Parametric versus nonparametric: The fitness coefficient

François Portier, Gildas Mazo

Received: 20 February 2019 | Revised: 14 February 2020 | Accepted: 8 September 2020

DOI: 10.1111/sjos.12495

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

Parametric versus nonparametric: The fitness coefficient

Gildas Mazo[1] | **François Portier**[2]

[1]MaIAGE, INRAE, Université Paris Saclay, France

[2]LTCI, Télécom ParisTech, University of Paris-Saclay, France

Correspondence
Gildas Mazo, MaIAGE, INRAE,
Université Paris Saclay, 78350
Jouy-en-Josas, France.
Email: gildas.mazo@inrae.fr

Abstract

Olkin and Spiegelman introduced a semiparametric estimator of the density defined as a mixture between the maximum likelihood estimator and the kernel density estimator. Due to the absence of any leave-one-out strategy and the hardness of estimating the Kullback–Leibler loss of kernel density estimate, their approach produces unsatisfactory results. This article investigates an alternative approach in which only the kernel density estimate is modified. From a theoretical perspective, the estimated mixture parameter is shown to converge in probability to one if the parametric model is true and to zero otherwise. From a practical perspective, the utility of the approach is illustrated on real and simulated data sets.

KEYWORDS

density estimation, goodness-of-fit, kernel smoothing, likelihood methods, semiparametric statistics

1 | INTRODUCTION

There exist two general approaches to density estimation. On the one hand, parametric methods are known to be precise but their results depend on the assumed statistical model. On the other hand, nonparametric methods often require more data but are free from model specification. In this context, Olkin and Spiegelman (1987), abbreviated OS in the present article, proposed to combine the two approaches by forming a convex combination $\alpha f_{\hat{\theta}_n} + (1 - \alpha)\hat{f}_n$ between a parametric density estimator $f_{\hat{\theta}_n}$ and a kernel density estimator $\hat{f}_n$. Given a set of data $X_1, \dots, X_n \subset \mathbb{R}^d$ where $d \geq 1$, the value of α is estimated by $\hat{\alpha}_n^{\mathrm{OS}}$, defined as the argmax over $\alpha \in [0,1]$ of the function

$$\sum_{i=1}^{n} \log(\alpha f_{\hat{\theta}_n}(X_i) + (1-\alpha)\hat{f}_n(X_i)).$$

This permits to get a density estimator robust to misspecification while retaining a performance comparable to parametric estimators when the true density is close to the model.

While the previous idea is quite seducing, it has little chance to succeed because of (i) a prohibitive bias caused by the absence of any leave-one-out estimation strategy and (ii) the inherent flaws of the kernel density estimator in Kullback–Leibler loss estimation (Hall, 1987). In practice, it is very sensitive to the choice of the bandwidth (Faraway, 1990; Rahman, Beaver, & Gokhale, 1997). This prevents the OS method from getting satisfactory theoretical and practical results, as we shall argue throughout this article.

In this article, we “repair” the kernel density estimator in order to improve the quality estimation of α. We introduce the estimate

$$\hat{\alpha}_n \in \underset{\alpha \in [0,1]}{\operatorname{argmax}} \sum_{i=1}^{n} \log(\alpha f_{\hat{\theta}_n}(X_i) + (1-\alpha)\hat{f}_{n,i}^{\mathrm{LR}}), \tag{1}$$

where $\hat{f}_{n,i}^{\mathrm{LR}}$ is called the *leave-and-repair* (LR) kernel density estimate and is given by

$$\hat{f}_{n,i}^{\mathrm{LR}} = \left(\frac{1}{(n-1)h_n^d} \sum_{j \neq i} K\left(\frac{X_i - X_j}{h_n} \right) \right) + \Delta_n q(X_i), \tag{2}$$

where $K : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ is the kernel, $h_n > 0$ is the bandwidth, $\Delta_n \geq 0$ and $q : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$. The LR estimate is a modification of the well-known *leave-one-out* (LOO) estimate usually employed in cross-validation procedures (Hall, 1987) and semiparametric estimation (Delecroix, Hristache, & Patilea, 2006). The estimator $\hat{\alpha}_n$ introduced in (1) is called the *fitness coefficient* because it may be interpreted as how well the model fits the data (see Section 5).

1.1 | Main contributions.

Under mild conditions, the fitness coefficient $\hat{\alpha}_n$ is shown to converge in probability to one if the model is true and zero otherwise, a property called *consistency*. Even if the fitness coefficient is maximizing some objective function (over $\alpha \in [0,1]$), the results from M-estimation theory do not apply directly because, when the model is true, the limiting objective function is independent from α. The proof follows from a fine comparison between the rates of convergence of $f_{\hat{\theta}_n}$ and $\hat{f}_n$. Using some real data as well as extensive simulations, we observe that the LR approach is more stable than the OS approach and leads to more accurate inference. Finally, some new perspectives on the use of the fitness coefficient for model evaluation are provided.

1.2 | Alternative approaches.

Following the idea of combining parametric and nonparametric estimator, some authors (Lee & Soleymani, 2015; Rahman et al., 1997; Soleymani & Lee, 2014) investigate different strategies based on the mean squared error between the combination $\alpha f_{\hat{\theta}_n} + (1-\alpha)\hat{f}_n$ and the true density, but then the solution depends on the unknown distribution and hence heavy bootstrap methods need to be employed. There exist also other approaches than that of forming a convex combination between the parametric and nonparametric estimators. The locally parametric nonparametric estimation is developed for instance in Hjort and Jones (1996), Hjort, McKeague, and Van Keilegom (2018), and Talamakrouni, El Ghouch, and Van Keilegom (2017), but is less appealing from the point of view of model quality assessment because they do not provide any "fitness coefficient."

1.3 | Outline.

In Section 2, we motivate the use of the LR estimator $\hat{f}_{i,n}^{\mathrm{LR}}$ to compute the fitness coefficient $\hat{\alpha}_n$. The consistency of the fitness coefficient is stated in Section 3 where some examples are given. In Section 4, numerical experiments are designed to measure the robustness of the fitness coefficient and the performance of the corresponding density estimators. Finally, some concluding remarks on the potential use of the fitness coefficient are given in Section 5. All the proofs are given in the Appendix.

2 | THE LEAVE-AND-REPAIR ESTIMATOR

The aim of this section is to motivate the use of the LR estimator $\hat{f}_n^{\mathrm{LR}}$ in the definition of the fitness coefficient $\hat{\alpha}_n$.

Let $(X_i)_{i\geq 1}$ be an independent and identically distributed sequence of $\mathbb{R}^d$-valued random variables having density f_0 with respect to the Lebesgue measure. The kernel density estimator of f_0 at $x \in \mathbb{R}^d$ is given by

$$\hat{f}_n(x) = \frac{1}{nh_n^d}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right).$$

For any $h > 0$, define the function f_h as the convolution product between $K_h(\cdot) = K(\cdot/h)/h^d$ and f_0, that is, $f_h(x) = (K_h \star f_0)(x)$, $x \in \mathbb{R}^d$. Note that $f_{h_n}(x) = \mathbb{E}[\hat{f}_n(x)]$. But since $\mathbb{E}[\hat{f}_n(X_i)|X_i] = f_{h_n}(X_i) + K(0)/(nh_n^d)$, we see that $\hat{f}_n(X_i)$ has a positive h_n-dependent bias when estimating $f_{h_n}(X_i)$, conditionally on X_i. When studying the estimator decomposition, this bias term spreads to the diagonal terms of some U-statistics and gives rise, in the end, to some nonnegligible terms. This phenomenon is common in semiparametric statistics, and has been noticed for instance in remark 4 in Portier and Segers (2018).

To overcome the undesirable effects caused by this bias term, the *leave-one-out* (LOO) estimator of $f_{h_n}(X_i)$, given by

$$\hat{f}_{n,i}^{\mathrm{LOO}} = \frac{1}{(n-1)h_n^d}\sum_{j\neq i} K\left(\frac{X_i - X_j}{h_n}\right),$$

has been successfully used in several cross-validation procedures aiming at selecting the bandwidth, either based on the likelihood (Habbema, Hermans, & Van Den Broeck, 1974; Hall, 1987;

Marron, 1985) or on the mean squared error (Rudemo, 1982; Stone, 1984) (see Marron (1987) for a comparison). Since then, LOO estimators have been frequently used in semiparametric studies (Delecroix et al., 2006).

The LR estimator proposed in this article is inspired, but different, from the LOO estimator. In view of (2), the LR estimator satisfies $\hat{f}_{n,i}^{\mathrm{LR}} = \hat{f}_{n,i}^{\mathrm{LOO}} + \Delta_n q(X_i)$. If $\Delta_n = 0$ the LR estimator is equal to the LOO estimator. If $q = K(0)$ and $\Delta_n = 1/((n-1)h_n^d)$ the LR estimator is equal to $(n/(n-1))\hat{f}_n(X_i)$. In this article, Δ_n shall be typically of order $1/n$ and q shall be a density satisfying $q(X_1) > 0$ almost surely.

The heuristic for using the LR estimator $\hat{f}_{n,i}^{\mathrm{LR}}$ instead of the LOO estimator $\hat{f}_{n,i}^{\mathrm{LOO}}$ is as follows. It is well known that the Kullback–Leibler divergence between kernel density estimates and the true density depends crucially on the tails of the true distribution f_0 (Hall, 1987; Schuster & Gregory, 1981). As shown in Hall (1987), if the tail is too heavy and the kernel $K(x)$ vanishes too quickly as x becomes large then the associated Kullback–Leibler divergence diverges to minus infinity. This is because some of the $\hat{f}_{n,i}^{\mathrm{LOO}}$, $i = 1, \ldots, n$, might have very small values (possibly zero), leading to very large values (possibly infinite) for some of the $\log(\hat{f}_{n,i}^{\mathrm{LOO}})$. We built the LR estimator to overcome this issue by simply adding $\Delta_n q(X_i)$ to the LOO estimator. We coined the term *leave-and-repair* because the term $\Delta_n q(X_i)$ *repairs* the LOO estimator. Since $\hat{f}_{n,i}^{\mathrm{LR}} \geq \Delta_n q(X_i)$, the LR estimator is not subjected to the difficulties of the LOO estimator. By adding the term $\Delta_n q(X_i)$ in (2), however, a bias is introduced: now one has $E[\hat{f}_{n,i}^{\mathrm{LR}}|X_i] - f_{h_n}(X_i) = \Delta_n q(X_i)$. Thus, there is a bias-variance trade-off controlled by the sequence Δ_n that must go to zero slowly enough to keep $\hat{f}_{n,i}^{\mathrm{LR}}$ away from zero but also fast enough to keep the bias as small as possible. The right compromise is given in the conditions in Theorem 1 (for instance $\Delta_n = 1/n$ is one possibility).

Let $\mathcal{P} = \{f_\theta \; : \; \theta \in \Theta\}$ be the parametric model where $\Theta \subset \mathbb{R}^p$ is such that for each $\theta \in \Theta, f_\theta \; : \; \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ and $\int f_\theta(x) \, dx = 1$. The maximum likelihood estimator of f_0 based on $\mathcal{P}$ and $X_1, \ldots, X_n$ is $f_{\hat{\theta}_n}$ where $\hat{\theta}_n$ (when it exists; this is further assumed) is defined as

$$\hat{\theta}_n \in \underset{\theta \in \Theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log(f_\theta(X_i)).$$

To conclude the section, we consider existence and uniqueness of the fitness coefficient $\hat{\alpha}_n$. The existence follows from the use of the LR estimator $\hat{f}_{n,i}^{\mathrm{LR}}$. Uniqueness of $\hat{\alpha}_n$ is obtained under the mild requirement that the parametric and nonparametric estimators are distinguishable on the observed data.

Proposition 1. *Suppose that $\Delta_n > 0$ and $q(X_1) > 0$ a.s. and $f_{\hat{\theta}_n}(X_i) \neq \hat{f}_{n,i}^{\mathrm{LR}}$ for at least one $i \in \{1, \ldots, n\}$. Then the fitness coefficient exists and is unique.*

The proof is given in Section A of the Appendix.

3 | CONSISTENCY OF THE FITNESS COEFFICIENT

3.1 | Assumptions and main result

Let $\|\cdot\|_2$ be the Euclidean norm and for any set $S \subset \mathbb{R}^d$ and any function $f \; : \; S \to \mathbb{R}$, define the sup-norm as $\|f\|_S = \sup_{x \in S} |f(x)|$. Denote by λ the Lebesgue measure on $\mathbb{R}^d$. Introduce the density level sets $S_t = \{x \in \mathbb{R}^d \; : \; f_0(x) > t\}$, $t \geq 0$. We shall assume the following.

(H1) The density f_0 is bounded and continuous on $\mathbb{R}^d$ and the gradient ∇f_0 of f_0 is bounded on $\mathbb{R}^d$, and satisfies, for every $x \in \mathbb{R}^d$ and $u \in [-1, 1]^d$,

$$|f_0(x+u) - f_0(x) - u^T \nabla f_0(x)| \leq \|u\|_2^2 g(x),$$

where g is positive, bounded, integrable and $\int g(x)^2/f_0(x)\, \mathrm{d}x < \infty$.

(H2) The kernel function $K : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ integrates to 1 and takes one of the two following forms,

$$(a) \quad K(x) \propto K^{(0)}(\|x\|_2) \quad \text{or} \quad (b) \quad K(x) \propto \prod_{k=1}^{d} K^{(0)}(|x_k|),$$

where $K^{(0)} : [0, 1] \to \mathbb{R}_{\geq 0}$ is of bounded variation. The sequence $(h_n)_{n \geq 1}$ is such that $nh_n^{2d+4} \to 0$, $nh_n^d / |\log(h_n)| \to \infty$.

Whereas (H1) and (H2) are rather classical in the kernel smoothing literature (see the remarks just below Theorem 1), the following assumption is specific to our approach. We shall see in Section 3.2 that this assumption is satisfied for densities with classical tails.

(H3) The function $q : \mathbb{R}^d \mapsto \mathbb{R}_{\geq 0}$ is bounded, integrable, and $\mathbb{E}[|\log(q(X_1))|] < \infty$. There exist $\beta \in (0, 1]$ and $c > 0$ such that $\int_{S_t^c} f_0(x)\, \mathrm{d}x \leq ct^\beta$ as $t \to 0$. For any $\gamma > 0$, $b_n = \gamma(nh_n^d)^{-1/\beta}$, there exists $C > 0$ such that, as $n \to \infty$,

$$\sup_{x \in S_{b_n}} \sup_{u \in [-1,1]^d} \frac{f_0(x + h_n u)}{f_0(x)} \leq C \quad \text{and} \quad h_n^d \lambda(S_{b_n}) \to 0.$$

For the sake of clarity, the assumptions dealing with the parametric model, namely (A1) and (A2), are postponed to Section C of the Appendix. They are taken from the monographs van der Vaart (1998) and Newey and McFadden (1994), and they mainly ensure the asymptotic normality of $\hat{\theta}_n$ whenever $f_0 \in \mathcal{P}$.

Theorem 1. *Suppose that assumptions (H1) to (H3), and (A1) are fulfilled.*

(i) When $f_0 \in \mathcal{P}$, under (A2) and if $(nh_n^d)\Delta_n \to 0$, it holds that $\hat{\alpha}_n \to 1$, in probability.

(ii) When $f_0 \notin \mathcal{P}$, if $\Delta_n \to 0$ and $\left(\sqrt{|\log(h_n)|/nh_n^d + h_n^2}\right)|\log(\Delta_n)|^{1/\beta} \to 0$, we have that $\hat{\alpha}_n \to 0$, in probability.

Section B of the Appendix is dedicated to the proof of Theorem 1. We did not follow the approach used in Olkin and Spiegelman (1987), which, we believe, is unsatisfactory because it does not consider the case when $\hat{\alpha}_n$ lies in the border of $[0, 1]$. Actually, this is not straightforwardly remedied as the event $\hat{\alpha}_n = 0$ or $\hat{\alpha}_n = 1$ has a nonnegligible probability (as illustrated in the numerical experiments in Section 4.2). The smoothness assumption stated in (H1) and the symmetries in the kernel function ensure a control of order h_n^2 of the bias $f_h(x) - f(x)$, uniformly in $x \in \mathbb{R}^d$ (see Lemma 5 stated in Section B.5 of the Appendix). Such a rate could be improved by using higher order kernels but this is not necessary here. Assumption (H2), (a) and (b), are borrowed from the empirical process literature; see among others (Einmahl & Mason, 2000; Giné & Guillou, 2002; Nolan & Pollard, 1987). They permit to bound, uniformly in $x \in \mathbb{R}^d$, the variance term $\hat{f}_n(x) - f_h(x)$. The fact that the kernel has a compact support can be alleviated at the price of additional technicalities in the proof and assuming that the tails of the kernel are light enough. We did not include this analysis in the article for reasons of clarity.

For any dimension $d \geq 1$, there exists a couple of sequences $(h_n, \Delta_n)_{n\geq 1}$ that fulfills the restrictions (i), (ii) of Theorem 1, and (H2). For instance, the bandwidth $h_n \propto n^{-1/(d+4)}$ and $\Delta_n = n^{-r}$ with $r \geq 1$ are such sequences. An interesting point in Theorem 1 is the two opposite roles played by the sequence Δ_n in (i) and (ii), respectively. The consistency when $f_0 \in \mathcal{P}$ requires Δ_n to be as small as possible, whereas when $f_0 \notin \mathcal{P}$, Δ_n must not be too close to 0. In the proof, the two cases $\Delta_n = 0$ (leave-one-out) as well as $\Delta_n q(X_i) = K(0)/(nh^d)$ (OS method) need to be excluded. However, as suggested by the assumptions on Δ_n, its rate of convergence to 0 might be very fast as it only appears as a logarithmic factor.

