

DIFFERENTIABLE FUNCTIONALS AND SMOOTHED BOOTSTRAP

ANTONIO CUEVAS¹ AND JUAN ROMO²

¹*Departamento de Matemáticas, Universidad Autónoma de Madrid, 28049-Madrid, Spain*

²*Departamento de Estadística y Econometría, Universidad Carlos III de Madrid,
28903-Getafe (Madrid), Spain*

(Received May 15, 1995; revised May 7, 1996)

Abstract. The differentiability properties of statistical functionals have several interesting applications. We are concerned with two of them. First, we prove a result on asymptotic validity for the so-called *smoothed bootstrap* (where the artificial samples are drawn from a density estimator instead of being re-sampled from the original data). Our result can be considered as a smoothed analog of that obtained by Parr (1985, *Stat. Probab. Lett.*, **3**, 97–100) for the standard, unsmoothed bootstrap. Second, we establish a result on asymptotic normality for estimators of type $T_n = T(\hat{f}_n)$ generated by a density functional $T = T(f)$, \hat{f}_n being a density estimator. As an application, a quick and easy proof of the asymptotic normality of $\int \hat{f}_n^2$, (the *plug-in* estimator of the integrated squared density $\int f^2$) is given.

Key words and phrases: Smoothed bootstrap, differentiable statistical functionals, bootstrap validity, smoothed empirical process, integrated squared densities.

1. Introduction

The idea of using the differentiability properties of functionals in statistics goes back to von Mises (1947). It has become an elegant and unified methodology which has proved to be particularly fruitful in robust statistics (see, e.g., Huber (1981), Hampel *et al.* (1986)). A major application arises in the study of asymptotic normality: assume that the estimators $T_n = T_n(X_1, \dots, X_n)$ are generated by a distribution functional T in the sense that $T_n = T(F_n)$, where F_n is the empirical distribution and T is a functional defined on a large enough space of distribution functions (which includes the empiricals). Typically the parameter of interest is $T(F)$, F being the underlying distribution. Then, if T is differentiable, one can establish the asymptotic normality

$$(1.1) \quad \sqrt{n}(T(F_n) - T(F)) \rightarrow N(0, \sigma^2), \quad \text{weakly,}$$

by performing a Taylor expansion of type

$$(1.2) \quad \sqrt{n}(T(F_n) - T(F)) = \text{linear term} + \text{remainder term},$$

where the *linear* term is asymptotically normal (from the central limit theorem) and the remainder term is shown to converge to zero in probability.

Different concepts of differentiability have been used in this setting, although all of them rely on the same basic idea indicated in (1.2). Some references are: Boos and Serfling (1980), Fernholz (1983), Parr (1985), Clarke (1986), Shao (1989).

In Section 3 we will prove (using Shao's (1989) notion of differentiability) a result (Theorem 3.1) of type (1.1) for *plug-in* estimators $T_n = T(\hat{f}_n)$ of a target parameter $T(f)$ defined as the value of a density functional T at the underlying density f ; here \hat{f}_n stands for a nonparametric density estimator of f (see, e.g., Devroye and Györfi (1985) or Silverman (1986) for background). Some examples of interest could be the integral of the squared probability density, $T(f) = \int f^2$, the entropy $E(f) = \int f \log f$ and the curvature $C(f) = \int (f'')^2$. A quick and easy proof of the asymptotic normality of $\int f_n^2$ is obtained from our result (see Section 4 below).

Our second application of differentiability techniques has to do with bootstrap theory; it is developed in Section 2. For the sake of clarity and completeness we summarize below some relevant ideas on this subject.

1.1 Differentiability and bootstrap: some background

In spite of the great success of Efron's bootstrap (see, e.g., Efron and Tibshirani (1993) and Hall (1992) for general overviews), the basic ideas behind this methodology are not always well understood. Although the popular clichés tend to insist on the resampling aspects, the essential feature of the method is maybe better described as a *plug-in* idea: replace *everywhere* the underlying distribution F by an estimate \hat{F} .

More precisely, if X_1, \dots, X_n denotes a random sample from the distribution F , the sampling distribution $\mathcal{L}(R)$ under F , of a statistic $R \equiv R(X_1, \dots, X_n; F)$ can be approximated by the *bootstrap law* \mathcal{L}^* , which is the law under \hat{F} , of

$$(1.3) \quad R^* \equiv R(X_1^*, \dots, X_n^*; \hat{F}),$$

where X_1^*, \dots, X_n^* denotes an artificial (bootstrap) sample drawn from \hat{F} .

Different types of bootstrap are obtained depending on the estimator \hat{F} used in (1.3). The most usual choice (*standard bootstrap*) is $\hat{F} = F_n$ (the empirical distribution). When F is absolutely continuous with density f , there is another natural alternative, called *smoothed bootstrap*, in which \hat{F} is a *smoothed empirical*, \hat{F}_h , associated with a density estimator of f , usually of kernel type,

$$\hat{f}_n(t; h) = \sum_{i=1}^n \frac{1}{n} K_h(t - X_i),$$

where $h = h_n$ is the sequence of *smoothing parameters* or *bandwidths*, K is a density function (called *kernel*) and $K_h(x) = h^{-1}K(x/h)$.

The notation \hat{f}_h will be generally used instead of $\hat{f}_n(\cdot; h)$. The bootstrap samples drawn from the smoothed empirical \hat{F}_h are denoted by X_1^0, \dots, X_n^0 . We reserve the more usual notation X_1^*, \dots, X_n^* for standard resamples (obtained from the empirical F_n). Observe that, in fact, the bootstrap samples, smoothed or not, are a triangular array since the underlying distribution depends on n . This fact will be made explicit in the notation, with a double subindex, when convenient.

