_κ values for all estimators on grids of candidate values for the constants $\kappa^f_j, \kappa^g_j, j = 1, 2, 3$ that are to be calibrated. Every row corresponds to the results of one estimator, every column to different values for n and μ and in each setting four different densities f are considered. Each curve is based on 1000 datasets.