The fitness coefficient can be used to improve upon the parametric and the nonparametric estimators. Define the mixture estimate

$$\tilde{f}_n(x) = \hat{\alpha}_n f_{\hat{\theta}_n}(x) + (1 - \hat{\alpha}_n)\hat{f}_n(x).$$

We have the following uniform consistency result which does not depend on the validity of the parametric model.

Proposition 2. *Suppose that assumptions (H1) to (H3), and (A1) are fulfilled and that* $\sup_{\theta\in\Theta} \|f_\theta\|_{\mathbb{R}^d} < \infty$. *If* $\Delta_n \to 0$ *and* $(\sqrt{|\log(h_n)|/nh_n^d} + h_n^2)|\log(\Delta_n)|^{1/\beta} \to 0$, *we have*

$$\|\tilde{f}_n - f_0\|_{\mathbb{R}^d} \to 0, \quad \text{in probability.}$$

3.2 | Distributions and bandwidths satisfying (H3)

For densities f_0 with unbounded supports, the verification of Assumption (H3) only depends on some tail function g_0 associated to the density f_0. The meaning of this is made precise in the following proposition.

Proposition 3. *Suppose that for any* $A > 0$, $\inf_{\|x\|\leq A} f_0(x) > 0$ *and that there exists a function* g_0 *such that* $f_0(x)/g_0(x) \to 1$ *as* $\|x\| \to \infty$. *Suppose that* $h_n \to 0$ *and* $nh_n^d \to \infty$. *If there exist* $c_2 > 0$ *and* $\beta \in (0, 1]$ *such that*

$$\int_{g_0(x)\leq t} g_0(x)\,\mathrm{d}x \leq c_2 t^\beta, \quad \text{as } t \to 0,$$

and if for any $\gamma > 0$, $b_n = \gamma(nh_n^d)^{-1/\beta}$, *there exists* $A > 0$, $C_2 > 0$ *such that*

$$\sup_{\|x\|>A,\ g_0(x)>b_n} \sup_{u\in[-1,1]} \frac{g_0(x + h_n u)}{g_0(x)} \leq C_2 \quad \text{and} \quad h_n^d \lambda(g_0(x) > b_n) \to 0, \tag{3}$$

as $n \to \infty$, *then (H3) is valid for* f_0 *with the same value of* β.

The proof of Proposition 3 is given in Section A of the Appendix. The function g_0 in Proposition 3, not necessarily a proper density function, represents the rate of decrease of $f_0(x)$ as $\|x\| \to \infty$.

Example 1 (Mixture of densities). Let $d = 1$. Let $f_0(x) = \pi_1 f_1(x) + \pi_2 f_2(x)$, $\pi_1 > 0$, $\pi_2 > 0$, $\pi_1 + \pi_2 = 1$, where f_1 and f_2 are densities such that $f_1(x)/f_2(x) \to 0$ as $|x| \to \infty$. Take $g_0(x) = \pi_2 f_2(x)$. Then, as $|x| \to \infty$,

$$\frac{f_0(x)}{g_0(x)} = \frac{\pi_1 f_1(x)}{\pi_2 f_2(x)} + 1 \to 1.$$

Hence the verification of (H3) by f_0 only depends on the component f_2.

Putting $g_0 \propto f_0$ (the symbol $\propto$ stands for proportionality) in Proposition 3 amounts to check (H3) directly, which is done in the following examples.

Example 2 (Gaussian tails). Let $d = 1$ and $g_0(x) = \kappa_1 \exp(-\kappa_2 x^2)$, with $\kappa_1 > 0$, $\kappa_2 > 0$. For clarity the computations are provided for $\kappa_1 = \kappa_2 = 1$ but can easily be extended for arbitrary values. As $\int_{\exp(-x^2) \le t} \exp(-x^2)\,\mathrm{d}x \le t$, as $t \to 0$, we have that $\beta = 1$. Moreover, for $0 < b_n < 1$, we have

$$\begin{aligned}\sup_{\exp(-x^2) > b_n} \sup_{u \in [-1,1]} \exp(-(x + h_n u)^2 + x^2) &\le \sup_{\exp(-x^2) > b_n} \sup_{u \in [-1,1]} \exp(-2h_n x u) \\ &= \sup_{\exp(-x^2) > b_n} \exp(2h_n |x|) \\ &\le \exp(2\sqrt{-\log(b_n)} h_n).\end{aligned}$$

Therefore, a sufficient condition on h_n guaranteeing (3) is that $h_n^2 \log(n) \to 0$, which is satisfied under (H2).

Example 3 (Exponential tails). Let $d = 1$ and $g_0(x) = \kappa_1 \exp(-\kappa_2 x)$, with $\kappa_1 > 0$, $\kappa_2 > 0$. The computations are very similar to the one presented in the Gaussian case. We find $\beta = 1$ and the condition on h_n becomes $h_n \log(n) \to 0$ which is always true under (H2). Hence, as for Gaussian tails, when the tails are exponential, (H3) is automatically satisfied under (H2).

Example 4 (Polynomial tails). Let $d = 1$ and $g_0(x) = \kappa_1 |x|^{-k}$ with $\kappa_1 > 0$, $k > 1$. For simplicity, as in the Gaussian example, we focus on $\kappa_1 = 1$. We find that $\beta = (k-1)/k$. For $h_n < |A|$, we have

$$\begin{aligned}\sup_{|x| > A,\ |x| \le b_n^{-1/k}} \sup_{u \in [-1,1]^d} \frac{|x|^k}{|x + h_n u|^k} &= \sup_{|x| > A,\ |x| \le b_n^{-1/k}} \frac{|x|^k}{(|x| - h_n)^k} \\ &= \frac{1}{(1 - h_n/A)^k} \overset{n \to \infty}{\to} 1.\end{aligned}$$

Finally, since $\lambda(g_0 > b_n) = 2b_n^{-1/k} = 2\gamma^{-1/k}(nh_n)^{1/(\beta k)}$, a sufficient condition on h_n guaranteeing (3) is that $nh_n^k \to 0$.

The three examples considered above are informative on the interplay between the tails of f_0 and the choice of h_n. For distributions with light enough tails, including Gaussian, exponential, and polynomial tails with $k \ge 6$, the conditions on h_n required by (H3) are already fulfilled when assuming (H2). Consequently, the rate $h_n \propto n^{-1/5}$, which is the optimal bandwidth corresponding to the kernel density estimator for estimating f_0 (Wand & Jones, 1994) is included by our set of assumptions. By contrast, as soon as $k < 6$ in the polynomial case, we have the additional condition that $nh_n^k \to 0$.

4 | NUMERICAL ILLUSTRATIONS

4.1 | Parameter tuning

The computation of the fitness coefficient depends on four tuning parameters: the function q, the sequence $(\Delta_n)_{n\geq 1}$, the kernel K, and the bandwidth $(h_n)_{n\geq 1}$. In our experiments, we have observed that the results are very robust to the choice of q and Δ_n. In particular, we have observed that $\Delta_n = n^{-r}$ with $r \geq 1$, and q being a large tail density function work fine in general. Concerning the kernel K, it is well known in nonparametric density estimation that the choice of K does not influence much the results (Silverman, 1998). As for the choice of the bandwidth, the values returned by the bandwidth selection methods of the literature are expected to give satisfactory results. Indeed, assuming that the tail of f_0 is not too heavy, these methods aim at estimating the optimal bandwidth $n^{-1/(d+4)}$, which satisfies the conditions of Theorem 2. Thus, a bandwidth of order $n^{-1/(d+4)}$ should be reasonable for computing the fitness coefficient.

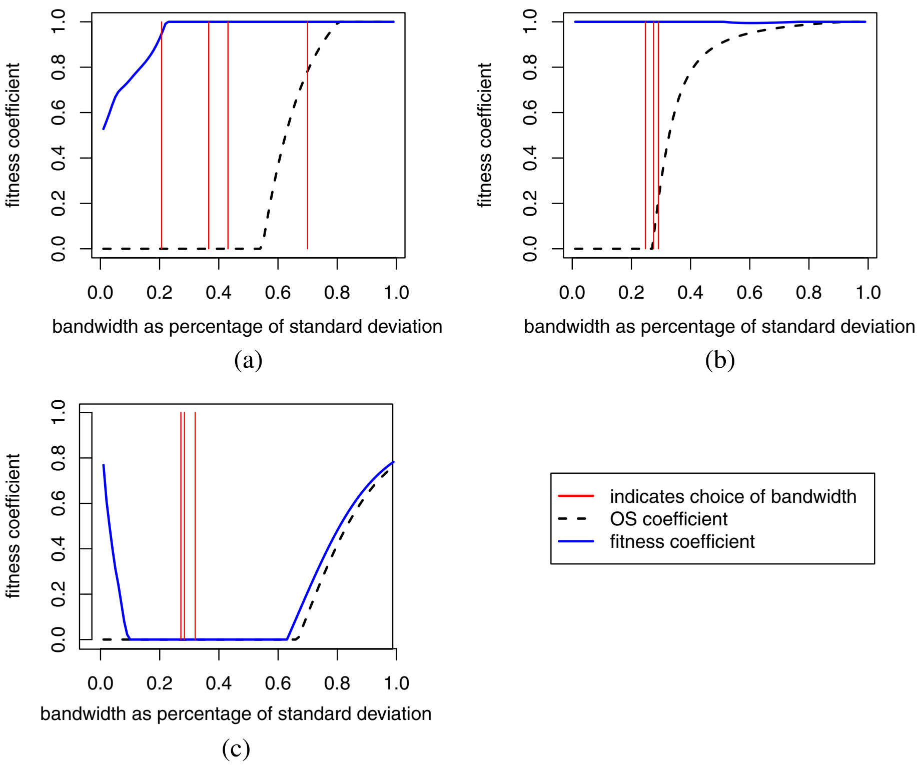
In all the simulation experiments, we set $\Delta_n = 1/n$, $K(x) \propto \exp(-x^2/2)$ and $q(x) = t_\nu((x-\mu_q)/\sigma_q)/\sigma_q^d$ where t_ν is the density of a Student-t distribution with $\nu = 3$ degrees of freedom, $\mu_q = 0$ and $\sigma_q = 100$. With such a large variance and heavy tails, this choice of q is noninformative. In all the experiments but those in Section 4.2, the bandwidth is chosen according to the well-known rule of thumb given in Silverman (1998, p. 48, equation (3.31)). All the numerical experiments were carried out with the `R` software.

4.2 | Sensitivity to the bandwidth: Comparison of the fitness coefficient and the OS coefficient

In this section, we study how a change in the bandwidth affects the fitness coefficient and the OS coefficient. We reanalyze the data set used in Olkin and Spiegelman (1987), consisting of yearly wind speed maxima taken in the north direction in Sheridan, Wyoming. There are 20 observations for the years 1958 to 1977: 70, 61, 61, 60, 61, 63, 61, 67, 61, 62, 47, 67, 61, 49, 55, 65, 57, 51, 47, and 56. The parametric model is a Gumbel model, that is, $\log f_\theta(x) = (x-\mu)/\sigma - \exp((x-\mu)/\sigma)$, where $\theta = (\mu, \sigma)$, μ is a real location parameter and $\sigma > 0$ a dispersion parameter obeying $\mathrm{var}_{f_\theta} = \pi^2\sigma^2/6$ and $\mathbb{E}_{f_\theta} = \mu - \sigma\gamma$, where $\gamma \approx 0.58$ is the Euler–Mascheroni constant. The maximum likelihood estimator is given by $\hat{\theta}_n \approx (62.1, 5.4)$.

Let h denote the bandwidth. In Olkin and Spiegelman (1987), it was arbitrarily chosen $h = 0.7s$, where s is the standard deviation of the data. This yields $\hat{\alpha}_n^{\mathrm{OS}} \approx 0.8$. But if $h \approx 0.43s$, $h \approx 0.37s$ or $h \approx 0.21s$ then one gets $\hat{\alpha}_n^{\mathrm{OS}} \approx 0$. All the above values for h are grounded by well-known bandwidth selection methods, see the textbook Silverman (1998, p. 47, equation (3.30) and p. 48, equation (3.31)) and Sheather and Jones (1991). By contrast, the fitness coefficient yields $\hat{\alpha}_n \approx 1$. These findings are summarized in Figure 1a, where the coefficients are represented as functions of h. We see that the OS coefficient is sensitive to the choice of the bandwidth: a slight difference in h can yield a large difference in $\hat{\alpha}_n^{\mathrm{OS}} = \hat{\alpha}_n^{\mathrm{OS}}(h)$ especially in the range $0.4 \leq h \leq 0.8$. On the opposite, the fitness coefficient is more robust: the estimated value for $\hat{\alpha}_n(h)$ remains close to one in a large range for h. In Figure 1a, the fitness coefficient and the OS coefficient contradict each other and no more credit can be given to any one of them because the ground truth is unknown.

To observe the behavior of the coefficients when the model is known to be true, we simulated $n = 400$ observations according to a Gumbel distribution with mean and SD equal to those of the wind speed data, that is, 59.1 and 6.55, respectively. The results are shown in Figure 1b. One has



(a)

(b)

(c)

FIGURE 1 Values of the fitness coefficient and the OS coefficient as a function of the bandwidth h, expressed as a proportion of the standard error of the data. Plain blue line: fitness coefficient. Dashed black line: OS coefficient. The red sticks indicate various bandwidth values chosen according to the literature (see text); (a) wind speed data, (b) simulations under a Gumbel model, and (c) simulations under a Gaussian model. In all cases the fitted model is a Gumbel model [Color figure can be viewed at wileyonlinelibrary.com]

$\hat{\alpha}_n(h) \approx 1$ whatever h while $\hat{\alpha}_n^{\mathrm{OS}}(h) \leq 0.2$ for all h chosen by the bandwidth selection methods of the literature. These results tend to indicate that the fitness coefficient is consistent but the OS coefficient is not. Let us note that Figure 1a,b is similar, making the Gumbel model plausible. The difference spotted in the range $0 \leq h \leq 0.2$ can be explained by the ties of the wind speed data. (When h is small, one can see that (1) is close to the likelihood of a Bernoulli trials experiment, the maximizer of which is given by the proportion of untied observations, here one-half.)

Whenever the model is wrong, we found on simulations that for most reasonable (i.e., found in the literature as above) values of h, the values of the coefficients are close to zero, as expected. This is illustrated in Figure 1c: the model is still Gumbel, but the $n = 400$ data points were generated according to a Gaussian distribution with mean 59.1 and SD 6.55.

4.3 | Performance of the methods when the model and the truth intertwine

Parametric estimators perform better than kernel density estimators when the model is approximately true, but worse otherwise. Can the semiparametric combination be uniformly best? Does the fitness coefficient goes to unity as the model approaches the truth?

We generated samples of size $n = 400$ according to a density f_t, for several values t in a certain index set, representing the "distance" between f_t and the model. The parametric model is given by $f_\theta \sim \mathcal{N}(\theta, 1)$ and the curve of true distributions is given by $f_t \sim \mathcal{N}(0, (1+t)^2)$. The intersection between the model $\{f_\theta\}$ and the family $\{f_t\}$ is given by $f_0 \sim \mathcal{N}(0, 1)$; that is, $\theta = t = 0$.

For each t, we compute the maximum likelihood estimator, the standard kernel density estimator, the fitness coefficient, the OS coefficient, and the semiparametric density estimator. The semiparametric density estimator is the combination between the maximum likelihood estimator and the kernel density estimator where the mixing coefficient can be either the fitness coefficient (LR method) or the OS coefficient (OS method). To assess the performances of the estimators, we compute the L2-distance to f_t. The above procedure is repeated 500 times so that the errors are averaged over the repetitions.

The errors for the parametric estimator, shown in Figure 2a, shrink sharply as the model and the truth intersect. The error for the nonparametric estimator is approximately constant. We see that the OS method performs poorly: it fails to give accurate estimates near the truth. This behavior is explained in Figure 2b, where we see that the values of the OS coefficient barely exceed 0.1. This is not the case for the fitness coefficient; the values stretch entirely the range $[0, 1]$ and are consistent with the proximity between the truth and the parametric model. As a consequence, coming back to Figure 2a, the error of the LR method is near the minimum of the parametric and nonparametric errors. This means that, in practice, however close our parametric model is to the truth, we never lose by choosing the LR method. Even more interestingly is the fact that in the region where the parametric and the nonparametric estimators perform similarly, the LR method performs better: this corresponds to the values $t \approx -0.10$ and $t \approx 0.15$. This fact is clearly seen in Figure 2c which pictures the averaged error integrated in the interval $[-t, t]$: the LR method always has the lowest curve.

The results for $n = 50, 100, 200$ and another setting, with $f_\theta \sim \mathcal{N}(0, \theta^2)$ and $f_t \sim \mathcal{N}(t, 1)$, are similar and not shown here to limit the length of the article.