Usually, the bootstrap law cannot be explicitly evaluated as a function of X_1, \dots, X_n , so it has to be in turn approximated by resampling from \hat{F} . Sometimes, however, the exact bootstrap moments can be obtained: see, e.g., Marron (1992), De Angelis and Young (1992). In any case, the use of bootstrap methods requires the theoretical support of validity results showing that, at least asymptotically, the method works. Results of this type have been established (always for the case of standard bootstrap), among others, by Bickel and Freedman (1981), Singh (1981) and, in the more general context of empirical processes, by Giné and Zinn (1990). Their general form is

$$(1.4) \quad d(\mathcal{L}(R), \mathcal{L}^*(R^*)) \rightarrow 0 \text{ (a.s. or in probability),}$$

where d denotes a metric between probability measures whose associated convergence coincides with (or is stronger than) the weak convergence. From now on, our convergence statements of type “weakly, a.s. (or in prob.)” should be understood in the sense of (1.4). Unless otherwise stated, the arrow \rightarrow will denote convergence as n tends to infinity.

The idea of the smoothed bootstrap is not new: in fact, it is already mentioned in Efron’s (1979) pioneering paper. However, the literature on this subject is not very large. It concerns mostly to applications and comparisons with the standard bootstrap. Some relevant references are Silverman (1981), Silverman and Young (1987), Hall *et al.* (1989), Wang (1989), Young (1990) and De Angelis and Young (1992). In particular, Hall *et al.* (1989) show that the smoothed bootstrap greatly improves the convergence rate of the standard bootstrap in the estimation of the median variance.

We are not aware of general validity results of type (1.4) for the smoothed bootstrap. This paper provides one of these results by using the differential methodology for statistical functionals as outlined in (1.2). In fact, such a methodology applies to the case (very common in parametric inference) where the statistic R to be bootstrapped is of the form

$$(1.5) \quad R(X_1, \dots, X_n; F) \equiv \sqrt{n}(T(F_n) - T(F)),$$

where T is a statistical functional defined on an appropriate subset of the distribution functions. Then, provided that the asymptotic normality (1.1) holds, a validity result as (1.4) would be equivalent to

$$(1.6) \quad \sqrt{n}(T(F_n^*) - T(F_n)) \rightarrow N(0, \sigma^2), \quad \text{weakly, a.s. (or in prob.).}$$

Some references on this approach, applied to the standard bootstrap, are Parr (1985), Shao (1989) and Arcones and Giné (1992). In Section 2 below we

obtain a result (Theorem 2.1) of type (1.6) for the smoothed bootstrap, that is, F_n^* is replaced by F_n^0 in (1.6). The proof is based on the use of the differential methodology (along similar lines to those of Parr (1985)) together with a result by van der Vaart (1994) about the distance between the empirical F_n and the smoothed empirical \hat{F}_h . Lemma 2.1 in Section 2 can be considered as an auxiliary result having some independent interest. It establishes the validity of smoothed bootstrap for statistics defined as an average of the form $\sum \Psi(X_i)/n$. The result for the ordinary sample mean ($\Psi(x) = x$) is trivial, but the general case requires some conditions on Ψ and \hat{f}_h . Some applications are briefly discussed in Section 4.

2. Smoothed bootstrap validity

Let us first establish some assumptions to be used below.

(K0) The kernel K is a density with $\int_{\mathbb{R}} uK(u)du = 0$ and $\int_{\mathbb{R}} |u|^3 K(u)du < \infty$.

(K1) K is a bounded and symmetric density with $\int_{\mathbb{R}} K^2(u)du < \infty$.

(H0) The bandwidth $h \equiv h_n$ satisfies $h = o(n^{-1/4})$.

(Ψ 0) Ψ is a bounded function with three derivatives $\Psi^{(j)}$, $1 \leq j \leq 3$, with $\Psi^{(3)}$ bounded. Moreover,

$$(2.1) \quad E_f[\Psi(X)] := \int_{\mathbb{R}} \Psi(x)f(x)dx = 0,$$

$$(2.2) \quad \sigma^2 = E_f[\Psi^2(X)] = V_f[\Psi(X)] = \int_{\mathbb{R}} \Psi^2(x)f(x)dx < \infty,$$

$$(2.3) \quad E_f[|\Psi^{(2)}(X)|] < \infty.$$

The following lemma will be used to handle the linear term in an expansion of type (1.2) which allows to establish (see Theorem 2.1 below) the validity of smoothed bootstrap. Moreover, this lemma has some independent interest since it could be considered as a (smoothed) bootstrap central limit theorem.

LEMMA 2.1. *Let $\hat{f}_n := \hat{f}_n(t; h)$ be a kernel density estimator based on i.i.d. observations $\{X_i\}$ from a density f . Suppose (K0), (H0) and (Ψ 0) hold. Then*

$$(2.4) \quad \sqrt{n} \left(\frac{\sum_{i=1}^n \Psi(X_i^0)}{n} - \frac{\sum_{i=1}^n \Psi(X_i)}{n} \right) \rightarrow N(0, \sigma^2), \quad \text{weakly, a.s.,}$$

and

$$(2.5) \quad \sqrt{n} \left(\frac{\sum_{i=1}^n \Psi(X_i^0)}{n} - \int_{\mathbb{R}} \Psi(x)\hat{f}_h(x)dx \right) \rightarrow N(0, \sigma^2), \quad \text{weakly, a.s.,}$$

where X_1^0, \dots, X_n^0 denotes a bootstrap sample drawn from \hat{f}_h .

PROOF. Denote by E^0 , D^0 the smoothed bootstrap mean and standard deviation, respectively (thus, for instance, $E^0[\Psi(X_i^0)] = \int \Psi(x)\hat{f}_h(x)dx$), and

$$Z_{in} = \Psi(X_i^0) - E^0\Psi(X_i^0) = \Psi(X_i^0) - \Psi(X) - E^0(\Psi(X_i^0) - \Psi(X)),$$

(recall that the observations X_i^0 are, in fact, a triangular array since $X_i^0 \sim \hat{f}_h$).