4.4 | Application to multivariate density estimation

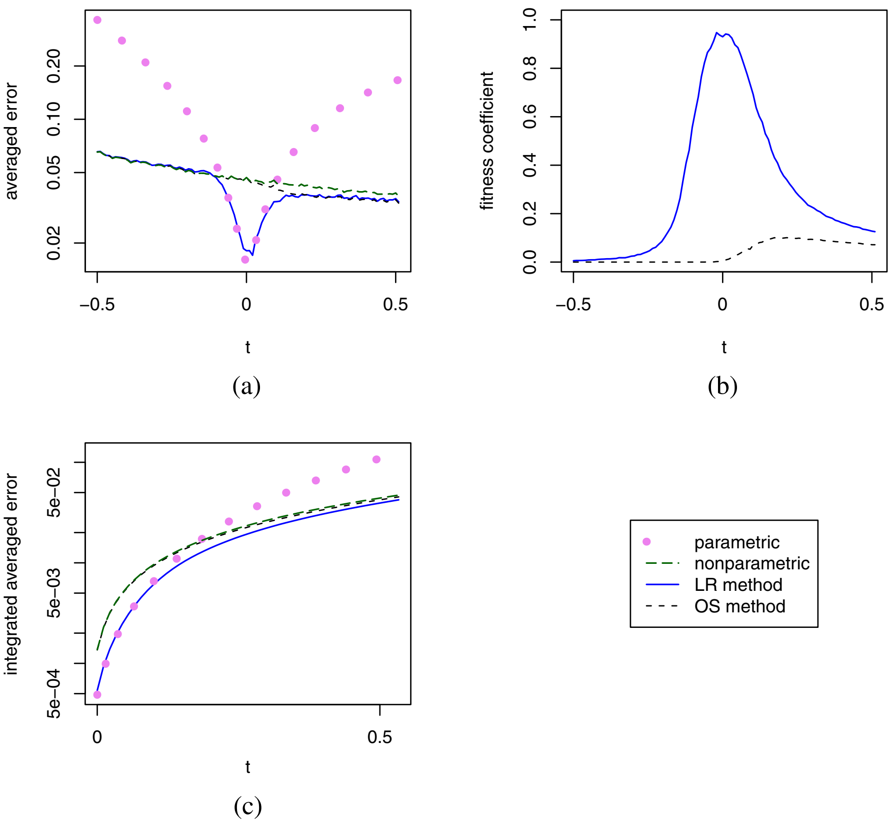
It is well known that building accurate multivariate parametric models is an uncertain and difficult task. One way of addressing this problem consists of decomposing the target density f_0 into a copula c and the marginal densities $f_1, \dots, f_d$, that is,

$$f(x_1, \dots, x_d) = c(F_1(x_1), \dots, F_d(x_d)) f_1(x_1) \dots f_d(x_d)$$

(here the $\{F_j\}$ stand for the distribution functions). This decomposition, also known as Sklar's theorem, is unique provided that the $\{F_j\}$ are continuous; for more details about copulas, see, for example, Genest and Favre (2007) or the books (Joe, 2014; Nelsen, 2006). The copula is assumed to belong to a parametric model $\{c_\xi, \ \xi \in \Xi\}$ and the true underlying parameter ξ is estimated (Genest, Ghoudi, & Rivest, 1995) by

$$\hat{\xi} = \arg\max_{\xi \in \Xi} \sum_{i=1}^{n} \log c_\xi \left(\frac{R_{i,1}}{n}, \dots, \frac{R_{i,d}}{n} \right),$$

where $R_{i,j}$ is the rank of $X_{i,j}$ among $(X_{1,j}, \dots, X_{n,j})$ and $X_{i,j}$ stands for the jth coordinate of the ith observation. The marginals are estimated in a separate step. If one of the marginals is



(a)

(b)

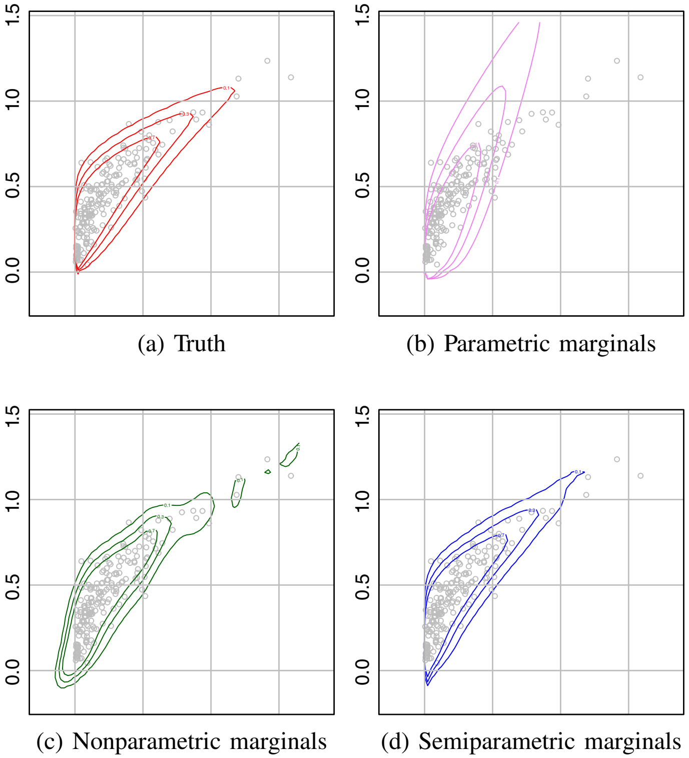
(c)

FIGURE 2 Performance of the methods when the truth $\{\mathcal{N}(0, (1+t)^2),\ -0.5 < t < 0.5\}$ approaches the model $\{\mathcal{N}(\theta, 1),\ -\infty < \theta < \infty\}$ until they intersect at $t = 0$. The L2-distance averaged over the replications is pictured in (a) for the parametric estimator, the nonparametric estimator, the OS method and the LR method. The integrated averaged distance is pictured in (c). Figure (b) pictures the values of the fitness coefficient and the OS coefficient averaged over the replications [Color figure can be viewed at wileyonlinelibrary.com]

misspecified, the estimation of the joint distribution is biased. In the following, a computer experiment illustrates that the LR method can help to reduce this bias by avoiding misspecification.

We have generated data sets of size $n = 25, 50, 100, 150, \dots, 500$ using a Gumbel copula with parameter $\xi = 3$ and marginals $f_1 \sim E(2)$, $f_2 \sim W(2, 1/2)$ where $E(\lambda)$ is an exponential distribution with mean $1/\lambda$ and $W(a, b)$ is a Weibull distribution with shape $a > 0$ and scale $b > 0$. For each of the simulated data sets, the copula parameter ξ was estimated as mentioned above and the marginals were estimated under three scenarios: using a kernel density estimator; a maximum likelihood estimate based on the exponential distribution for both margins; and using the semiparametric combination based on the LR method.

Figure 3 shows the estimation for the bivariate joint density for $n = 200$. In Figure 3b we see that one marginal misspecification led to a poor estimation of the joint density, especially in the joint tails. Figure 3c shows the estimated joint density with the nonparametric strategy for the marginals. Drawbacks of nonparametric estimation are easily spotted: the estimated density is multimodal and assumes positive values where it should be null. Visually, the best performance is



(a) Truth

(b) Parametric marginals

(c) Nonparametric marginals

(d) Semiparametric marginals

FIGURE 3 Contour plots of the true (a) and the estimated joint densities with the parametric (b), nonparametric (c), and semiparametric (d) strategies. The size of the data set is $n = 200$ [Color figure can be viewed at wileyonlinelibrary.com]

achieved with the semiparametric strategy in Figure 3d. The figures for $n = 50, 100, 500$ are similar and not shown to limit the length of the article.

The squared L2-distances between the true joint density and the estimators are shown in Figure 4. The semiparametric strategy performs best for all sample sizes.

5 | DISCUSSION

In the previous sections, we have mostly focused on the estimation properties related to the fitness coefficient and we have demonstrated that the difficulties (practical and theoretical) related to the initial OS method can be overcome by using the LR estimate of the density (as illustrated for instance in Figure 1). Apart from the use of the fitness coefficient in building robust (against misspecification) density estimate, not so much has been said about other potential applications of the fitness coefficient. The aim of this section is to show that it may be used (i) to assess the quality of a given model, (ii) to detect when the dimension is too large to make use of nonparametric estimates, and (iii) to achieve model selection. Finally we shall see that (iv) the fitness coefficient may be extended to the regression setting.

5.1 | Model assessment with the fitness coefficient

The fitness coefficient, which follows from a competition between the parametric and the nonparametric approaches so as to maximize the likelihood of the data, may be interpreted as a

FIGURE 4 Squared L2-distances between the true joint density and the estimators in function of the sample size. From bottom to top, the plain blue line, green dashed line, and violet dotted line are the semiparametric, the nonparametric, and the parametric error curves, respectively [Color figure can be viewed at wileyonlinelibrary.com]

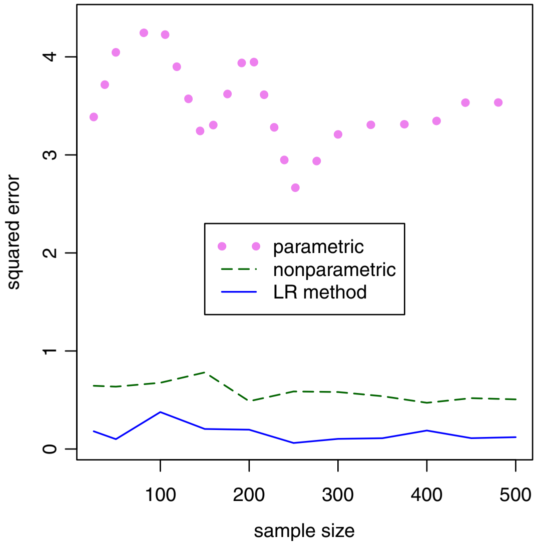
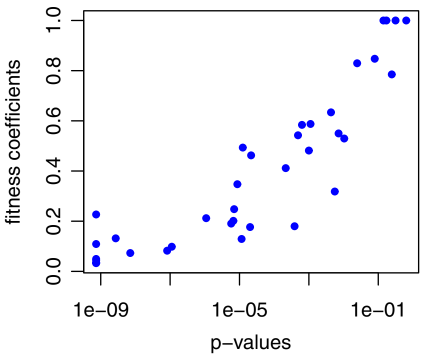


FIGURE 5 Estimated p-values (on a logarithmic scale) of the Cramer–von Mises goodness-of-fit normality test against the values of the fitness coefficient for the CAC40 data [Color figure can be viewed at wileyonlinelibrary.com]



measure of how well the model fits the data with regard to the nonparametric alternative. This intuition has been confirmed by Theorem 1 in which it is shown that $\hat{\alpha}_n$ goes to 1 when the model is true and to 0 when the model is wrong.

To support the interpretation of $\hat{\alpha}_n$ as a measure of model plausibility, we compare it with p-values on a real data example. Specifically, we have $n = 409$ financial returns of 38 companies from the French stock market CAC40 and we wish to measure the quality of the normal model for each of these companies. On the one hand a goodness-of-fit test based on the Cramer–von Mises statistic (D'Agostino & Stephens, 1986) is carried out. On the other hand, the fitness coefficient defined by (1) is computed with a Gaussian kernel K, $\Delta_n = 1/n$, $q(x) = t_\nu((x - \mu_q)/\sigma_q)/\sigma_q^d$ (as in the previous section), and $h_n = 1.06An^{-1/5}$, where A is the minimum between the standard deviation of the data and, the interquartile range divided by 1.34. This value for the bandwidth is a standard choice, see Silverman (1998, p. 48). In Figure 5, we plotted the values of the fitness coefficient against the p-values on a logarithmic scale. We can see a clear positive dependence relationship suggesting that, if one had used p-values to assess the fitness of the normal model, he or she could have done so with the fitness coefficient.

Given the value $\hat{\alpha}_n \in [0, 1]$, to assess whether the model is true or not is an interesting and yet still open question. To address this type of problem, the general approach consists in deriving, under the null hypothesis (in our context when the model is true), the weak convergence of the sequence $r_n(\hat{\alpha}_n - 1)$, for some sequence $r_n \to \infty$. The knowledge of the limiting distribution (or some approximation) might help to define appropriate reject regions according to the desired level. Deriving the limiting distribution of $r_n(\hat{\alpha}_n - 1)$ is not straightforward. The classical asymptotic theory of Z-estimation does not apply here and one possible avenue of investigation would be to use the so called *argmax theorem* (van der Vaart & Wellner, 1996, chapter 3.2) which requires to obtain the rate of convergence r_n as a first step. Following the approach taken in our consistency proof, the key quantities comes from a Taylor expansion of the likelihood. They take the form

$$\sum_{i=1}^{n} \left(\frac{\hat{f}^{\mathrm{LR}}_{n,i} - f_{0,i}}{f_{0,i}} \right)^k, \quad k = 1, 2.$$

Unfortunately, our analysis of these terms, conducted in Lemmas 1 and 2 in the Appendix, is not enough to obtain the rate of convergence r_n nor the weak convergence of $r_n(\hat{\alpha}_n - 1)$. We believe that this question is beyond the scope of the present article but represents an avenue for further research.

Note that the quality criterion induced by the fitness coefficient is different than that of information criteria (Burnham & Anderson, 2003; Claeskens & Hjort, 2008) such as the Akaike information criterion (Akaike, 1974) or the Bayesian information criterion (Schwarz, 1978) which focus on the *relative* performance *between* models (e.g., p. 30 in Claeskens and Hjort (2008) for their relationship with the Kullback–Leibler distance). Note also that convex parametric combinations recently have been proposed in the Bayesian literature (Kamary, Mengersen, Robert, & Rousseau, 2014) to assess the fitness of a certain parametric model against another.

5.2 | The problem of higher dimensions

Although our results hold for any value of d, the accuracy of kernel methods (and hence that of the LR) deteriorates in high dimensions. This phenomenon, called the *curse of dimensionality*, might prevent one from using the fitness coefficient as a quality criterion (as developed in the previous section) when the dimension is high.

As an illustration, suppose that the density of the observed data is $f_0 = f_{01}/2 + f_{02}/2$, where $f_{01} \sim \mathcal{N}(\mu_{01}, I_d/d)$, $\mu_{01} = (3, 0, \dots, 0)$, and $f_{02} \sim \mathcal{N}(\mu_{02}, I_d/d)$, $\mu_{02} = (-3, 0, \dots, 0)$. Note that the variances of each densities f_{01} and f_{02} have been scaled down such that the dispersion around their mean does not depend on the dimension. The model is $f_\theta \sim \mathcal{N}(\mu, \sigma^2 I_d)$, $\theta = (\mu, \sigma^2) \in \mathbb{R}^d \times (0, \infty)$. For each dimension $d = 1, \dots, 15$, and each sample size $n = 1000, 2000, 5000$, we compute the fitness coefficient over $M = 50$ Monte Carlo experiments. The resulting boxplots are depicted in Figure 6 where d has been restricted to $\{6, \dots, 14\}$. They illustrate nicely the curse of dimensionality because, even though the model is wrong, the fitness coefficient increases with the dimension. When d is small, for example, $d \leq 9$ and $n = 5000$, the kernel density estimator dominates the MLE but when the dimension increases, the MLE (based on a wrong model) takes over.

With regard to the difficulties faced by the LR method in high dimensions, we would like to raise two points.

(i) When the dimension is too large compared to n, which could be inferred based on simulations (as in Figure 6), one may replace the kernel density estimate by other nonparametric density estimator that suffer less from the curse of dimensionality. Such candidates include,

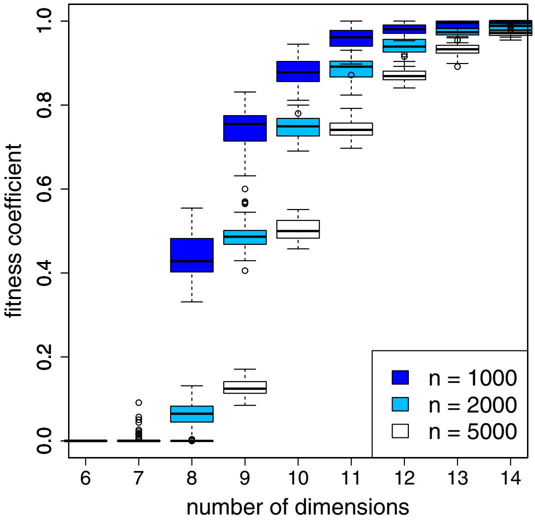


FIGURE 6 Boxplots of the fitness coefficients for the model $f_{01}/2 + f_{02}/2$. The x-line corresponds to different dimensions going from 6 to 14. The sample sizes are $n = 1000$ (blue), $n = 2000$ (light blue), and $n = 5000$ (white) [Color figure can be viewed at wileyonlinelibrary.com]

among others, projection-pursuit density estimators (Hwang, Lay, & Lippman, 1994), maximum likelihood log-concave density estimators (Cule, Samworth, & Stewart, 2010), nonparametric density estimation with a parametric start (Hjort & Glad, 1995), penalized likelihood methods (Silverman, 1998) and structured nonparametric methods, such as nonparametric vine copulas methods (Nagler & Czado, 2016). Could still one get a consistency result as Theorem 1? The proof of Theorem 1 essentially consists of checking the conditions of Theorems 2 and 3, the statements of which are valid for any nonparametric estimator $\xi_{n,i}$. Thus, if one is able to check those conditions for other nonparametric estimators, then Theorem 1 will remain true. To check the conditions of Theorems 2 and 3, one needs to compare the linear and quadratic errors associated with the nonparametric estimators to their parametric counterparts. To do so, however, we heavily relied on rates valid for kernel density estimators, which may not be extended trivially to other nonparametric estimators.