We use Lyapounov's theorem for triangular arrays (see, e.g., Billingsley (1986, p. 371)): denote $S_n := \sum_{i=1}^n Z_{in}$; if

$$(2.6) \quad \lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n E^0[|Z_{in}|^{2+\delta}]}{D^0(S_n)^{2+\delta}} = 0, \quad \text{a.s. for some } \delta > 0,$$

then

$$(2.7) \quad \frac{S_n}{D^0(S_n)} \rightarrow N(0, 1), \quad \text{weakly, a.s.}$$

We will check condition (2.6) for an arbitrary $\delta > 0$: since the Z_{in} are identically distributed for $1 \leq i \leq n$, we have

$$(2.8) \quad \sum_{i=1}^n E^0[|Z_{in}|^{2+\delta}] = nE^0[|Z_{1n}|^{2+\delta}].$$

Since Ψ is bounded, the sequence $E^0[|Z_{1n}|^{2+\delta}]$ is also bounded and (2.8) entails

$$(2.9) \quad \sum_{i=1}^n E^0[|Z_{in}|^{2+\delta}] = O(n) \quad \text{a.s.}$$

On the other hand, the denominator in (2.6) is

$$(2.10) \quad D^0(S_n)^{2+\delta} = V^0(S_n)^{(2+\delta)/2} = (nV^0(Z_{1n}))^{(2+\delta)/2} = (nE^0(Z_{1n}^2))^{(2+\delta)/2}.$$

Since, from the law of large numbers, $(1/n) \sum_{i=1}^n \Psi(X_i) \rightarrow E_f[\Psi(X)] = 0$ a.s., in addition to the fact that Ψ (and hence Ψ^2) are continuous and bounded and the distribution given by \hat{f}_h converges weakly a.s. to the one given by f , we have

$$\begin{aligned} \lim_{n \rightarrow \infty} E^0(Z_{1n}^2) &= \lim_{n \rightarrow \infty} \int_{\mathbb{R}} (\Psi(x) - \overline{\Psi(X)})^2 \hat{f}_h(x) dx \\ &\quad - \lim_{n \rightarrow \infty} \left[\int_{\mathbb{R}} (\Psi(x) - \overline{\Psi(X)}) \hat{f}_h(x) dx \right]^2 \\ &= \int_{\mathbb{R}} \Psi^2(x) f(x) dx - \left(\int_{\mathbb{R}} \Psi(x) f(x) dx \right)^2 = \sigma^2 \quad \text{a.s.} \end{aligned}$$

Thus, (2.10) implies

$$D^0(S_n)^{2+\delta} = O(n^{(2+\delta)/2}), \quad \text{a.s.}$$

So, (2.6) holds and (2.7) follows. Moreover, since

$$\lim_{n \rightarrow \infty} \frac{D^0(S_n)}{\sqrt{n}} = \lim_{n \rightarrow \infty} E^0(Z_{1n}^2)^{1/2} = \sigma \quad \text{a.s.},$$

we have in fact proved that

$$(2.11) \quad \sqrt{n} \left(\frac{\sum_{i=1}^n \Psi(X_i^0)}{n} - \frac{\sum_{i=1}^n \Psi(X_i)}{n} \right) - \sqrt{n} E^0(\Psi(X_1^0) - \overline{\Psi(X)}) \rightarrow N(0, \sigma^2),$$

weakly, a.s.

Let us study the bias term; we perform a standard change of variable ($x = hu + X_i$) in each integral, together with a second order Taylor expansion (denote by ζ_i a point between X_i and x):

$$(2.12) \quad \begin{aligned} \lim_{n \rightarrow \infty} \sqrt{n} E^0(\Psi(X_1^0) - \overline{\Psi(X)}) &= \lim_{n \rightarrow \infty} \sqrt{n} \left(\int_{\mathbb{R}} \Psi(x) \hat{f}_h(x) dx - \overline{\Psi(X)} \right) \\ &= \lim_{n \rightarrow \infty} \sqrt{n} \left(\sum_{i=1}^n \int_{\mathbb{R}} \Psi(x) \frac{1}{n} K_h(x - X_i) dx - \overline{\Psi(X)} \right) \\ &= \lim_{n \rightarrow \infty} \sqrt{n} \left(\sum_{i=1}^n \frac{1}{n} \Psi(X_i) + \left(\int_{\mathbb{R}} u^2 K(u) du \right) \frac{h^2}{2n} \sum_{i=1}^n \Psi^{(2)}(X_i) \right. \\ &\quad \left. + \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}} \Psi^{(3)}(\zeta_i) \frac{h^3 u^3}{6} K(u) du - \overline{\Psi(X)} \right) \\ &= \lim_{n \rightarrow \infty} \sqrt{n} h^2 \left(\int_{\mathbb{R}} u^2 K(u) du \right) \frac{1}{2n} \sum_{i=1}^n \Psi^{(2)}(X_i) \\ &\quad + \lim_{n \rightarrow \infty} \sqrt{n} o(h^2) = 0 \quad \text{a.s.} \end{aligned}$$

We have used the boundedness of $\Psi^{(3)}$ and the fact that $\int |u|^3 K(u) du < \infty$ to ensure that the remainder terms, for every X_i , in the Taylor expansions are uniformly bounded by an expression of type $o(h^2)$. So, their average is also $o(h^2)$. Then, (2.12) follows by using $nh^4 \rightarrow 0$ (assumption (H0)) and

$$\frac{1}{n} \sum_{i=1}^n \Psi^{(2)}(X_i) \rightarrow E_f[\Psi^{(2)}(X)] \quad \text{a.s.}$$

Finally, (2.12) and (2.11) lead (by Slutsky's theorem) to the desired conclusion (2.4).