(ii) In high-dimensions, the fitness coefficient may no longer be interpreted as a measure of model quality because the nonparametric estimator will perform badly anyway. Note however that the fitness coefficient remains useful as a selection criterion because it allows to compare between any two given approaches.

5.3 | Model selection

The approach of this article could be adapted to the case where an arbitrary number of parametric models are in competition with each other. Suppose that $\mathcal{F}_1, \dots, \mathcal{F}_p$ are p families of the form $\mathcal{F}_i = \{f_i(\cdot;\theta_i),\ \theta_i \in \Theta_i\}$. Suppose furthermore that f_0, the true underlying density, belongs to only one model $\mathcal{F}_k$ among the collection of models, that is, $f_0 \in \mathcal{F}_k$ and $f_0 \notin \mathcal{F}_j$, for all $j \neq k$. Then the selection of one of the models could be performed by estimating each models $f_j(\cdot;\hat{\theta}_j), j = 1, \dots, p$, and then maximizing

$$\sum_{i=1}^{n} \log(\alpha_1 f_1(X_i;\hat{\theta}_1) + \dots + \alpha_p f_p(X_i;\hat{\theta}_p))$$

over all $\alpha_1, \dots, \alpha_p$ such that $\sum_{j=1}^{p} \alpha_j = 1$ and $\alpha_j \geq 0$ for all $j = 1, \dots, p$. In this situation, we expect that the previous maximizer $(\hat{\alpha}_{n,1}, \dots, \hat{\alpha}_{n,p})$ would converge to $(0, \dots, 0, 1, 0, \dots, 0)$, where the 1 is placed at the kth coordinate. Indeed the identification condition ensures that one and only one of the $f_i(\cdot; \hat{\theta}_i)$ will be close to f_0. Thus the "multivariate" fitness coefficient would be a pointer to the true model. Even though all the estimators could possibly have the same rates of convergence, the proof of the convergence to the multivariate fitness coefficient should be feasible because only one density estimator goes to the true density; in this respect this is similar to the case studied in Theorem 1(ii) (actually simpler since we only have to deal with parametric models).

5.4 | Extension to the regression setting

The approach of this article could be extended to the problem of regression. Let $Y_i = m_0(x_i) + \varepsilon_i$, $i = 1, 2, \dots$, where the $x_i \in \mathbb{R}^d$ are some covariates, ε_i some centered random errors with common variance σ^2 and m_0 is an unknown function to estimate based on n independent but not identically distributed observations $Y_1, \dots, Y_n$. On the one hand, one can consider a parametric model of the form $\{m_\theta : \mathbb{R}^d \to \mathbb{R} \ : \ \theta \in \Theta\}$, for example, the linear model is given by $m_\theta(x) = \theta^T x$ with $\Theta = \mathbb{R}^d$, and compute the estimator

$$\hat{\theta}_n = \operatorname{argmin}_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} (Y_i - m_\theta(x_i))^2.$$

On the other hand, one can choose a nonparametric estimator. For instance, the well-known Nadaraya–Watson estimator is given by

$$\hat{m}_n(x) = \frac{\sum_{i=1}^{n} K_h(x - x_i) Y_i}{\sum_{i=1}^{n} K_h(x - x_i)},$$

for some bandwidth $h > 0$. To get a tradeoff between these two approaches, we could solve

$$\hat{\alpha}_n = \underset{\alpha \in [0,1]}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \alpha m_{\hat{\theta}_n}(x_i) - (1 - \alpha) \hat{m}_n^{(i)}(x_i))^2, \tag{4}$$

where $\hat{m}_n^{(i)}$ is the leave-one-out version of $\hat{m}_n$. We expect $\hat{\alpha}_n$ be close to one when m_0 belongs to the parametric model and near zero otherwise. Heuristically, when the model is valid, since $m_{\hat{\theta}_n}$ converges to m_0 faster than $\hat{m}_n$, the former should be quite close to m_0 while, comparatively, the later should stay away from it. As a consequence, the above objective function should be similar to $\alpha \mapsto E(Y_1 - m_0(x_1))^2 + (1 - \alpha)^2 \frac{1}{n} \sum_{i=1}^{n} (m_0(x_i) - \hat{m}_n^{(i)}(x_i))^2$, whose argmin is $\alpha = 1$. Indeed, we have that the objective function in (4) with m_0 in place of $m_{\hat{\theta}_n}$ is equal to

$$\frac{1}{n} \sum_i (Y_i - m_0(x_i))^2 + (1 - \alpha)^2 \frac{1}{n} \sum_i (m_0(x_i) - \hat{m}_n^{(i)}(x_i))^2$$
$$+ 2(1 - \alpha) \frac{1}{n} \sum_i (Y_i - m_0(x_i))(m_0(x_i) - \hat{m}_n^{(i)}(x_i)).$$

The first term goes to $E(Y_1 - m_0(x_1))^2$ and it is reasonable to expect the last term to go to zero thanks to the existence of laws of large numbers for sequences of independent but nonidentically distributed random variables, see, for example, theorem 3 in Feller (1971, p. 239). To avoid bad effect due to overfitting, it might be helpful to split the data into two independent samples. The first sample would be used to compute $\hat{\theta}_n$ and $\hat{m}_n$, and the second one would be used for $\hat{\alpha}_n$.

ACKNOWLEDGEMENTS

The author gratefully acknowledge *"Fonds de la Recherche Scientifique—FNRS" (Belgium)*. The author gratefully acknowledges support from the Fonds de la Recherche Scientifique (FNRS) A4/5 FC 2779/2014-2017 No. 22342320, from the contract "Projet d'Actions de Recherche Concertées" No. 12/17-045 of the "Communauté française de Belgique." The authors thank two anonymous referees for their helpful remarks and suggestions. The authors are grateful to Ingrid Van Keilegom, Anouar El Ghouch, and Sylvie Huet for useful comments and references. The authors are also grateful to Christian Robert for some motivating remarks at the very beginning of this work.

ORCID

Gildas Mazo https://orcid.org/0000-0003-3189-6818

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Burnham, K. P., & Anderson, D. R. (2003). *Model selection and multimodel inference: A practical information-theoretic approach*. Berlin, Germany: Springer Science & Business Media.

Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging* (Vol. *330*). Cambridge, MA: Cambridge University Press.

Cule, M., Samworth, R., & Stewart, M. (2010). Maximum likelihood estimation of a multi-dimensional log-concave density. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(5), 545–607.

D'Agostino, R. B., & Stephens, M. A. (1986). *Goodness-of-fit techniques*. New York, NY: Marcel Dekker Inc.

Delecroix, M., Hristache, M., & Patilea, V. (2006). On semiparametric M-estimation in single-index regression. *Journal of Statistical Planning and Inference*, *136*(3), 730–769.

Einmahl, U., & Mason, D. M. (2000). An empirical process approach to the uniform consistency of kernel-type function estimators. *Journal of Theoretical Probability*, *13*(1), 1–37.

Faraway, J. (1990). Implementing semiparametric density estimation. *Statistics and Probability Letters*, *10*(2), 141–143.

Feller, W. (1971). *An introduction to probability theory and its applications*. Hoboken, NJ: Wiley.

Genest, C., & Favre, A. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, *12*(4), 347–368.

Genest, C., Ghoudi, K., & Rivest, L. P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, *82*(3), 543–552.

Giné, E., & Guillou, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, *38*(6), 907–921.

Grenander, U. (1981). *Abstract inference*. New York, NY: John Wiley & Sons, Inc.

Habbema, J. D. F., Hermans, J., & Van Den Broeck, K. (1974). *A stepwise discriminant analysis program using density estimation*. Vienna, Austria: Physica Verlag.

Hall, P. (1987). On Kullback-Leibler loss and density estimation. *The Annals of Statistics*, *15*(4), 1491–1519.

Hjort, N. L., & Glad, I. K. (1995). Nonparametric density estimation with a parametric start. *The Annals of Statistics*, *23*(3), 882–904.

Hjort, N. L., & Jones, M. C. (1996). Locally parametric nonparametric density estimation. *The Annals of Statistics*, *24*(4), 1619–1647.

Hjort, N. L., McKeague, I. W., & Van Keilegom, I. (2018). Hybrid combinations of parametric and empirical likelihoods. *Statistica Sinica*, *28*(4), 2389.

Hwang, J.-N., Lay, S.-R., & Lippman, A. (1994). Nonparametric multivariate density estimation: A comparative study. *IEEE Transactions on Signal Processing*, *42*(10), 2795–2810.

Joe, H. (2014). *Dependence modeling with copulas*. Boca Raton, FL: Chapman & Hall.

Kamary, K., Mengersen, K., Robert, C. P., & Rousseau, J. (2014). Testing hypotheses via a mixture estimation model. *arXiv preprint arXiv:1412.2044*.

Lee, S. S. M., & Soleymani, M. (2015). A simple formula for mixing estimators with different convergence rates. *Journal of the American Statistical Association*, *110*(512), 1463–1478.

Marron, J. S. (1985). An asymptotically efficient solution to the bandwidth problem of kernel density estimation. *The Annals of Statistics*, *13*(3), 1011–1023.

Marron, J. S. (1987). A comparison of cross-validation techniques in density estimation. *The Annals of Statistics*, *15*(1), 152–162.

Murphy, S. A. (1994). Consistency in a proportional hazards model incorporating a random effect. *The Annals of Statistics*, *22*(2), 712–731.

Nagler, T., & Czado, C. (2016). Evading the curse of dimensionality in nonparametric density estimation with simplified vine copulas. *Journal of Multivariate Analysis*, *151*, 69–89.

Nelsen, R. B. (2006). *An introduction to copulas*. New York, NY: Springer.

Newey, W. K., & McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics*, *4*, 2111–2245.

Nolan, D., & Pollard, D. (1987). U-processes: Rates of convergence. *The Annals of Statistics*, *15*(2), 780–799.

Olkin, I., & Spiegelman, C. H. (1987). A semiparametric approach to density estimation. *Journal of the American Statistical Association*, *82*(399), 858–865.

Pollard, D. (2000). Asymptopia. *Unpublished book*.

Portier, F., & Segers, J. (2018). On the weak convergence of the empirical conditional copula under a simplifying assumption. *Journal of Multivariate Analysis*, *166*, 160–181.

Rahman, M., Beaver, R. J., & Gokhale, D. V. (1997). A note on estimating the combining constant in semiparametric density estimation. *Brazilian Journal of Probability and Statistics*, *11*(1), 37–50.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, *9*(2), 65–78.

Schuster, E. F., & Gregory, G. G. (1981). *On the nonconsistency of maximum likelihood nonparametric density estimators*. Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface (pp. 295–298). Springer, New York, NY.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Sheather, S. J., & Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, *53*(3), 683–690.

Silverman, B. W. (1998). *Density estimation for statistics and data analysis*. Boca Raton, FL: Chapman & Hall.

Soleymani, M., & Lee, S. M. S. (2014). A bootstrap procedure for local semiparametric density estimation amid model uncertainties. *Journal of Statistical Planning and Inference*, *153*, 75–86.

Stone, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics*, *12*(4), 1285–1297.

Talamakrouni, M., El Ghouch, A., & Van Keilegom, I. (2017). Parametrically guided local quasi-likelihood with censored data. *Electronic Journal of Statistics*, *11*(2), 2773–2799.

van der Vaart, A., & Wellner, J. A. (2000). *Preservation theorems for Glivenko-Cantelli and uniform Glivenko-Cantelli classes*. In *High dimensional probability II* (pp. 115–133). New York, NY: Springer.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge, NY: Cambridge University Press.

van der Vaart, A. W., & Wellner, J. A. (1996). *Weak convergence and empirical processes. With applications to statistics*. New York, NY: Springer-Verlag.

Wand, M. P., & Jones, M. C. (1994). Multivariate plug-in bandwidth selection. *Computational Statistics*, *9*(2), 97–116.

How to cite this article: Mazo G, Portier F. Parametric versus nonparametric: The fitness coefficient. *Scand J Statist*. 2020;1–40. https://doi.org/10.1111/sjos.12495

APPENDIX A. PROOFS OF THE PROPOSITIONS

We define the mixture likelihood function $L_n : [0, 1] \to [-\infty, +\infty)$ as

$$L_n(\alpha) = \sum_{i=1}^{n} \log(\alpha f_{\hat{\theta}_n}(X_i) + (1-\alpha)\hat{f}_{n,i}^{\mathrm{LR}}).$$

The fitness coefficient $\hat{\alpha}_n$ in (1) is then defined as a maximizer of $L_n(\alpha)$ over $[0, 1]$.

A.1 Proof of Proposition 1

The presence of $\Delta_n q(X_i)$ in $\hat{f}_{n,i}^{\mathrm{LR}}$ allows for $L_n(\alpha) > -\infty$ for all $\alpha \in [0, 1)$. If for all $i, f_{\hat{\theta}_n}(X_i) > 0$, that is, $L_n(1) > -\infty$, then L_n is continuous on $[0, 1]$ and the extreme value theorem yields the existence of $\hat{\alpha}_n$. Else, if $L_n(1) = -\infty$, there exists $\delta > 0$ such that $\sup_{\alpha\in[0,1-\delta]} L_n(\alpha) > \sup_{\alpha\in(1-\delta,1]} L_n(\alpha)$, meaning that the maximum is over $[0, 1-\delta]$ and exists in virtue of the extreme value theorem. Whenever $f_{\hat{\theta}_n}(X_i)$ is not identically equal to $\hat{f}_{n,i}^{\mathrm{LR}}$ for all $i = 1, \ldots, n$, the function L_n is strictly concave and so comes the unicity.

A.2 Proof of Proposition 2

The proof follows from the decomposition

$$\begin{aligned} |\tilde{f}_n(x) - f_0(x)| &= |\hat{\alpha}_n(f_{\hat{\theta}_n}(x) - f_0(x)) + (1-\hat{\alpha}_n)(\hat{f}_n(x) - f_0(x))| \\ &\leq \hat{\alpha}_n |f_{\hat{\theta}_n}(x) - f_0(x)| + (1-\hat{\alpha}_n)|\hat{f}_n(x) - f_0(x)|. \end{aligned}$$

Under (H1) and (H2), we have (Portier & Segers, 2018, proposition C.1) (see also theorem 2.1 in Giné and Guillou (2002))

$$\|\hat{f}_n - f_{h_n}\|_{\mathbb{R}^d} = O_{\mathbb{P}}\left(\sqrt{\frac{|\log(h_n)|}{nh_n^d}}\right).$$

Using Lemma 7, we obtain that

$$\|\hat{f}_n - f_0\|_{\mathbb{R}^d} = O_{\mathbb{P}}\left(\sqrt{\frac{|\log(h_n)|}{nh_n^d}}\right) + O(h_n^2). \tag{A1}$$

As shown in (C2), when $f_0 \in \mathcal{P}$ and under (A1), we have that $\|f_{\hat{\theta}_n} - f_0\|_{\mathbb{R}^d} \to 0$, in probability. Consequently, if $f_0 \in \mathcal{P}$, the conclusion holds regardless of the convergence of $\hat{\alpha}_n$. Suppose that

$f_0 \notin \mathcal{P}$, then

$$|\tilde{f}_n(x) - f_0(x)| \leq \hat{\alpha}_n(\sup_{\theta\in\Theta} \|f_\theta\|_{\mathbb{R}^d} + \|f_0\|_{\mathbb{R}^d}) + (1-\hat{\alpha}_n)\|\hat{f}_n - f_0\|_{\mathbb{R}^d}.$$

The previous bound goes to 0 in virtue of (A1) and because $\alpha_n \to 0$.

A.3 Proof of Proposition 3

Let $0 < \epsilon < 1$. By assumption, there exists $\tilde{A} > 0$, such that for all $\|x\|_2 > \tilde{A}$, we have

$$(1-\epsilon)g_0(x) \leq f_0(x) \leq g_0(x)(1+\epsilon).$$

For $t > 0$ small enough (i.e., taking any $t < \inf_{\|x\|_2 \leq \tilde{A}} f_0(x)$ implies that $\{\|x\|_2 \leq \tilde{A}\} \subset \{f_0(x) > t\}$, or equivalently that $S_t^c \subset \{\|x\|_2 > \tilde{A}\}$), it holds

$$\int_{S_t^c} f_0(x)\ \mathrm{d}x = \int_{f_0(x)\leq t} f_0(x)\ \mathrm{d}x \leq (1+\epsilon)\int_{(1-\epsilon)g_0(x)\leq t} g_0(x)\ \mathrm{d}x.$$

Consequently, we obtain that $\int_{S_t^c} f_0(x)\ \mathrm{d}x \leq t^\beta c_2(1+\epsilon)/(1-\epsilon)^\beta$.

Remark that

$$\lambda(S_t) \leq \lambda(\{\|x\|_2 \leq \tilde{A}\}) + \lambda(\{\|x\|_2 > \tilde{A}\} \cap S_t) \leq \lambda(\{\|x\|_2 \leq \tilde{A}\}) + \int_{(1+\epsilon)g_0(x)>t} dx,$$

which is enough to obtain the last point of (H3).

Suppose that $0 < h_n \leq 1$. By enlarging $\tilde{A}$ (i.e., taking $\tilde{A} := \tilde{A} + \sqrt{d}$), we have, for all $\|x\|_2 > \tilde{A}$ and $u \in [-1,1]^d$,

$$(1-\epsilon)g_0(x+h_nu) \leq f_0(x+h_nu) \leq g_0(x+h_nu)(1+\epsilon).$$

Let $b_n = \gamma(nh_n^d)^{-1/\beta}$ and $A_1 = \max(A, \tilde{A})$. As soon as $\|x\|_2 \leq A_1$,

$$\sup_{u\in[-1,1]^d} \frac{f_0(x+h_nu)}{f_0(x)} \leq \frac{\|f_0\|_{\mathbb{R}^d}}{\inf_{\|x\|_2\leq A_1} f_0(x)}.$$

Otherwise,

$$\sup_{\|x\|_2 > A_1,\ f_0(x)>b_n} \sup_{u\in[-1,1]^d} \frac{f_0(x+h_nu)}{f_0(x)} \leq \frac{(1+\epsilon)}{(1-\epsilon)} \sup_{\|x\|_2>A_1,\ (1+\epsilon)g_0(x)>b_n} \sup_{u\in[-1,1]^d} \frac{g_0(x+h_nu)}{g_0(x)}.$$

We conclude by remarking that the previous is bounded, by $(1+\epsilon)C_2/(1-\epsilon)$.

APPENDIX B. PROOF OF THEOREM 1

Theorem 1 follows from the application of two high-level results, corresponding respectively to the well-specified and misspecified case. Both high-level results take place in the following general framework: given a triangular sequence of nonnegative real numbers $\xi_{n,i}, i = 1, \dots, n$, $n \geq 1$, we consider the mixture likelihood function given by

$$L_n(\alpha) = \sum_{i=1}^{n} \log(\alpha f_{\hat{\theta}_n}(X_i) + (1-\alpha)\xi_{n,i}).$$

Here the sequence $(\xi_{n,i})$ is left unspecified in order to highlight the assumptions that we need on the nonparametric part. This random sequence could be the nonparametric estimator evaluated at X_i, that is, $\hat{f}_n(X_i)$, the LOO estimate $\hat{f}_{n,i}^{\text{LOO}}$ or the LR estimate $\hat{f}_{n,i}^{\text{LR}}$ with $\Delta_n > 0$. In this slightly new context, we define $\hat{\alpha}_n$ as

$$\hat{\alpha}_n \in \operatorname{argmax}_{\alpha \in [0,1]} L_n(\alpha).$$

In both cases, respectively, the misspecified and well-specified case, the approach taken is similar. We compare the empirical likelihood of the mixture to the one of the parametric estimate (in the well-specified case) or the nonparametric estimate (in the misspecified case).

For ease of readability, we introduce the short-cut notation g_i for $g(X_i)$, for any function g.

B.1 Case (i): The model is well specified

We are based on some restricted mean quadratic error

$$Q_n^{(p)}(S) = \sum_{i=1}^{n} \left(\frac{f_{\hat{\theta}_n,i} - f_{0,i}}{f_{0,i}} \right)^2 1_{\{X_i \in S\}},$$

$$Q_n^{\text{np}}(S) = \sum_{i=1}^{n} \left(\frac{\xi_{n,i} - f_{0,i}}{f_{0,i}} \right)^2 1_{\{X_i \in S\}},$$

and some averaged linear error

$$M_n^{(p)} = \sum_{i=1}^{n} \left(\frac{f_{\hat{\theta}_n,i} - f_{0,i}}{f_{0,i}} \right), \quad M_n^{\text{np}} = \sum_{i=1}^{n} \left(\frac{\xi_{n,i} - f_{0,i}}{f_{0,i}} \right).$$

The proof of the following theorem is given in Section B.3.

Theorem 2. *Suppose that $f_0 \in \mathcal{P}$ and let $S \subset \mathbb{R}^d$ and $b > 0$ be such that for all $x \in S$, $f_0(x) > b$. If the following convergences hold in probability, as $n \to \infty$,*

$$\|f_{\hat{\theta}_n} - f_0\|_S \to 0, \qquad \max_{i=1,\dots,n,\ :\ X_i \in S} |\xi_{n,i} - f_{0,i}| \to 0, \tag{B1}$$

$$\frac{Q_n^{(p)}(S)}{Q_n^{\text{np}}(S)} \to 0, \qquad \frac{|M_n^{(p)}| + |M_n^{\text{np}}|}{Q_n^{\text{np}}(S)} \to 0, \tag{B2}$$

then, $\hat{\alpha}_n \to 1$ as $n \to \infty$, in probability.

We now verify the conditions of the previous theorem when $\xi_{n,i}$ is the LR sequence $\hat{f}_{n,i}^{\mathrm{LR}}$ and when (H1), (H2), (H3), (A1), (A2) and $nh_n^d\Delta_n \to 0$ are fulfilled.

Condition (B1)

The first convergence in (B1) holds in virtue of (C2) established in Section C1. For the second one, it holds that

$$\hat{f}_{n,i}^{\mathrm{LR}} = \left(\frac{n}{n-1}\right)\left(\hat{f}_n(X_i) - \frac{K(0)}{nh_n^d}\right) + \Delta_n q(X_i).$$

Using (A1), we get

$$\begin{aligned}\max_{i=1,\dots,n} |\hat{f}_{n,i}^{\mathrm{LR}} - f_{0,i}| &\le \left(\frac{n}{n-1}\right)\left(\|\hat{f}_n - f_0\|_{\mathbb{R}^d} + \frac{K(0)}{nh_n^d}\right) + \Delta_n\|q\|_{\mathbb{R}^d} + \frac{\|f_0\|_{\mathbb{R}^d}}{n-1}\\ &= O_{\mathbb{P}}\left(\sqrt{\frac{|\log(h_n)|}{nh_n^d}} + h_n^2 + \Delta_n\right).\end{aligned} \tag{B3}$$

The latter bound indeed goes to 0, in probability, as $n \to \infty$.

Condition (B2)

We proceed as follows, with $\xi_{n,i} = \hat{f}_{n,i}^{\mathrm{LR}}$:

(a) By Lemma 1, stated in Section B.4, there exists $c > 0$ such that with probability going to 1, $h_n^d Q_n^{\mathrm{np}}(S) \ge c$. The set S is chosen equal to $\{f_0(x) > b\}$ where $b > 0$ is such that it is nonempty.
(b) We show in Lemma 2 that $h_n^d |M_n^{\mathrm{np}}| \to 0$ in probability.
(c) In Lemma 3 (resp. Lemma 4), it is established, under (A1) and (A2), that $Q_n^{(p)}(S) = O_{\mathbb{P}}(1)$ (resp. $|M_n^{(p)}| = O_{\mathbb{P}}(1)$).

All this together implies that (B2) holds true.

B.2 Case (ii): The model is misspecified

When the model is misspecified, that is, $f_0 \notin \mathcal{P}$, the following high-level conditions are enough to ensure the convergence in probability $\alpha_n \to 0$. These conditions are easily implied by (A1), (H1), (H2), (H3) and

$$|\log(\Delta_n)|^{1/\beta}(\sqrt{|\log(h_n)|/nh_n^d} + h_n^2) \to 0,$$

as demonstrated below. The proof of this theorem is given in Section B.3.

Theorem 3. *Suppose that $f_0 \notin \mathcal{P}$ and $\mathbb{E}[|\log(f_{0,1})|] < \infty$, that the class $\mathcal{P}$ is Glivenko–Cantelli (i.e., (C1) holds) with Θ compact, that the envelop F_Θ is such that $\mathbb{E}[\log(F_{\Theta,1})] < +\infty$, and that for every $x \in \mathbb{R}^d$, $\theta \mapsto f_\theta(x)$ is a continuous function defined on Θ. Suppose that there exists $\beta \in (0, 1]$ and $c > 0$*

such that, as $t \to 0$, $\int_{S_t^c} f_0(x) \leq ct^\beta$, *and* $q : \mathbb{R}^d \to \mathbb{R}^+$ *such that* $\mathbb{E}[|\log(q(X_1))|] < \infty$, $\xi_{n,1} \geq \Delta_n q_1$ *a.s., and*

$$|\log(\Delta_n)|^{1/\beta} \max_{i=1,\dots,n} |\xi_{n,i} - f_0(X_i)| \to 0,$$

then, $\hat{\alpha}_n \to 0$, *as* $n \to \infty$, *in probability.*

We already argued that (C1) is implied by (A1). The continuity of f_θ is deduced from the continuity of $\log(f_\theta)$ provided by (A1). The bound given in (B3) together with $|\log(\Delta_n)|^{1/\beta}(\sqrt{|\log(h_n)|/nh_n^d} + h_n^2) \to 0$ implies the stated convergence with $\xi_{n,i} = \hat{f}_{n,i}^{\mathrm{LR}}$.

B.3 Proofs of the high-level theorems

Proof of Theorem 2

Define $\tilde{L}_n$, the normalized version of $L_n(\alpha)$, given by

$$\tilde{L}_n(\alpha) = \sum_{i=1}^n \log\left(\frac{\alpha f_{\hat{\theta}_n,i} + (1-\alpha)\xi_{n,i}}{f_{0,i}}\right).$$

Because $f_0 \in \mathcal{P}$, it holds that $f_{\hat{\theta}_n,i} > 0$ for all $i = 1, \dots, n$, which guarantees the existence of a maximizer $\hat{\alpha}_n$ (as explained in the proof of Proposition 1). By definition of $\hat{\theta}_n$, $\sum_{i=1}^n \log(f_{0,i}) \leq \sum_{i=1}^n \log(f_{\hat{\theta}_n,i})$. Consequently, $\max_{\alpha\in[0,1]} \tilde{L}_n(\alpha) \geq 0$ and for every $\epsilon > 0$, the event $\max_{\alpha\in[0,1-\epsilon]} \tilde{L}_n(\alpha) < 0$ implies that $\hat{\alpha}_n > 1 - \epsilon$. Thus, let $\epsilon \in (0, 1)$, the proof will be completed by showing that with probability going to 1,

$$\sup_{\alpha\in[0,1-\epsilon]} \tilde{L}_n(\alpha) < 0.$$

A useful notation in the following is

$$\hat{x}_{i,n} = 1 + \frac{\alpha(f_{\hat{\theta}_n,i} - f_{0,i})}{f_{0,i}} + \frac{(1-\alpha)(\xi_{n,i} - f_{0,i})}{f_{0,i}}.$$

A useful technical detail is there exists a sequence $\epsilon_n \to 0$ such that the event

$$\{\max_{i=1\dots,n\ :\ X_i\in S} f_{0,i}|\hat{x}_{i,n} - 1| \leq \epsilon_n\}$$

has probability going to 1 as $n \to \infty$. This is a consequence of (B1). As we are establishing a result in probability, we can further suppose that this event is realized.

A key step in our approach is the following inequality, reminiscent of the Taylor development of the logarithm around 1,

$$\log(x) - (x-1) \leq \begin{cases} -\frac{1}{4}(x-1)^2 & \text{if } 1/2 < x < 3/2 \\ 0 & \text{else} \end{cases},$$

which might be derived by studying the concerned function. This kind of inequality is commonly used for studying likelihood methods (Grenander, 1981; Murphy, 1994). Applied to $\hat{x}_{i,n}$, it gives

$$\begin{aligned}
\tilde{L}_n(\alpha) - (\alpha M_n^{(p)} + (1-\alpha)M_n^{\mathrm{np}}) &= \sum_{i=1}^{n} (\log(\hat{x}_{i,n}) - (\hat{x}_{i,n} - 1)) \\
&\le -\frac{1}{4} \sum_{i=1}^{n} (\hat{x}_{i,n} - 1)^2 1_{\{|\hat{x}_{i,n}-1|<1/2\}} \\
&\le -\frac{1}{4} \sum_{i=1}^{n} (\hat{x}_{i,n} - 1)^2 1_{\{X_i \in S,\ |\hat{x}_{i,n}-1|<1/2\}}.
\end{aligned}$$

Note that whenever $X_i \in S$, because it holds $f_{0,i}|\hat{x}_{i,n} - 1| \le \epsilon_n$, we have (for n small enough) that $|\hat{x}_{i,n} - 1| < 1/2$. This means that, for all $i = 1, \dots, n$, $1_{\{X_i \in S\}} \le 1_{\{|\hat{x}_{i,n}-1|<1/2\}}$, and it follows

$$\begin{aligned}
&\tilde{L}_n(\alpha) - (\alpha M_n^{(p)} + (1-\alpha)M_n^{\mathrm{np}}) \\
&\le -\frac{1}{4} \sum_{i=1}^{n} (\hat{x}_{i,n} - 1)^2 1_{\{X_i \in S\}} \\
&= -\frac{1}{4} \{(1-\alpha)^2 Q_n^{\mathrm{np}}(S) + \alpha^2 Q_n^{(p)}(S) + 2\alpha(1-\alpha)U_n\} \\
&\le -\frac{1}{4}(1-\alpha)^2 Q_n^{\mathrm{np}}(S) \left\{ 1 - \frac{2\alpha|U_n|}{(1-\alpha)Q_n^{\mathrm{np}}(S)} \right\},
\end{aligned}$$

where

$$U_n = \sum_{i=1}^{n} \frac{(f_{\hat{\theta}_n,i} - f_{0,i})(\xi_{n,i} - f_{0,i})}{f_{0,i}^2} 1_{\{X_i \in S\}}.$$

Bounding the right-hand side with respect to $\alpha \in [0, 1-\epsilon]$ gives

$$\begin{aligned}
&\sup_{\alpha \in [0,1-\epsilon]} \{\tilde{L}_n(\alpha) - (\alpha M_n^{(p)} + (1-\alpha)M_n^{\mathrm{np}})\} \\
&\quad \le -\frac{1}{4}\epsilon^2 Q_n^{\mathrm{np}}(S) \left(1 - 2\epsilon^{-1} \frac{|U_n|}{Q_n^{\mathrm{np}}(S)} \right).
\end{aligned}$$

By assumption, we have that $Q_n^{(p)}(S)/Q_n^{\mathrm{np}}(S) \to 0$ in probability. From the Cauchy–Schwartz inequality we get that $|U_n| \le \sqrt{Q_n^{\mathrm{np}}(S)Q_n^{(p)}(S)}$, leading to $|U_n|/Q_n^{\mathrm{np}}(S) \to 0$, in probability. Consequently, we obtain that

$$\sup_{\alpha \in [0,1-\epsilon]} \tilde{L}_n(\alpha) \le -\frac{1}{4}\epsilon^2 Q_n^{\mathrm{np}}(S) \left(1 - 2\epsilon^{-1} \frac{|U_n|}{Q_n^{\mathrm{np}}(S)} - 4 \frac{|M_n^{(p)}| + |M_n^{\mathrm{np}}|}{\epsilon^2 Q_n^{\mathrm{np}}(S)} \right).$$

The term between brackets goes to 1, in probability, implying that for every $\delta > 0$, with probability going to 1,

$$\sup_{\alpha \in [0,1-\epsilon]} \tilde{L}_n(\alpha, \hat{\theta}) \le -\frac{1}{4}\epsilon^2 Q_n^{\mathrm{np}}(S)(1-\delta).$$

Hence it remains to note that, by (B2), with probability going to 1, $Q_n^{\mathrm{np}}(S) > 0$.

Proof of Theorem 3

Note that $\hat{\alpha}_n \in \mathrm{argmax}_{\alpha\in[0,1]}\tilde{L}_n(\alpha)$ exists because $\xi_{n,i} > 0$ for all i, as explained in the proof of Proposition 1. Let $\epsilon > 0$. The proof requires to show that with probability going to 1, $\hat{\alpha}_n < \epsilon$. This event is realized as soon as $\max_{\alpha\in[\epsilon,1]}\tilde{L}_n(\alpha) < \tilde{L}_n(0)$. We analyze both terms separately. First we show that

$$\tilde{L}_n(0) \to 0,$$

in probability, and then that there exists $\delta > 0$ such that, with probability going to 1,

$$\sup_{\alpha\in[\epsilon,1]} \tilde{L}_n(\alpha) \le -\delta. \tag{B4}$$

Let $\eta > 0$, $b_n = (\eta/|\log(\Delta_n)|)^{1/\beta}$ and $c_n = \max_{i=1,\dots,n}|\xi_{n,i} - f_{0,i}|$. We assume further that $b_n + c_n < 1$ and $\Delta_n < 1$. We have

$$\begin{aligned}
&|\tilde{L}_n(0)| \\
&\le \left| n^{-1}\sum_{i=1}^{n} \log\left(\frac{\xi_{n,i}}{f_{0,i}}\right) 1_{\{f_{0,i}>b_n\}} \right| + n^{-1}\sum_{i=1}^{n} \left|\log\left(\frac{\xi_{n,i}}{f_{0,i}}\right)\right| 1_{\{f_{0,i}\le b_n\}} \\
&\le \left| n^{-1}\sum_{i=1}^{n} \log\left(\frac{\xi_{n,i}}{f_{0,i}}\right) 1_{\{f_{0,i}>b_n\}} \right| \\
&\quad + n^{-1}\sum_{i=1}^{n}(|\log(\Delta_n q_i)|1_{\{f_{0,i}\le b_n\}} + |\log(f_{0,i})|1_{\{f_{0,i}\le b_n\}}) \\
&\le \left| n^{-1}\sum_{i=1}^{n} \log\left(\frac{\xi_{n,i}}{f_{0,i}}\right) 1_{\{f_{0,i}>b_n\}} \right| + |\log(\Delta_n)| n^{-1}\sum_{i=1}^{n} 1_{\{f_{0,i}\le b_n\}} \\
&\quad + n^{-1}\sum_{i=1}^{n}(|\log(q_i)| + |\log(f_{0,i})|)1_{\{f_{0,i}\le b_n\}}.
\end{aligned}$$

The expectation of the term in the right is smaller than $\mathbb{E}[(|\log(q_1)| + |\log(f_{0,1})|)1_{\{f_{0,i}\le b_n\}}]$, which goes to 0 because $|\log(q_1)|$ and $|\log(f_{0,1})|$ are integrable. For the term in the left, the mean-value theorem gives

$$\begin{aligned}
\left| n^{-1}\sum_{i=1}^{n} \log\left(\frac{\xi_{n,i}}{f_{0,i}}\right) 1_{\{f_{0,i}>b_n\}} \right| &\le \frac{\max_{i=1,\dots,n} |\xi_{n,i} - f_{0,i}|}{\inf_{i=1,\dots,n\ :\ f_{0,i}>b_n} \inf_{t\in[0,1]} t\xi_{n,i} + (1-t)f_{0,i}} \\
&\le \frac{c_n}{b_n - c_n} \to 0.
\end{aligned}$$

Conclude remarking that the expectation of the term in the middle is bounded by $|\log(\Delta_n)|\mathbb{P}(f_0(X_1) \le b_n)$. This is of order $|\log(\Delta_n)|b_n^{\beta} = \eta$, by assumption, but η is arbitrarily small.