Remark 1. This lemma is conceptually related with the approach of Silverman and Young (1987) and De Angelis and Young (1992). These authors consider the bootstrap estimation (in the plug-in sense we commented above) of linear functionals of type $A(F) = \int \Psi(x) dF(x)$. In other words, they study the properties of $A(\hat{F}_h)$ as an estimator of $A(F)$, whereas our Lemma 2.1 provides conditions under which the approximations

$$(2.13) \quad \mathcal{L}(\sqrt{n}(A(F_n) - A(F))) \simeq \mathcal{L}^0(\sqrt{n}(A(F_n^0) - A(F_n)))$$

and

$$(2.14) \quad \mathcal{L}(\sqrt{n}(A(F_n) - A(F))) \simeq \mathcal{L}^0(\sqrt{n}(A(F_n^0) - A(\hat{F}_h)))$$

hold. Thus, our conclusion (2.14) can be seen as a further result in the approach of Silverman and Young (1987) and De Angelis and Young (1992). Note also that in (2.13) the smoothed empirical \hat{F}_h is used only for generating the artificial sample whose corresponding empirical, F_n^0 , is plugged-in into $A(\cdot)$.

Remark 2. The validity of the smoothed bootstrap for the sample mean directly follows from the convolution structure of the smoothed observations: indeed, $X_i^0 = X_i^* + \varepsilon_i$, where the $\varepsilon_i \sim K_h$ is independent of X_i^* , and $K_h = h^{-1}K(\cdot/h)$, K being a density with mean zero and finite variance. Thus,

$$\sqrt{n}(\bar{X}^0 - \bar{X}) = \sqrt{n}(\bar{X}^* - \bar{X}) + \sqrt{n}\bar{\varepsilon}.$$

The first term converges weakly a.s. (Bickel and Freedman (1981), Singh (1981)) to $N(0, \sigma^2)$ and $\sqrt{n}\bar{\varepsilon} \rightarrow 0$ in probability, whenever $h \rightarrow 0$ (as $n \rightarrow \infty$): this follows from Chebichev's inequality since $h \rightarrow 0$ if and only if the variance of K_h converges to zero (recall that K_h is the density of hU where U is a random variable with density K). Note that in this particular case the condition $nh^4 \rightarrow 0$ is not needed.

Remark 3. Hypothesis (H0) in Lemma 2.1 imposes a bandwidth order of type $o(n^{-1/4})$, which is slightly smaller than the typical optimum ($n^{-1/5}$) in density estimation (see, e.g., Silverman (1986), p. 40). This comparison, however, is not completely adequate since this optimum corresponds to a different problem: the global (L_2) estimation of the underlying density. This is in the the same spirit of De Angelis and Young's ((1992), p. 48) remark: "...even if smoothing is considered worthwhile in the bootstrap estimation, the optimal amount of smoothing will generally be small compared to that appropriate for estimating the underlying density". In fact, the optimal bandwidths obtained by these authors for the estimation of a wide class of linear functionals are of type $O(n^{-1/2})$; they also found, however, that the classical order $O(n^{-1/5})$ turns out to be the optimum for the bootstrap estimation of quantiles. The optimal smoothing orders obtained by Silverman and Young (1987) for some particular functionals of linear type, also range from $n^{-1/2}$ to $n^{-1/5}$.

Let us finally mention that Hall *et al.* (1989) use a bandwidth of type $O(n^{-1/5})$ for the bootstrap estimation of the variance of the sample quantile.

Remark 4. As stated in Lemma 2.1, $h_n = o(n^{-1/4})$ is just a sufficient condition for the validity of the smoothed bootstrap, but one could intuitively expect that some smallness requirement would also be necessary. The reason is that if the bandwidth is very large, the kernel estimator is too far away from the sample as to provide relevant information by resampling. The sample mean case considered in Remark 2 would be an exception because, when using a zero mean kernel, the estimator mean ($\int x \hat{f}_h(x) dx$) coincides with the sample mean, so that no bias is introduced by smoothing.

Remark 5. Conclusions (2.4) and (2.5) can be seen as two different statements of validity for smoothed bootstrap, where the centering values are the sample mean and the smoothed sample mean, respectively. Only (2.4) will be used below.

We next present the notion of *Fréchet differentiability* required for Theorem 2.1.

DEFINITION 2.1. Denote by $\|\cdot\|$ the supremum norm defined on a space \mathcal{F} of distribution functions (which is assumed to include the empirical distributions). A functional T defined on \mathcal{F} is said to be differentiable at $F \in \mathcal{F}$ if there exists a function $\Psi : \mathbb{R} \rightarrow \mathbb{R}$ (depending on F) such that

$$T(G) - T(F) + \int_{\mathbb{R}} \Psi(x)d(G(x) - F(x)) + o(\|G - F\|), \text{ as } \|G - F\| \rightarrow 0.$$

Observe that the function Ψ (usually called the *influence function* of T at F) is defined up to an additive constant. Thus, we may (and will) assume throughout that $\int \Psi dF = 0$. The following result provides the asymptotic validity of the smoothed bootstrap for distribution functionals.

THEOREM 2.1. Assume that the distribution functional T is differentiable (according to Definition 2.1) at an absolutely continuous distribution F whose density f is continuously differentiable with bounded derivative. Assume further that conclusion (2.4) in Lemma 2.1 is fulfilled, Ψ being the influence function of T at F and $\sigma^2 = \int \Psi^2 dF$. Let X_1^0, \dots, X_n^0 be a sample drawn from a kernel density estimator $\hat{f}_h := \hat{f}_n(t; h)$ (based on a sample X_1, \dots, X_n of f) such that the bandwidth h and the kernel K satisfy (H0) and (K0), respectively.

Then, the “bootstrap law” (under \hat{f}_h) of $\sqrt{n}(T(F_n^0) - T(F_n))$ converges (in the supremum norm) to $N(0, \sigma^2)$, in probability, where F_n^0 and F_n denote the empirical distributions corresponding to the samples X_1^0, \dots, X_n^0 and X_1, \dots, X_n .

PROOF. By using the differentiability condition of Definition 2.1 we obtain

$$(2.15) \quad T(F_n) = T(F) + \int_{\mathbb{R}} \Psi(x)dF_n(x) + o(\|F_n - F\|)$$

and

$$(2.16) \quad T(F_n^0) = T(F) + \int_{\mathbb{R}} \Psi(x)dF_n^0(x) + o(\|F_n^0 - F\|).$$

So,

$$(2.17) \quad \begin{aligned} \sqrt{n}(T(F_n^0) - T(F_n)) &= \sqrt{n} \int_{\mathbb{R}} \Psi(x)d(F_n^0(x) - F_n(x)) \\ &\quad + \sqrt{n}o(\|F_n^0 - F\|) + \sqrt{n}o(\|F_n - F\|). \end{aligned}$$

Denote by U_n and V_n the remainder terms in (2.15) and (2.16), respectively. By Donsker’s theorem, we have

$$(2.18) \quad \sqrt{n}\|F_n - F\| = O_P(1).$$

Thus, $\sqrt{n}U_n \rightarrow 0$, in probability.

On the other hand, if we denote by \hat{F}_h the distribution function corresponding to \hat{f}_h ,

$$\sqrt{n}V_n = o(A_n + B_n + C_n),$$

where $A_n = \sqrt{n}\|F_n^0 - \hat{F}_h\|$, $B_n = \sqrt{n}\|\hat{F}_h - F_n\|$ and $C_n = \sqrt{n}\|F_n - F\|$. But

$$(2.19) \quad A_n \text{ is stochastically bounded.}$$

This follows from Dvoretzky *et al.* (1956) inequality which implies $\sqrt{n}\|F_n^0 - \hat{F}_h\| = O_P(1)$, where $O_P(1)$ refers here to the conditional probabilities generated by the bootstrap randomization (observe that F_n^0 is the empirical corresponding to \hat{F}_h). Also,

$$(2.20) \quad B_n \rightarrow 0, \quad \text{in probability.}$$

This follows as a direct consequence of a result recently proved by van der Vaart (1994). This author establishes (under conditions (2.22) and (2.23) indicated below) that

$$(2.21) \quad \sqrt{n}\|\hat{F}_h - F_n\| = o_P(1).$$

In fact, van der Vaart's result is much more general since it holds (even for the case of random h) in the broader sense of empirical processes when the supremum is taken over a Donsker class: in our case it suffices to consider the Donsker class of half-lines. Thus, the associated supremum leads to the norm $\|\cdot\|$, and van der Vaart's conclusion is exactly (2.21); the required hypotheses in this case are (see conditions (1.1) and (1.2) in van der Vaart (1994), keeping in mind that we assume a deterministic h):

$$(2.22) \quad \sup_{a \in \mathbb{R}} E \left(\int_{\mathbb{R}} (I_{(-\infty, a]}(X + y) - I_{(-\infty, a]}(X)) K_h(y) dy \right)^2 \rightarrow 0,$$

and

$$(2.23) \quad \sup_{a \in \mathbb{R}} \sqrt{n} \left| E \int_{\mathbb{R}} (I_{(-\infty, a]}(X + y) - I_{(-\infty, a]}(X)) K_h(y) dy \right| \rightarrow 0.$$

Let us check that these conditions hold in our case (we denote by μ_n the probability measure associated with K_h):

$$\begin{aligned} & \sup_{a \in \mathbb{R}} E \left(\int_{\mathbb{R}} (I_{(-\infty, a]}(X + y) - I_{(-\infty, a]}(X)) K_h(y) dy \right)^2 \\ &= \sup_{a \in \mathbb{R}} \int_{\mathbb{R}} \left(\int_{\mathbb{R}} (I_{(-\infty, a]}(x + y) - I_{(-\infty, a]}(x)) d\mu_n(y) \right)^2 dF(x) \\ &= \sup_{a \in \mathbb{R}} \int_{\mathbb{R}} \left(\int_{\mathbb{R}} I_{(-\infty, a-x]}(y) d\mu_n(y) - I_{(-\infty, a]}(x) \right)^2 dF(x) \\ &= \sup_{a \in \mathbb{R}} \left(\int_{\{x > a\}} (\mu_n(-\infty, a - x])^2 dF(x) \right) \end{aligned}$$

$$\begin{aligned}
 & + \int_{\{x < a\}} (1 - \mu_n(-\infty, a - x])^2 dF(x) \Big) \\
 = & \sup_{a \in \mathbb{R}} \left(\int_{\{x > a\}} (\mu_n(-\infty, a - x])^2 dF(x) \right. \\
 & \left. + \int_{\{x < a\}} (\mu_n(a - x, +\infty))^2 dF(x) \right)
 \end{aligned}$$

Since F is absolutely continuous, for all $\epsilon > 0$, there exists $\delta > 0$ such that $|t - s| \leq \delta$ implies $P\{[s, t]\} < \epsilon$ (where P is the probability measure associated with F). Also, from the weak convergence of μ_n to 0, it follows that there is $n_0 \in \mathbb{N}$ such that for $n \geq n_0$, $\mu_n((-\infty, -\delta] \cup [\delta, +\infty)) < \epsilon$. Now the expression above can be written as

$$\begin{aligned}
 & \sup_{a \in \mathbb{R}} \left(\int_{\{a-x < -\delta\}} (\mu_n(-\infty, a - x])^2 dF(x) \right. \\
 & \quad + \int_{\{-\delta < a-x < 0\}} (\mu_n(-\infty, a - x])^2 dF(x) \\
 & \quad + \int_{\{a-x > \delta\}} (\mu_n(a - x, +\infty))^2 dF(x) \\
 & \quad \left. + \int_{\{0 < a-x < \delta\}} (\mu_n(a - x, +\infty))^2 dF(x) \right) \\
 & \leq 2\epsilon^2 + \sup_{a \in \mathbb{R}} \{P\{x : -\delta < a - x < 0\} + P\{x : 0 < a - x < \delta\}\} \\
 & = 2\epsilon^2 + \sup_{a \in \mathbb{R}} \{P(a, a + \delta) + P(a - \delta, a)\} \leq 2\epsilon^2 + 2\epsilon,
 \end{aligned}$$

for $n \geq n_0$.