Now we establish (B4) by obtaining one-sided inequalities. Consider $b_n = (1/|\log(\Delta_n)|)^{1/\beta}$, suppose that $b_n + c_n < 1$, and use the monotonicity of the logarithm, to get that

$$
\begin{aligned}
& n^{-1}\sum_{i=1}^{n}\log\left(\frac{\alpha f_{\hat\theta_n,i}+(1-\alpha)\xi_{n,i}}{f_{0,i}}\right)1_{\{f_{0,i}\le b_n\}}\\
& \le n^{-1}\sum_{i=1}^{n}\log\left(\frac{F_{\Theta,i}+b_n+c_n}{f_{0,i}}\right)1_{\{f_{0,i}\le b_n\}}\\
& \le n^{-1}\sum_{i=1}^{n}|\log\left(\frac{F_{\Theta,i}+1}{f_{0,i}}\right)|1_{\{f_{0,i}\le b_n\}}.
\end{aligned}
$$

Taking the expectation, we find a bound in $\mathbb{E}[|\log\left(\frac{F_{\Theta,1}+1}{f_{0,1}}\right)|1_{\{f_{0,1}\le b_n\}}]$ which goes to 0 as $n\to\infty$ in virtue of the Lebesgue dominated convergence theorem. Let $\eta\in(0,1)$. It holds that

$$
\begin{aligned}
\tilde L_n(\alpha) & \le n^{-1}\sum_{i=1}^{n}\log\left(\frac{\alpha f_{\hat\theta_n,i}+(1-\alpha)\xi_{n,i}}{f_{0,i}}\right)1_{\{f_{0,i}>b_n\}}+o_p(1)\\
& \le n^{-1}\sum_{i=1}^{n}\log\left(\frac{\alpha(f_{\hat\theta_n,i}+\eta f_{0,i})+(1-\alpha)\xi_{n,i}}{f_{0,i}}\right)1_{\{f_{0,i}>b_n\}}+o_p(1).
\end{aligned}
$$

The first term in the right-hand side is decomposed according to

$$
\begin{aligned}
& n^{-1}\sum_{i=1}^{n}\log\left(\frac{\alpha(f_{\hat\theta_n,i}+\eta f_{0,i})+(1-\alpha)\xi_{n,i}}{\alpha(f_{\hat\theta_n,i}+\eta f_{0,i})+(1-\alpha)f_{0,i}}\right)1_{\{f_{0,i}>b_n\}}\\
& +n^{-1}\sum_{i=1}^{n}\log\left(\frac{\alpha(f_{\hat\theta_n,i}+\eta f_{0,i})+(1-\alpha)f_{0,i}}{f_{0,i}}\right)1_{\{f_{0,i}>b_n\}}.
\end{aligned}
$$

By the mean value theorem, the term on the left is bounded by

$$
\frac{(1-\alpha)\max_{i=1,\dots,n}|\xi_{n,i}-f_{0,i}|}{(\eta b_n)\wedge(b_n-c_n)},
$$

which goes to 0, by assumption. For the term on the right, notice that $\{\alpha(f_\theta+\eta f_0)+(1-\alpha)f_0\ :\ \alpha\in[\epsilon,1],\ \theta\in\Theta\}$ is Glivenko–Cantelli with envelop $F_\Theta+2f_0$. Then applying theorem 3 in van der Vaart and Wellner (2000), the class formed by $\log(\alpha(f_\theta+\eta f_0)+(1-\alpha)f_0)$ is still Glivenko–Cantelli. Since for all $\theta\in\Theta$, $\alpha\in[\epsilon,1]$,

$$
\log(\epsilon\eta f_0)\le\log(\alpha(f_\theta+\eta f_0)+(1-\alpha)f_0)\le\log(F_\Theta+2f_0+1),
$$

the function $|\log(\epsilon\eta f_0)|+\log(F_\Theta+2f_0+1)$ is an integrable envelop. Using again theorem 3 in van der Vaart and Wellner (2000), the class formed by $\log(\alpha(f_\theta+\eta f_0)+(1-\alpha)f_0)1_{\{f_0>b\}}$, $\theta\in\Theta$, $\alpha\in[\epsilon,1]$, $0<b<1$, is still Glivenko–Cantelli with the same envelop. This implies that

$$
\begin{aligned}
\sup_{\alpha\in[\epsilon,1],\ \theta\in\Theta}\Bigg| & n^{-1}\sum_{i=1}^{n}\log\left(\frac{\alpha(f_{\theta,i}+\eta f_{0,i})+(1-\alpha)f_{0,i}}{f_{0,i}}\right)1_{\{f_{0,i}>b_n\}}\\
& -\mathbb{E}\left[\log\left(\frac{\alpha(f_{\theta,1}+\eta f_{0,1})+(1-\alpha)f_{0,1}}{f_{0,1}}\right)1_{\{f_{0,1}>b_n\}}\right]\Bigg|\to 0.
\end{aligned}
$$

The integrability of the envelop and the fact that $b_n \to 0$ implies that

$$\sup_{\alpha\in[\epsilon,1],\ \theta\in\Theta} \mathbb{E}\left[\log\left(\frac{\alpha(f_{\theta,1}+\eta f_{0,1})+(1-\alpha)f_{0,1}}{f_{0,1}}\right)1_{\{f_{0,1}\le b_n\}}\right] \to 0.$$

It remains to use the inequality $\log(x) \le 2(\sqrt{x}-1)$ to obtain that

$$\sup_{\alpha\in[\epsilon,1],\ \theta\in\Theta} \mathbb{E}\left[\log\left(\frac{m_{\theta,\alpha,\eta}}{f_{0,1}}\right)\right] \le \sup_{\alpha\in[\epsilon,1],\ \theta\in\Theta} 2\int(\sqrt{m_{\theta,\alpha,\eta}f_0}-f_0)\,\mathrm{d}\lambda,$$

where $m_{\theta,\alpha,\eta} = \alpha(f_\theta+\eta f_0)+(1-\alpha)f_0$. Since $(m_{\theta,\alpha,\eta}-m_{\theta,\alpha,0})f_0 = \eta f_0^2$ and using that $\sqrt{a+b} \le \sqrt{a}+\sqrt{b}$, $a \ge 0$, $b \ge 0$, we obtain

$$\begin{aligned}\sup_{\alpha\in[\epsilon,1],\ \theta\in\Theta} \mathbb{E}\left[\log\left(\frac{m_{\theta,\alpha,\eta}}{f_{0,1}}\right)\right] &\le 2\sqrt{\eta}+2\sup_{\alpha\in[\epsilon,1],\ \theta\in\Theta}\int(\sqrt{m_{\theta,\alpha,0}f_0}-f_0)\,\mathrm{d}\lambda\\ &= 2\sqrt{\eta}-\inf_{\alpha\in[\epsilon,1],\ \theta\in\Theta}\int(\sqrt{m_{\theta,\alpha,0}}-\sqrt{f_0})^2\,\mathrm{d}\lambda.\end{aligned}$$

Using standard results about the Hellinger distance (Pollard, 2000, chapter 3) we obtain

$$\begin{aligned}\sup_{\alpha\in[\epsilon,1],\ \theta\in\Theta} \mathbb{E}\left[\log\left(\frac{m_{\theta,\alpha,\eta}}{f_{0,1}}\right)\right] &\le 2\sqrt{\eta}-(1/4)\inf_{\alpha\in[\epsilon,1],\ \theta\in\Theta}\alpha^2\left(\int|f_\theta-f_0|\,\mathrm{d}\lambda\right)^2\\ &\le 2\sqrt{\eta}-(\epsilon^2/4)\left(\inf_{\theta\in\Theta}\int|f_\theta-f_0|\,\mathrm{d}\lambda\right)^2.\end{aligned}$$

Since $f_0 \notin \mathcal{P}$ and by the continuity assumption on f_θ, it holds that $\inf_{\theta\in\Theta}\int|f_\theta-f_0|\mathrm{d}\lambda > 0$. Then, as η is arbitrary, the proof of (B4) is complete.

B.4 Linear and quadratic error of parametric and nonparametric estimate

Important tools for dealing with the terms involving $\hat{f}_{n,i}^{\mathrm{LR}}$ are coming from U-statistic theory. We call U-statistic of order p with kernel $w : \mathbb{R}^p \to \mathbb{R}$, any quantity of the kind

$$\sum_{i_1,\dots,i_p\in D} w(X_{i_1},\dots,X_{i_p}),$$

where the summation is taken over the subset D formed by the $(i_1,\dots,i_p)\in\{1,\dots,n\}^p$ such that $i_k \ne i_\ell$, $\forall k \ne \ell$. The number of terms in the summation is then $n(n-1)\dots(n-p+1)$. When the kernel w is such that, for every $k\in\{1,\dots,p\}$, $\mathbb{E}[w(X_1,\dots,X_p)|X_1,\dots X_{k-1},X_{k+1},\dots,X_p]=0$, it is called a degenerate U-statistic. In the proofs, we shall rely on the so-called Hajek decomposition (van der Vaart, 1998, lemma 11.11).

To establish the two following lemmas, Lemmas 1 and 2, we are based on (H1), (H2), and (H3). One might note that the expressions (a) or (b) in (H2) on the kernel are not used in any of these lemmas.

Lemma 1. *Under assumptions (H1), (H2), and (H3), if $nh_n^d\Delta_n \to 0$, for any $\delta > 0$ and any set $S \subset \mathbb{R}^d$ such that* $\inf_{x\in S} f_0(x) > b$*, we have with probability going to 1,*

$$h_n^d \sum_{i=1}^{n} \left(\frac{\hat{f}_{n,i}^{\mathrm{LR}} - f_{0,i}}{f_{0,i}} \right)^2 1_{\{X_i \in S\}} \geq (1-\delta) v_K \lambda(S),$$

where $v_K = \int K(u)^2 \, \mathrm{d}u$.

Proof. Note that

$$\begin{aligned} &\mathbb{E}[\hat{f}_{n,i}^{\mathrm{LR}} | X_i] \\ &= (n-1)^{-1} \sum_{j \neq i}^{n} \mathbb{E}\left[h_n^{-d} K\left(\frac{X_i - X_j}{h_n} \right) | X_i \right] + \Delta_n q_i = f_{h_n,i} + \Delta_n q_i. \end{aligned}$$

The proof follows from the decomposition

$$\sum_{i=1}^{n} \left(\frac{\hat{f}_{n,i}^{\mathrm{LR}} - f_{0,i}}{f_{0,i}} \right)^2 1_{\{X_i \in S\}} = A_n + B_n + 2C_n,$$

where

$$\begin{aligned} A_n &= \sum_{i=1}^{n} \left(\frac{\hat{f}_{n,i}^{\mathrm{LR}} - \mathbb{E}[\hat{f}_{n,i}^{\mathrm{LR}} | X_i]}{f_{0,i}} \right)^2 1_{\{X_i \in S\}}, \\ B_n &= \sum_{i=1}^{n} \left(\frac{f_{h_n,i} + \Delta_n q_i - f_{0,i}}{f_{0,i}} \right)^2 1_{\{X_i \in S\}}, \\ C_n &= \sum_{i=1}^{n} \frac{(\hat{f}_{n,i}^{\mathrm{LR}} - \mathbb{E}[\hat{f}_{n,i}^{\mathrm{LR}} | X_i])(f_{h_n,i} + \Delta_n q_i - f_{0,i})}{f_{0,i}^2} 1_{\{X_i \in S\}}. \end{aligned}$$

We will show that $h_n^d A_n \to v_K \lambda(S)$, in probability and that $h_n^d C_n \to 0$, in probability. This will be enough as $B_n \geq 0$, almost surely.

Proof that $h_n^d A_n \to v_K \lambda(S)$ in probability Introduce the notation, for any $h > 0$,

$$\begin{aligned} a_h(x,y) &= \frac{K_h(x-y) - f_h(x)}{f_0(x)}, \\ u_h(x,y,z) &= a_h(x,y) a_h(x,z) 1_{\{x \in S\}}. \end{aligned}$$

Developing, we find

$$A_n = (n-1)^{-2} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \sum_{k \neq i}^{n} u_{h_n}(i,j,k),$$

where $u_{h_n}(i,j,k)$ is as short-cut for $u_{h_n}(X_i, X_j, X_k)$. We treat A_n relying on the Hajek projection of U-statistics. Up to a centering term, which is $\mathbb{E}[u_{h_n}(i,j,k)|X_j, X_k]$, the U-statistic A_n is a degenerate U-statistic. In the following we voluntary introduce this centering term in the summation to handle separately a degenerate U-statistic and another summation with less indices. By introducing, for any $h > 0$,

$$v_h(j,k) = \mathbb{E}[u_h(i,j,k)|X_j, X_k],$$
$$w_h(i,j,k) = u_h(i,j,k) - v_h(j,k),$$

we obtain

$$\begin{aligned} A_n &= (n-1)^{-2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\sum_{k\neq i}^{n} w_{h_n}(i,j,k) + (n-1)^{-2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\sum_{k\neq i}^{n} v_{h_n}(j,k) \\ &= (n-1)^{-2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\sum_{k\neq i,\ k\neq j}^{n} w_{h_n}(i,j,k) + (n-1)^{-2}\sum_{i=1}^{n}\sum_{j\neq i}^{n} w_{h_n}(i,j,j) \\ &\quad + (n-1)^{-2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\sum_{k\neq i}^{n} v_{h_n}(j,k). \end{aligned} \tag{B5}$$

Treatment of the first term in (B5). Note that $w_{h_n}(i,j,k)$ defines a degenerate U-statistic, that is,

$$\mathbb{E}[w_{h_n}(i,j,k)|X_i, X_j] = \mathbb{E}[w_{h_n}(i,j,k)|X_i, X_k] = \mathbb{E}[w_{h_n}(i,j,k)|X_j, X_k] = 0.$$

Note that

$$\sum_{i=1}^{n}\sum_{j\neq i}^{n}\sum_{k\neq i,\ k\neq j}^{n} w_{h_n}(i,j,k) = \sum_{i=1}^{n}\sum_{j>i}^{n}\sum_{k>j}^{n} \overline{w}_{h_n}(i,j,k),$$

where $\overline{w}_h$ is the symmetrized version of w_h, that is, for any triplet (x_1, x_2, x_3) of $\overline{w}_h(x_1, x_2, x_3) = \sum_\sigma w_h(x_{\sigma(1)}, x_{\sigma(2)}, x_{\sigma(3)})$ where the sum is over all the 3! possible permutations of the set $\{1, 2, 3\}$. Using that the U-statistic with kernel $\overline{w}_{h_n}$ is degenerate, some algebra gives that

$$\begin{aligned} &\mathbb{E}\left[\left((n-1)^{-2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\sum_{k\neq i,\ k\neq j}^{n} w_{h_n}(i,j,k)\right)^2\right] \\ &= (n-1)^{-4}\sum_{i=1}^{n}\sum_{j>i}^{n}\sum_{k>j}^{n}\mathbb{E}[\overline{w}_{h_n}(i,j,k)^2] \\ &= O(n^{-1})\mathbb{E}[\overline{w}_{h_n}(1,2,3)^2]. \end{aligned}$$

We have, using Minkowski's inequality and the definition of the conditional expectation, that

$$\sqrt{E[\overline{w}_{h_n}(1,2,3)^2]} \leq 3!\sqrt{E[w_{h_n}(1,2,3)^2]} \leq 3!\sqrt{E[u_{h_n}(1,2,3)^2]}.$$

Consequently, in virtue of (B10) in Lemma 5, we have shown that

$$\mathbb{E}\left[\left((n-1)^{-2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\sum_{k\neq i,\ k\neq j}^{n} w_{h_n}(i,j,k)\right)^2\right] = O(n^{-1}h_n^{-2d}).$$

The previous rate, multiplied by h_n^{2d}, goes to 0, hence, this term is negligible.

Treatment of the second term in (B5). We continue the study of A_n by considering

$$
\begin{aligned}
&(n-1)^{-2}\sum_{i=1}^{n}\sum_{j\neq i}^{n} w_{h_n}(i,j,j) \\
&\quad = (n-1)^{-2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}(w_{h_n}(i,j,j) - \mathbb{E}[u_{h_n}(i,j,j)|X_i] + \mathbb{E}[u_{h_n}(1,2,2)]) \\
&\qquad + (n-1)^{-1}\sum_{i=1}^{n}(\mathbb{E}[u_{h_n}(i,j,j)|X_i] - \mathbb{E}[u_{h_n}(1,2,2)]).
\end{aligned}
$$

The first term is a degenerate U-statistic of order 2 whose order 2 moments satisfy

$$
\begin{aligned}
&\mathbb{E}\left[\left((n-1)^{-2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}(w_{h_n}(i,j,j) - \mathbb{E}[u_{h_n}(i,j,j)|X_i] + \mathbb{E}[u_{h_n}(1,2,2)])\right)^2\right] \\
&\quad = O(n^{-2})\mathbb{E}[u_{h_n}(1,2,2)^2] = O(n^{-2}h_n^{-3d}).
\end{aligned}
$$

This is obtained by following exactly the same lines as in the treatment of the U-statistic w_n and using (B13) in Lemma 5. As $n^{-2}h_n^{-3d} \times h_n^{2d} \to 0$, the previous term is negligible. The second term is a sum of centered independent random variables with variance smaller than, in virtue of (B9) in Lemma 5,

$$
n(n-1)^{-2}\mathbb{E}[\mathbb{E}[u_{h_n}(1,2,2)|X_1]^2] = O(n^{-1}h_n^{-2d}).
$$

This is the same rate as the rate obtained for the (negligible) U-statistic of order 3 with kernel w_n.

Treatment of the third term in (B5). The study of A_n continues by considering

$$
\begin{aligned}
&(n-1)^{-2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\sum_{k\neq i}^{n} v_{h_n}(j,k) \\
&\quad = (n-1)^{-2}\sum_{i=1}^{n}\sum_{j\neq i}^{n} v_{h_n}(j,j) + (n-1)^{-2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\sum_{k\neq i,\ k\neq j}^{n} v_{h_n}(j,k) \\
&\quad = (n-1)^{-1}\sum_{j=1}^{n} v_{h_n}(j,j) + (n-1)^{-2}(n-2)\sum_{j=1}^{n}\sum_{k\neq j}^{n} v_{h_n}(j,k).
\end{aligned}
$$

The term associated with double summation over j and k is a degenerate U-statistic, as $\mathbb{E}[v_{h_n}(j,k)|X_k] = \mathbb{E}[v_{h_n}(j,k)|X_j] = 0$. Consequently, following the same lines as in the treatment of the first term of A_n, and using (B14) in Lemma 5, we get

$$
\mathbb{E}\left[\left((n-1)^{-2}(n-2)\sum_{j=1}^{n}\sum_{k\neq j}^{n} v_{h_n}(j,k)\right)^2\right] = O(1)\mathbb{E}[v_{h_n}(1,2)^2] = O(h_n^{-d}),
$$

which goes to 0, when multiplied by h_n^{2d}. The remaining term is a sum of independent and identically distributed random variables. We have, by computing the variance of the centered average,

$$(n-1)^{-1}\sum_{j=1}^{n}(v_{h_n}(j,j)=O_{\mathbb{P}}(n^{-1/2})\sqrt{\mathbb{E}v_{h_n}(1,1)^2}+n(n-1)^{-1}\mathbb{E}[v_{h_n}(1,1)],$$

where the first term, using (B12) in Lemma 5, is $O(n^{-1/2}h_n^{-d})$ which goes to 0 when multiplied by h_n. The dominating term is in fact the last one, as by Lemma 6, it holds that $h_n^d\mathbb{E}[v_{h_n}(1,1)]\to v_K\lambda(S)$.

Proof that $h_n^dC_n\to 0$ in probability We are based on similar decompositions as for A_n involving U-statistics. Let $\ell_{h_n}(x)=(f_{h_n}(x)-f_0(x)+\Delta_n q(x))/f_0(x)$ and note that in virtue of Lemma 7, it holds

$$\|\ell_{h_n}\|_S\le b^{-1}\left(h_n^2\|g\|_{\mathbb{R}^d}\int\|u\|_2^2K(u)\,\mathrm{d}u+\Delta_n\|q\|_{\mathbb{R}^d}\right).$$

Then

$$C_n=(n-1)^{-1}\sum_{i=1}^{n}\sum_{j\neq i}^{n}(a_{h_n}(i,j)\ell_{h_n,i}1_{\{X_i\in S\}}-b_{h_n}(j))+\sum_{i=1}^{n}b_{h_n}(i),$$

with $b_{h_n}(j)=\mathbb{E}[a_{h_n}(i,j)\ell_{h_n,i}1_{\{X_i\in S\}}|X_j]$. The term on the left is a degenerate U-statistic for which it holds

$$\mathbb{E}\left[\left((n-1)^{-1}\sum_{i=1}^{n}\sum_{j\neq i}^{n}(a_{h_n}(i,j)\ell_{h_n,i}1_{\{X_i\in S\}}-b_{h_n}(j))\right)^2\right]$$
$$=O(1)\mathbb{E}[a_{h_n}(1,2)^2\ell_{h_n,1}^2 1_{\{X_1\in S\}}].$$

Using (B8) in Lemma 5 and the previous bound for $\|\ell_{h_n}\|_S$, we find

$$\mathbb{E}[a_{h_n}(1,2)^2\ell_{h_n,1}^2 1_{\{X_1\in S\}}]=\mathbb{E}\left[\left(\frac{V_{h_n}(X_1)\ell_{h_n,1}^2}{f_{0,1}^2}\right)1_{\{X_1\in S\}}\right]$$
$$=O(h_n^{-d}(h_n^2+\Delta_n)),$$

where V_h is defined in (B7). The previous bound multiplied by h_n^{2d} goes to 0. Using that $\mathbb{E}[a_{h_n}(1,2)\ell_{h_n,1}1_{\{X_1\in S\}}]=0$ and (B11), the variance of the term on the right in C_n is smaller than

$$n\mathbb{E}[\mathbb{E}[a_{h_n}(1,2)\ell_{h_n,1}1_{\{X_1\in S\}}|X_2]^2]\le n\mathbb{E}[\mathbb{E}[|a_{h_n}(1,2)||X_2]^2]\|\ell_{h_n}\|_S^2$$
$$=O(n(h_n^4+\Delta_n^2)),$$

which, multiplied by h_n^{2d}, goes to 0 by hypothesis. Hence $h_n^dC_n\to 0$, in probability and the proof is complete. ■

Lemma 2. *Under Assumptions (H1), (H2), and (H3), if $nh_n^d\Delta_n\to 0$, we have*

$$h_n^d\sum_{i=1}^{n}\left(\frac{\hat{f}_{n,i}^{\mathrm{LR}}-f_{0,i}}{f_{0,i}}\right)=o_{\mathbb{P}}(1).$$

Proof. The decomposition is as follows

$$
\begin{aligned}
h_n^d \sum_{i=1}^n \left(\frac{\hat{f}_{n,i}^{\mathrm{LR}} - f_{0,i}}{f_{0,i}} \right) \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(B6)}\\
= h_n^d (n-1)^{-1} \sum_{i=1}^n \sum_{j \neq i}^n \left(\frac{K_{h_n}(i,j) - f_{h_n,i}}{f_{0,i}} \right) + h_n^d \sum_{i=1}^n \left(\frac{f_{h_n,i} - f_{0,i}}{f_{0,i}} \right) \\
+ h_n^d \Delta_n \sum_{i=1}^n \frac{q_i}{f_{0,i}}.
\end{aligned}
$$

The expectation of the last term is $nh_n^d \Delta_n \int q(x)\, \mathrm{d}x$ which goes to 0 by assumption. We can now focus on the first and second term of the decomposition.

Treatment of the second term in (B6). Using that $\int K(u)\, \mathrm{d}u = 1$, the considered term is a centered empirical sum. Using Lemma 7, its variance is then bounded by

$$
\begin{aligned}
\mathbb{E}\left[\left(h_n^d \sum_{i=1}^n \frac{f_{h_n,i} - f_{0,i}}{f_{0,i}} \right)^2 \right] &\leq nh_n^{2d} \int \frac{(f_{h_n}(x) - f_0(x))^2}{f_0(x)}\, \mathrm{d}x \\
&\leq nh_n^{2d+4} \int \frac{g(x)^2}{f_0(x)}\, \mathrm{d}x \left(\int u^2 K(u)\, \mathrm{d}u \right)^2,
\end{aligned}
$$

which goes to 0.

Treatment of the first term in (B6). Using that $\int K(u)\, \mathrm{d}u = 1$, one can verify that it is a degenerate U-statistic. Here the variance cannot be computed directly because the leading term $\mathbb{E}\left[\frac{(K_{h_n}(1,2) - f_{h_n,1})^2}{f_{0,1}^2} \right]$ is not necessarily finite. Hence we decompose according to the X_i in S_{b_n} and the others, with $b_n = (\epsilon / nh_n^d)^{1/\beta}$ where β is given in (H3) and $\epsilon > 0$. We introduce

$$
k(x,y) = \frac{K_{h_n}(x,y)}{f_0(x)},
$$

and define the linear operator $Q_{\mathbb{P}} : L_2(\mathbb{P}) \to L_2(\mathbb{P})$ as

$$
Q_{\mathbb{P}}[w](x,y) = w(x,y) - \mathbb{E}[w(x, X_1)] - \mathbb{E}[w(X_1, y)] + \mathbb{E}[w(X_1, X_2)].
$$

Because $\mathbb{E}[k(X_1, y)] = \mathbb{E}[k(X_1, X_2)] = 1$ for all $y \in \mathbb{R}^d$, one sees that

$$
\begin{aligned}
\sum_{i=1}^n \sum_{j \neq i}^n \left(\frac{K_{h_n}(i,j) - f_{h_n,i}}{f_{0,i}} \right) &= \sum_{i=1}^n \sum_{j \neq i}^n Q_{\mathbb{P}}(k)_{i,j} \\
&= \sum_{i=1}^n \sum_{j \neq i}^n (Q_{\mathbb{P}}(k 1_{S_{b_n}})_{i,j} + Q_{\mathbb{P}}(k 1_{S_{b_n}^c})_{i,j}).
\end{aligned}
$$

Because the summation over $Q_{\mathbb{P}}(k 1_{S_{b_n}})$ is a degenerate U-statistics, we get that

$$\mathbb{E}\left[\left(h_n^d(n-1)^{-1}\sum_{i=1}^{n}\sum_{j\neq i}^{n}Q_{\mathbb{P}}(k1_{S_{b_n}})_{i,j}\right)^2\right]$$
$$= O(h_n^{2d})\mathbb{E}\left[\frac{(K_{h_n}(1,2)-f_{h_n,1})^2}{f_{0,1}^2}1_{S_{b_n}}(X_1)\right].$$

Defining the kernel $\tilde{K}=K^2/v_K$ and $\tilde{f}_h=f_0\star\tilde{K}_h$, we obtain

$$\mathbb{E}\left[\frac{(K_{h_n}(1,2)-f_{h_n,1})^2}{f_{0,1}^2}1_{S_{b_n}}(X_1)\right]$$
$$\le \mathbb{E}\left[\frac{\mathbb{E}[K_{h_n}(1,2)^2|X_1]}{f_{0,1}^2}1_{S_{b_n}}(X_1)\right]$$
$$= v_K h_n^{-d}\int_{S_{b_n}}\frac{\tilde{f}_{h_n}(x)}{f_0(x)}\,\mathrm{d}x$$
$$= v_K h_n^{-d}\left(\int_{S_{b_n}}\int\frac{f(x-h_nu)}{f_0(x)}\tilde{K}(u)\mathrm{d}u\,\mathrm{d}x\right)$$
$$\le v_K h_n^{-d}\lambda(S_{b_n})\left(\sup_{x\in S_{b_n}}\sup_{u\in[-1,1]^d}\frac{f(x+h_nu)}{f_0(x)}\right).$$

For the term with $Q_{\mathbb{P}}(k1_{S^c_{b_n}})$, we obtain that

$$\mathbb{E}\left[\left|\sum_{i=1}^{n}\sum_{j\neq i}^{n}Q_{\mathbb{P}}(k1_{S^c_{b_n}})_{i,j}\right|\right]\le n(n-1)\mathbb{E}[|Q_{\mathbb{P}}(k1_{S^c_{b_n}})_{1,2}|]$$
$$\le 4n(n-1)\mathbb{E}[|k_{1,2}|1_{S^c_{b_n}}(X_1)]$$
$$= 4n(n-1)\int_{S^c_{b_n}}f_h(x)\,\mathrm{d}x.$$

From Lemma 8, we deduce that $\int_{S^c_{b_n}}f_h(x)\,\mathrm{d}x\le c_2b_n^{\beta}=c_2\epsilon/nh_n^d$. To conclude, we have shown that there exists a constant $\tilde{C}>0$ such that

$$\mathbb{E}\left[\left|h_n^d(n-1)^{-1}\sum_{i=1}^{n}\sum_{j\neq i}^{n}Q_{\mathbb{P}}(k)\right|\right]\le\tilde{C}(\sqrt{h_n^d\lambda(S_{b_n})}+nh_n^db_n^{\beta})$$
$$=\tilde{C}(\sqrt{h_n^d\lambda(S_{b_n})}+\epsilon).$$

Invoking (H3) and because ϵ is arbitrarily small, the limit as $n\to\infty$ is 0. ■

Lemma 3. *Under (A1) and (A2), we have*

$$\sum_{i=1}^{n}\left(\frac{f_{\hat{\theta}_n,i}-f_{0,i}}{f_{0,i}}\right)^2 1_{\{X_i\in S\}}=O_{\mathbb{P}}(1).$$

Proof. Using (C2), we have that, with probability going to 1,

$$\begin{aligned}
&\sum_{i=1}^{n}\left(\frac{f_{\hat\theta_n,i}-f_{0,i}}{f_{0,i}}\right)^2 1_{\{X_i\in S\}}\\
&\le \left(n^{-1}\sum_{i=1}^{n}\frac{\dot\ell(X_i)^2\sup\limits_{\theta\in B(\theta_0,\delta)} f_\theta(X_i)^2}{f_{0,i}^2}1_{\{X_i\in S\}}\right) n\|\hat\theta_n-\theta_0\|_2^2\\
&\le \|\dot\ell\sup_{\theta\in B(\theta_0,\delta)} f_\theta\|_{\mathbb{R}^d}^2 b^{-2} n\|\hat\theta_n-\theta_0\|_2^2,
\end{aligned}$$

which is a tight sequence because of (C3). ■

Lemma 4. *Under (A1) and (A2), we have*

$$\sum_{i=1}^{n}\left(\frac{f_{\hat\theta_n,i}-f_{0,i}}{f_{0,i}}\right)=O_{\mathbb{P}}(1).$$

Proof. In virtue of (A2), the map $\theta\mapsto\log f_\theta(x)$ is differentiable at θ_0, for P-almost every $x\in\mathbb{R}^d$ with derivative $\dot\ell_{\theta_0}(x)$ (this is obtained in van der Vaart (1998), in the proof of theorem 5.39). Using stability properties for the composition, the map $\theta\mapsto f_\theta(x)=\exp(\log(f_\theta(x)))$ is differentiable at θ_0, for P-almost every $x\in\mathbb{R}^d$ with derivative $\dot\ell_{\theta_0}(x)f_0(x)$. We are in position to apply lemma 19.31 in van der Vaart (1998), with $r_n=\sqrt{n}$ and $m_\theta=f_\theta/f_0$. From the mentioned lemma, as $\sqrt{n}(\hat\theta_n-\theta_0)$ is tight, defining

$$\begin{aligned}
T_i(\theta)&=\left[\left(\frac{f_{\theta,i}-f_{0,i}}{f_{0,i}}\right)-(\theta-\theta_0)^T\dot\ell_{\theta_0,i}\right],\\
t(\theta)&=\int[f_\theta(x)-f_0(x)-(\theta-\theta_0)^T\dot\ell_{\theta_0}(x)f_0(x)]\,\mathrm{d}x,
\end{aligned}$$

we obtain

$$\left|\sum_{i=1}^{n}\{T_i(\hat\theta_n)-t(\hat\theta_n)\}\right|=o_{\mathbb{P}}(1).$$

Actually, recalling that $\int\dot\ell_{\theta_0}(x)f_0(x)\,\mathrm{d}x=0$, we find that, for all $\theta\in\Theta$, $t(\theta)=0$. Hence, we obtain

$$\sum_{i=1}^{n}\left(\frac{f_{\hat\theta_n,i}-f_{0,i}}{f_{0,i}}\right)=n^{1/2}(\hat\theta_n-\theta_0)^T\left[n^{-1/2}\sum_{i=1}^{n}\dot\ell_{\theta_0,i}\right]+o_{\mathbb{P}}(1).$$

where the first term is a $O_{\mathbb{P}}(1)$. ■

B.5 Auxiliary results

Recall some definitions, for any $h>0$,

$$V_h(X_1)=\mathbb{E}[(K_h(1,2)-f_{h,1})^2|X_1],$$

$$a_h(x,y) = \frac{K_h(x-y) - f_h(x)}{f_0(x)},$$

$$u_h(x,y,z) = a_h(x,y)a_h(x,z)1_{\{x \in S\}}, \tag{B7}$$

as well as the short-cut $g(i,j,k)$ for $g(X_i, X_j, X_k)$.