With regard to condition (2.23), van der Vaart (1994) shows that it is fulfilled, under (K0) and the assumption imposed on f , when the bandwidth is $h = o(n^{-1/4})$. The reason is that the expression in (2.23) can be bounded by $C\sqrt{nh^2}E(U^2)$, where C is a constant and U is a random variable with density K . Thus the theorem in van der Vaart (1994) implies (2.20). Then, we conclude the convergence (in probability) to zero of the remainder term $\sqrt{n}V_n$ since C_n is also stochastically bounded by (2.18).

Recall, finally, that we have assumed the conclusion (2.4) of Lemma 2.1, that is, the linear term in (2.17) satisfies

$$\begin{aligned}
 (2.24) \quad \sqrt{n} \int_{\mathbb{R}} \Psi(x) d(F_n^0(x) - F_n(x)) &= \sqrt{n} \left(\frac{\sum_{i=1}^n \Psi(X_i^0)}{n} - \frac{\sum_{i=1}^n \Psi(X_i)}{n} \right) \\
 &\rightarrow N(0, \sigma^2), \quad \text{weakly, a.s.}
 \end{aligned}$$

Now the result directly follows from (2.17)–(2.24).

Remark 6. It is worth noting that Theorem 2.1 could be adapted to the case of random h . The method of proof would be essentially the same since van der Vaart's (1994) result $\sqrt{n}\|\hat{F}_h - F_n\| = o_P(1)$ also holds for the case of random h . This extension is very important from a practical point of view since the bandwidth h is empirically chosen (and, hence, it is random) in most cases. We have preferred, however, to present Theorem 2.1 in the current simpler form.

Remark 7. An alternative way of proving the convergence to zero of term B_n is by using Theorem 2.3 in Fernholz (1991). This result establishes $\sqrt{n}\|\hat{F}_h - F_n\| \rightarrow 0$ a.s. by assuming that F is uniformly Lipschitz (i.e., $|F(x) - F(y)| \leq C|x - y|$ for some constant C) and $\int_{|u| > b_n/h} K_h(u) du = o(n^{-1/2})$, for some positive sequence $\{b_n\}$ such that $b_n = o(n^{-1/2})$. However, this condition implies that $h_n = o(n^{-1/2})$ which is much more restrictive than assumption (H0).

3. Asymptotic normality of density functionals

We now adapt Shao's (1989) definition of locally Lipschitz differentiability to density functionals. This property will allow us to establish an asymptotic normality result for these functionals.

DEFINITION 3.1. *A real functional T , defined on a subspace \mathcal{D} of the space of square-integrable densities, is said to be locally Lipschitz differentiable at $f \in \mathcal{D}$ with respect to the L_2 -norm ($\|\cdot\|_2$) if*

$$(3.1) \quad T(g) = T(f) + \int_{\mathbb{R}} \Psi(x)(g(x) - f(x))dx + O(\|g - f\|_2^2),$$

as $\|g - f\|_2 \rightarrow 0$, for $g \in \mathcal{D}$.

In the sequel we will assume that every density functional T is defined on a large enough subspace \mathcal{D} (which includes all the required kernel density estimates as well as the underlying density f).

THEOREM 3.1. *Let T be a density functional locally Lipschitz differentiable at a square-integrable density f , such that the function Ψ of (3.1) fulfils condition ($\Psi 0$). Assume that*

(D1) *f has two continuous derivatives and $f^{(2)} \in L_2(\mathbb{R})$ (i.e., the second derivative of f is square integrable on \mathbb{R}).*

Let $\hat{f}_n := \hat{f}_n(t; h)$ be a sequence of kernel density estimators of f (as defined in (1)), where the kernel K satisfies (K0) and (K1) and the bandwidth fulfils

(H1) *$h_n^{-1} = o(n^{1/2})$ and $h_n = o(n^{-1/4})$.*

Then,

$$(3.2) \quad \sqrt{n}(T(\hat{f}_n) - T(f)) \rightarrow N(0, \sigma^2) \quad \text{weakly,}$$

where $\sigma^2 = E_f[\Psi^2(X)] = V_f[\Psi(X)]$.

PROOF. Since $E_f[\Psi(X)] = 0$, differentiability condition (3.1) implies that

$$(3.3) \quad \sqrt{n}(T(\hat{f}_h) - T(f)) = \sqrt{n} \int_{\mathbb{R}} \Psi(x) \hat{f}_h(x) dx + \sqrt{n} O(\|\hat{f}_h - f\|_2^2).$$

We first study the main term

$$\int_{\mathbb{R}} \Psi(x) \hat{f}_h(x) dx = \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}} \Psi(x) K_h(x - X_i) dx.$$

In a similar way to that of expression (2.12) in the proof of Lemma 2.1, we use the change of variable $x = hu + X_i$ in the i -th integral as well as a second order Taylor expansion to get

$$(3.4) \quad \begin{aligned} \int_{\mathbb{R}} \Psi(x) \hat{f}_h(x) dx &= \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}} \Psi(hu + X_i) K(u) du \\ &= \frac{1}{n} \sum_{i=1}^n \Psi(X_i) \\ &\quad + \frac{h^2}{2} \left(\int_{\mathbb{R}} u^2 K(u) du \right) \left(\frac{1}{n} \sum_{i=1}^n \Psi^{(2)}(X_i) \right) + o(h^2). \end{aligned}$$

Again, the first-order term vanishes because $\int uK(u)du = 0$; the boundedness of $\Psi^{(3)}$ entails that the remainder terms are uniformly bounded by a term $o(h^2)$ so that their average is also $o(h^2)$.