Lemma 5. *Under (H1) and (H2), if $S \subset \mathbb{R}^d$ is such that for all $x \in S$, $f_0(x) > b > 0$, we have, for any $h > 0$,*

$$V_h(X_1) \leq h^d C_0, \tag{B8}$$

$$\mathbb{E}[\mathbb{E}[u_h(1,2,2)|X_1]^2] \leq h^{-2d} C_1, \tag{B9}$$

$$\mathbb{E}[u_h(1,2,3)^2] \leq h^{-2d} C_1, \tag{B10}$$

$$\mathbb{E}[|a_h(1,2)||X_2] \leq 2, \tag{B11}$$

$$\mathbb{E}[\mathbb{E}[u_h(1,2,2)|X_2]^2] \leq h^{-2d} C_2 + C_3, \tag{B12}$$

$$\mathbb{E}[u_h(1,2,2)^2] \leq h^{-3d} C_4 + C_5, \tag{B13}$$

$$\mathbb{E}[\mathbb{E}[u_h(1,2,3)|X_2,X_3]^2] \leq h^{-d} C_6 + C_6, \tag{B14}$$

where the constants C_k, $k = 0, \ldots, 7$ depend on K and f_0 only.

Proof. Remark that because K is bounded and $\int |K(u)| \,\mathrm{d}u < \infty$, we have $\int |K(u)|^k \,\mathrm{d}u < \infty$, for any $k \geq 1$. Note that, for every $h > 0$,

$$V_h(X_1) \leq \mathbb{E}[K_h(2,1)^2 | X_1] \leq h^{-d} v_K \|f_0\|_{\mathbb{R}^d}.$$

We obtain (B9) by writing

$$\mathbb{E}[\mathbb{E}[u_h(1,2,2)|X_1]^2] = \mathbb{E}\left[\frac{V_h(X_1)^2}{f_{0,1}^4} 1_{\{X_1 \in S\}}\right] \leq h^{-2d} v_K^2 \|f_0\|_{\mathbb{R}^d}^2 b^{-4}.$$

To establish (B10), note that

$$\mathbb{E}[u_h(1,2,3)^2] = \mathbb{E}\left[\frac{V_h(X_1)^2}{f_{0,1}^4} 1_{\{X_1 \in S\}}\right] = \mathbb{E}[\mathbb{E}[u_h(1,2,2)|X_1]^2].$$

For (B11), write

$$\mathbb{E}[|a_h(1,2)||X_2] = \int |K_h(x - X_2) - f_h(x)| \,\mathrm{d}x$$
$$\leq \int K_h(x - X_2) \,\mathrm{d}x + \int f_h(x) \,\mathrm{d}x = 2.$$

Inequality (B12) follows from the lines

$$
\begin{aligned}
\mathbb{E}[\mathbb{E}[u_h(1,2,2)|X_2]^2] &= \int \left(\int \frac{(K_h(x-y)-f_h(x))^2}{f_0(x)} 1_{\{x\in S\}}\ \mathrm{d}x \right)^2 f_0(y)\ \mathrm{d}y \\
&\le 2 \int \left(\int \frac{K_h(x-y)^2+f_h(x)^2}{f_0(x)} 1_{\{x\in S\}}\ \mathrm{d}x \right)^2 f_0(y)\ \mathrm{d}y \\
&\le 2b^{-2} \int \left(\int K_h(x-y)^2+f_h(x)^2\ \mathrm{d}x \right)^2 f_0(y)\ \mathrm{d}y \\
&\le 2b^{-2} \int (h^{-d}v_K + \|f_h\|_{\mathbb{R}^d})^2 f_0(y)\ \mathrm{d}y \\
&\le 4b^{-2}(h^{-2d}v_K^2 + \|f_0\|_{\mathbb{R}^d}^2).
\end{aligned}
$$

To show (B13), write

$$
\mathbb{E}[u_h(1,2,2)^2] = \mathbb{E}\left[\frac{(K_h(1,2)-f_{h,1})^4}{f_{0,1}^4} 1_{\{X_1\in S\}} \right].
$$

Using that $(a+b)^4 \le 8(a^4+b^4)$, we obtain

$$
\begin{aligned}
\mathbb{E}[u_h(1,2,2)^2] &\le 8\mathbb{E}\left[\left(\frac{K_h(1,2)^4+f_{h,1}^4}{f_{0,1}^4} \right) 1_{\{X_1\in S\}} \right] \\
&\le 8b^{-4}\left(h^{-3d}\|f_0\|_{\mathbb{R}^d} \int K(u)^4\ \mathrm{d}u + \|f_0\|_{\mathbb{R}^d}^4 \right).
\end{aligned}
$$

For (B14), we have

$$
\begin{aligned}
&\mathbb{E}[\mathbb{E}[u_h(1,2,3)|X_2,X_3]^2\ ] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(\frac{K_h(3,1)-f_{h,3}}{f_{0,3}} \right) \left(\frac{K_h(3,2)-f_{h,3}}{f_{0,3}} \right) 1_{\{X_3\in S\}} | X_1, X_2 \right]^2 \right].
\end{aligned}
$$

We develop and compute bounds for each term. The larger term will be the one associated with the product of the kernels. We have, by Jensen's inequality, for any $(y,z) \in \mathbb{R}^d \times \mathbb{R}^d$,

$$
\begin{aligned}
\psi_h(y,z) :&= \left(\int \left(\frac{K_h(x-y)K_h(x-z)}{f_0(x)} \right) 1_{\{x\in S\}}\ \mathrm{d}x \right)^2 \\
&\le b^{-2}\left(\int K_h(x-y)K_h(x-z)\ \mathrm{d}x \right)^2 \\
&= b^{-2}h^{-2d}\left(\int K(u)K((y-z)/h+u)\ \mathrm{d}u \right)^2 \\
&\le b^{-2}h^{-2d} \int K(u)K((y-z)/h+u)^2\ \mathrm{d}u.
\end{aligned}
$$

Then we obtain

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left[\left(\frac{K_h(3,1)}{f_{0,3}}\right)\left(\frac{K_h(3,2)}{f_{0,3}}\right)1_{\{X_3\in S\}}|X_1,X_2\right]^2\right]\\
&= \int\int \psi_h(y,z)f_0(y)f_0(z)\ \mathrm{d}y\mathrm{d}z\\
&\le b^{-2}h^{-2d}\int\int\int K(u)K((y-z)/h+u)^2 f_0(y)f_0(z)\ \mathrm{d}y\mathrm{d}z\mathrm{d}u\\
&= h^{-d}b^{-2}\int\int\int K(u)K(v+u)^2 f_0(z+hv)f_0(z)\ \mathrm{d}v\mathrm{d}z\mathrm{d}u\\
&\le h^{-d}b^{-2}\|f_0\|_{\mathbb{R}^d}\int\int K(u)K(v+u)^2\ \mathrm{d}v\mathrm{d}u\\
&= h^{-d}b^{-2}\|f_0\|_{\mathbb{R}^d}v_K.
\end{aligned}$$

Moreover, as

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{K_h(3,1)}{f_{0,3}}\right)\left(\frac{f_{h,3}}{f_{0,3}}\right)1_{\{X_3\in S\}}|X_1,X_2\right]\\
&= \int\left(\frac{K_h(x-X_1)f_h(x)}{f_0(x)}\right)1_{\{x\in S\}}\ \mathrm{d}x\\
&\le \|f_0\|_{\mathbb{R}^d}b^{-1},
\end{aligned}$$

and

$$\mathbb{E}\left[\left(\frac{f_{h,3}}{f_{0,3}}\right)^2 1_{\{X_3\in S\}}|X_1,X_2\right] = \int\frac{f_h(x)^2}{f_0(x)}1_{\{x\in S\}}\ \mathrm{d}x \le \|f_0\|_{\mathbb{R}^d}b^{-1}.$$

we finally obtain the result. ■

Lemma 6. *Under (H1) and (H2), if $S\subset\mathbb{R}^d$ is such that for all $x\in S$, $f_0(x)>b>0$, we have that*

$$\lim_{h\to 0} h^d\mathbb{E}\left[\left(\frac{K_h(1,2)-f_{h,1}}{f_{0,1}}\right)^2 1_{\{X_1\in S\}}\right] = v_K\lambda(S).$$

Proof. Write

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{K_h(1,2)-f_{h,1}}{f_{0,1}}\right)^2 1_{\{X_1\in S\}}\right]\\
&= \mathbb{E}\left[\frac{V_h(X_1)}{f_{0,1}^2}1_{\{X_1\in S\}}\right]\\
&= \mathbb{E}\left[\frac{\mathbb{E}[K_h(1,2)^2|X_1]}{f_{0,1}^2}1_{\{X_1\in S\}}\right] - \mathbb{E}\left[\frac{f_{h,1}^2}{f_{0,1}^2}1_{\{X_1\in S\}}\right].
\end{aligned}$$

The right-hand side is bounded by $b^{-2}\|f_0\|_{\mathbb{R}^d}$, hence its participation in the stated limit is 0. For the left-hand side term, use $\tilde{K} = K^2/v_K$ and write

$$\begin{aligned}
&h^d \mathbb{E}\left[\frac{\mathbb{E}[K_h(1,2)^2|X_1]}{f_{0,1}^2} 1_{\{X_1\in S\}}\right] \\
&\quad = h^d \int\int \frac{f_0(y)}{f_0(x)} K_h(x-y)^2 1_{\{x\in S\}}\, \mathrm{d}y\mathrm{d}x \\
&\quad = \int\int \frac{f_0(x-hu)}{f_0(x)} K(u)^2 1_{\{x\in S\}}\, \mathrm{d}u\mathrm{d}x \\
&\quad = v_K \int\int \frac{f_0(x-hu)}{f_0(x)} \tilde{K}(u) 1_{\{x\in S\}}\, \mathrm{d}u\mathrm{d}x \\
&\quad = v_K \lambda(S) + v_K \int\int \frac{(f_0(x-hu)-f_0(x))}{f_0(x)} \tilde{K}(u) 1_{\{x\in S\}}\, \mathrm{d}u\mathrm{d}x.
\end{aligned}$$

It remains to note that the term in the right goes to 0, as $h \to 0$, in virtue of the Lebesgue dominated convergence theorem. ■

Lemma 7. *Under (H1) and (H2), we have, for every $x \in \mathbb{R}^d$ and $h > 0$,*

$$|f_0 \star K_h(x) - f_0(x)| \le g(x)h^2 \int \|u\|_2^2 K(u)\, \mathrm{d}u.$$

Proof. Note that $\int K(u)\, \mathrm{d}u = 1$ and, by symmetry, $\int uK(u)\, \mathrm{d}u = 0$. Write

$$\begin{aligned}
|f_0 \star K_h(x) - f_0(x)| &= \left|\int (f_0(x-hu) - f_0(x))K(u)\, \mathrm{d}u\right| \\
&= \left|\int (f_0(x-hu) - f_0(x) - (hu)^T \nabla f_0(x))K(u)\, \mathrm{d}u\right| \\
&\le \int |f_0(x-hu) - f_0(x) - (hu)^T \nabla f_0(x)|\, K(u)\, \mathrm{d}u,
\end{aligned}$$

and use (H1) to conclude. ■

Lemma 8. *Under (H1), (H2), and (H3), there exists $c_2 > 0$ such that $\int_{S_{b_n}^c} f_{h_n}(x)\, \mathrm{d}x \le c_2 b_n^\beta$.*

Proof. Note that

$$\begin{aligned}
\int_{S_{b_n}^c} f_h(x)\, \mathrm{d}x &= \mathbb{P}(S_{b_n}^c) + \int_{S_{b_n}^c} (K_{h_n} \star f_0 - f_0)\, \mathrm{d}x \\
&= \mathbb{P}(S_{b_n}^c) + \int (K_{h_n} \star 1_{S_{b_n}^c}(x) - 1_{S_{b_n}^c}(x)) f_0(x)\, \mathrm{d}x.
\end{aligned}$$

The term in the left is bounded by cb_n^β as supposed in (H3). For the term in the right, define $S_{b_n,h_n} = \{y + h_n u \ :\ u \in [-1,1]^d,\ y \in S_{b_n}\}$. Note that, by (H2), as soon as $x \notin S_{b_n,h_n}$, $K_{h_n} \star 1_{S_{b_n}^c}(x) = 1$, hence

$$|K_{h_n} \star 1_{S_{b_n}^c}(x) - 1_{S_{b_n}^c}(x)| \le 1_{S_{b_n,h_n}}.$$

Moreover, for any $x \in S_{b_n,h_n}$, we have, by (H3), that

$$f_0(x) \le \sup_{y\in S_{b_n}} \sup_{u\in[-1,1]^d} f_0(y + h_n u) \le b_n \sup_{y\in S_{b_n}} \sup_{u\in[-1,1]^d} \frac{f_0(y+h_n u)}{f_0(y)} = b_n C,$$

hence, $1_{S_{b_n,h_n}} \leq 1_{f_0(x)\leq Cb_n}$, leading to

$$\left|\int (K_{h_n} \star 1_{S^c_{b_n}}(x) - 1_{S^c_{b_n}}(x)) f_0(x)\,\mathrm{d}x\right| \leq \int 1_{f_0(x)\leq Cb_n} f_0(x)\,\mathrm{d}x \leq (Cb_n)^{\beta}.$$

■

APPENDIX C. PARAMETRIC MAXIMUM LIKELIHOOD ESTIMATOR

In this section are reported some classical results on the maximum likelihood estimator of the density. When the model is well specified, we need the consistency and the asymptotic normality of the estimated parameter θ_0.

(A1) The set $\Theta \subset \mathbb{R}^q$ is compact. The model $\mathcal{P} = \{f_\theta \; : \; \theta \in \Theta\}$, a collection of densities on $\mathbb{R}^d$, is identifiable, that is, for every $\theta_1 \neq \theta_2$ in $\Theta, f_{\theta_1} \neq f_{\theta_2}$ and the envelop $F_\Theta(x) = \sup_{\theta\in\Theta} f_\theta(x)$ is such that $\mathbb{E}[\log(F_{\Theta,1})] < +\infty$. There exists an $\mathbb{R}^+$-valued measurable function $\dot{\ell}$ with $E\dot{\ell}(X_1)^2 < \infty$ for every $x \in \mathbb{R}^d$, for every θ_1 and θ_2 in Θ,

$$\lvert \log(f_{\theta_1}(x)) - \log(f_{\theta_2}(x))\rvert \leq \dot{\ell}(x)\|\theta_1 - \theta_2\|_2.$$

There exists $\delta > 0$ such that the function $\dot{\ell} \times \sup_{\theta\in B(\theta_0,\delta)} f_\theta$ is bounded.

It follows from (A1) that the class of functions $\mathcal{P}$ is Glivenko–Cantelli (van der Vaart & Wellner, 1996, theorem 2.7.11), that is,

$$\sup_{\theta\in\Theta}\left| n^{-1}\sum_{i=1}^{n}(\log(f_\theta(X_i) - E[\log(f_\theta(X_1))])\right| \to 0. \tag{C1}$$

Whenever $f_0 \in \mathcal{P}$, it holds that $\hat{\theta}_n \to \theta_0$, in probability (Newey & McFadden, 1994, theorem 2.1) or (van der Vaart, 1998, lemma 5.35). For now, asking the above Lipschitz condition to guarantee the Glivenko–Cantelli might seem a bit restrictive (Newey & McFadden, 1994, lemma 2.4), but this condition will also be required to derive asymptotic normality of $\hat{\theta}_n$ as well as to obtain uniform convergence (over $x \in \mathbb{R}^d$) of $f_{\hat{\theta}_n}(x)$ to $f_{\theta_0}(x)$. Indeed, we have that for any $\delta > 0$, with probability going to 1, $\hat{\theta}_n \in B(\theta_0, \delta)$. Hence, using the mean-value theorem, we find

$$\lvert f_{\hat{\theta}_n}(x) - f_{\theta_0}(x)\rvert \leq \|\hat{\theta}_n - \theta_0\|_2 \dot{\ell}(x) \sup_{\theta\in B(\theta_0,\delta)} f_\theta(x) \tag{C2}$$

for every $x \in \mathbb{R}^d$. Conclude using that $\dot{\ell} \times \sup_{\theta\in B(\theta_0,\delta)} f_\theta$ is bounded and the convergence in probability of $\hat{\theta}_n$ to θ_0.

(A2) The true parameter θ_0 an interior point of $\Theta \subset \mathbb{R}^q$. The model $\mathcal{P}$ is differentiable in quadratic mean at θ_0, that is, there exists a measurable vector-valued function $\dot{\ell}_{\theta_0}$, with $E[\|\dot{\ell}_{\theta_0}(X_1)\|_2^2]$, such that

$$\int \left[\sqrt{f_\theta} - \sqrt{f_{\theta_0}} - \frac{1}{2}(\theta - \theta_0)^T \dot{\ell}_{\theta_0}\sqrt{f_{\theta_0}}\right]^2 \mathrm{d}\lambda = o(\|\theta - \theta_0\|_2^2).$$

The matrix $\mathcal{I} = E[\dot{\ell}_{\theta_0}(X_1)\dot{\ell}_{\theta_0}(X_1)^T]$ is invertible.

As a consequence of the previous set of conditions (van der Vaart, 1998, lemma 5.39), we have

$$n^{1/2}(\hat{\theta}_n - \theta_0) = \mathcal{I}^{-1} n^{-1/2} \sum_{i=1}^{n} \dot{\ell}_{\theta_0}(X_i) + o_{\mathbb{P}}(1). \tag{C3}$$

where $E[\dot{\ell}_{\theta_0}(X_1)] = 0$. In particular, it holds that $\sqrt{n}(\hat{\theta}_n - \theta_0) = O_{\mathbb{P}}(1)$.