On the other hand, the standard central limit theorem and the strong law of large numbers yield, respectively,

$$(3.5) \quad \frac{1}{\sqrt{n}} \sum_{i=1}^n \Psi(X_i) \rightarrow N(0, \sigma^2) \quad \text{weakly,}$$

and

$$(3.6) \quad \sqrt{nh^2} \frac{\sum_{i=1}^n \Psi^{(2)}(X_i)}{n} \rightarrow 0, \quad \text{a.s.,}$$

since $\sqrt{nh^2} \rightarrow 0$. Now, (3.4)–(3.6) imply

$$(3.7) \quad \sqrt{n} \left(\int_{\mathbb{R}} \Psi(x) \hat{f}_h(x) dx \right) \rightarrow N(0, \sigma^2), \quad \text{weakly.}$$

As for the remainder term in (3.3), hypothesis (D1) and the assumptions made on K establish precise conditions under which the mean integrated square error

$$MISE(h) := E_f \|\hat{f}_h - f\|_2^2$$

has the usual structure

$$(3.8) \quad MISE(h) = O(h^4) + O(n^{-1}h^{-1}),$$

where $O(h^4)$ and $O(n^{-1}h^{-1})$ correspond, respectively, to the bias and variance term. This well-known result can be proved by a standard argument based on Taylor expansions (see, e.g., Silverman ((1986), pp. 39–40), Cao ((1990), pp. 32–36)). A much more detailed analysis of the structure of $MISE(h)$ is carried out, under supplementary assumptions, by Cao ((1993), Lemma 1).

By assumption (H1), the bias and variance terms in (3.8) are of type $o(n^{-1})$ and $o(n^{-1/2})$, respectively. Thus $MISE(h)$ is also $o(n^{-1/2})$ and, by Markov's inequality,

$$(3.9) \quad \sqrt{n}\|\hat{f}_h - f\|_2^2 = o_P(1).$$

Finally, from (3.3), (3.7) and (3.9) we get (3.2).

Remark 8. Condition (H1) is fulfilled, e.g., by bandwidths of type $h_n = Cn^{-p}$, with $\frac{1}{4} < p < \frac{1}{2}$, which also fulfil condition (H0) in Lemma 2.1 and Theorem 2.1.

4. Examples and final comments

In this section we present several applications of the previous results both for distribution and density functionals. Finally, we provide some remarks on the results contained in the paper.

4.1 Bootstrap validity for M - and L -estimators

Shao (1989) provides conditions ensuring differentiability (as required in Theorem 2.1) of the distribution functionals which generate the well-known M - and L -estimators.

Let $\rho(x, t)$ be a real function on \mathbb{R}^2 . The M -functional $T = T(F)$ is defined as a solution of

$$(4.1) \quad \int_{\mathbb{R}} \rho(x, T(F)) dF(x) = \min_t \int_{\mathbb{R}} \rho(x, t) dF(x).$$

$T(F_n)$ is the M -estimator of $T(F)$. Suppose that $T(F)$ is the unique minimum of (4.1) and for each t ,

$$\phi(x, t) = \partial \rho(x, t) / \partial t$$

exists a.e. and ϕ is increasing in t . Assume that

$$\partial \int_{\mathbb{R}} \rho(x, t) dG(x) / \partial t = \int_{\mathbb{R}} \phi(x, t) dG(x)$$

and that the function $\lambda_F(t) = \int \phi(x, t) dF(x)$ has positive derivative at $T(F)$. Assume further that either ρ is bounded or ϕ is bounded and continuous and $\lambda_F(t)$ has a unique root. Denote by $\|\cdot\|_V$ the total variation norm (see, e.g., Natanson (1961)). If there exists a neighborhood \mathcal{N} of $T(F)$ such that $\|\phi(\cdot, t)\|_V < \infty$ for $t \in \mathcal{N}$ and $\|\phi(\cdot, t) - \phi(\cdot, T(F))\|_V \rightarrow 0$ as $t \rightarrow T(F)$ then T is differentiable

according to Definition 2.1 as needed in Theorem 2.1; moreover, the influence function is given by $\psi(x) = -\phi(x, T(F))/\lambda'_F(T(F))$.

The corresponding result for L -estimators can be also seen in Shao (1989). Let T be an L -functional defined by $T(G) = \int xJ(G(x))dG(x)$, $G \in \mathcal{F}$, where J is a function defined on $[0, 1]$. See, e.g., Serfling (1980) and Lehmann (1983) for examples. If $J(t) = 0$ for $t < \alpha$ or $t > \beta$, $0 < \alpha < \beta < 1$ and J is bounded and continuous a.e. (Lebesgue) and a.e. (F^{-1}), then T is differentiable as required in Theorem 2.1 and the influence function is given by $\psi(x) = -\int (I_{\{x \leq y\}} - F(y))J(F(y))dy$.

So, the classes of M - and L -estimators provide examples where Theorem 2.1 could be applied. Thus, this result would support the validity of asymptotic confidence intervals or, in general, asymptotic inference based on smoothed bootstrap resampling, under the appropriate conditions.

4.2 Asymptotic normality of $\int \hat{f}_n^2$

An important example of density functional is the integral of the squared probability density, $T(f) = \int f^2$. It appears in a number of interesting statistical problems as, for instance, in rank-based nonparametric statistics. Some references are Aubuchon and Hettmansperger (1984), Birgé and Massart (1995), Hall and Marron (1987), Prakasa Rao (1983), Sheather *et al.* (1994), Wu (1995).

The locally Lipschitz differentiability of T at f follows easily from the identity $\int (f - g)^2 = \int f^2 + \int g^2 - \int 2fg$, because this gives that

$$\int_{\mathbb{R}} g^2 - \int_{\mathbb{R}} f^2 = - \int_{\mathbb{R}} 2f^2 + \int_{\mathbb{R}} 2fg + \|g - f\|_2^2 = \int_{\mathbb{R}} 2f(g - f) + O(\|g - f\|_2^2)$$

and so, the influence function is $\Psi(x) = 2f(x) - 2 \int f^2$ (recall that $\Psi(x)$ is defined up to additive constants and we take $E\Psi = 0$). Thus, our Theorem 3.1 applies to this functional, establishing its asymptotic normality.

4.3 Other norms

Note that Definition 3.1 of locally Lipschitz differentiability is given for the L_2 -norm but other choices as the L_1 - or the sup norms could be also considered; this would allow to establish the corresponding differentiability and the asymptotic normality for new density functionals. In particular, the L_1 -norm seems to be specially suitable for this purpose because universal bounds for the L_1 -error are available for this metric: see Datta (1992) and references therein. The entropy *plug-in* estimator is a possible candidate to be considered in this framework.

Acknowledgements

We are very grateful to Ricardo Fraiman and Ricardo Cao for helpful comments concerning technical details in the proofs. We also gratefully acknowledge the careful report of an anonymous referee. This work was partially supported by Spanish grants PB94-0179 (A. Cuevas) and PB93-0232 (J. Romo).

REFERENCES

- Arcones, M. A. and Giné, E. (1992). On the bootstrap of M-estimators and other statistical functionals, *Proc. IMS Bootstrap Conference, East Lansing, 1990*.
- Aubuchon, C. and Hettmansperger, T. P. (1984). A note on the estimation of the integral of $f^2(x)$, *J. Statist. Plann. Inference*, **9**, 321–331.
- Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap, *Ann. Statist.*, **9**, 1196–1217.
- Billingsley, P. (1986). *Probability and Measure*, second edition, Wiley, New York.
- Birgé, L. and Massart, P. (1995). Estimation of integral functionals of a density, *Ann. Statist.*, **23**, 11–29.
- Boos, D. D. and Serfling, R. J. (1980). A note on differentials and the CLT and LIL for statistical functions with application to M-estimates, *Ann. Statist.*, **8**, 618–624.
- Cao, R. (1990). Aplicaciones y nuevos resultados del método bootstrap en la estimación no paramétrica de curvas, Ph.D. Thesis, Universidad de Santiago de Compostela.
- Cao, R. (1993). Bootstrapping the mean integrated squared error, *J. Multivariate Anal.*, **45**, 137–160.
- Clarke, B. R. (1986). Nonsmooth analysis and Fréchet differentiability of M-functionals, *Probab. Theory Related Fields*, **73**, 197–209.
- Datta, S. (1992). Some non-asymptotic bounds for L_1 density estimation using kernels, *Ann. Statist.*, **20**, 1658–1667.
- De Angelis, D. and Young, G. A. (1992). Smoothing the bootstrap, *Internat. Statist. Rev.*, **60**, 45–56.
- Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: The L_1 -View*, Wiley, New York.
- Dvoretzky, A., Kiefer, J. and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator, *Ann. Math. Statist.*, **27**, 642–669.
- Efron, B. (1979). Bootstrap methods: another look at the jackknife, *Ann. Statist.*, **7**, 1–26.
- Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*, Chapman and Hall, New York.
- Fernholz, L. T. (1983). *von Mises Calculus for Statistical Functionals*, Springer, New York.
- Fernholz, L. T. (1991). Almost sure convergence of smoothed empirical distribution functions, *Scand. J. Statist.*, **18**, 255–262.
- Giné, E. and Zinn, J. (1990). Bootstrapping general empirical measures, *Ann. Probab.*, **18**, 851–869.
- Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*, Springer, New York.
- Hall, P. and Marron, J. S. (1987). Estimation of integrated squared density derivatives, *Stat. Probab. Lett.*, **6**, 109–115.
- Hall, P., DiCiccio, T. J. and Romano, J. P. (1989). On smoothing and the bootstrap, *Ann. Statist.*, **17**, 692–704.
- Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics. The Approach Based on Influence Functions*, Wiley, New York.
- Huber, P. J. (1981). *Robust Statistics*, Wiley, New York.
- Lehmann, E. L. (1983). *Theory of Point Estimation*, Wiley, New York.
- Marron, J. S. (1992). Bootstrap bandwidth selection. *Exploring the Limits of Bootstrap* (eds. R. Lepage and L. Billard), 249–262, Wiley, New York.
- Natanson, I. P. (1961). *Theory of Functions of a Real Variable*, **1**, Ungar, New York, rev. ed.
- Parr, W. C. (1985). The bootstrap: some large sample theory and connections with robustness, *Stat. Probab. Lett.*, **3**, 97–100.
- Priakasa Rao, B. L. S. (1983). *Nonparametric Functional Estimation*, Academic Press, New York.
- Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*, Wiley, New York.
- Shao, J. (1989). Functional calculus and asymptotic theory for statistical analysis, *Stat. Probab. Lett.*, **8**, 397–405.

- Sheather, S. J., Hettmansperger, T. P. and Donald, M. R. (1994). Data-based bandwidth selection for kernel estimators of the integral of f^2 , *Scand. J. Statist.*, **21**, 265–275.
- Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality, *J. Roy. Statist. Soc. Ser. B*, **43**, 97–99.
- Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York.
- Silverman, B. W. and Young, G. A. (1987). The bootstrap: to smooth or not to smooth? *Biometrika*, **74**, 469–479.
- Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap, *Ann. Statist.*, **9**, 1187–1195.
- van der Vaart, A. (1994). Weak convergence of smoothed empirical processes, *Scand. J. Statist.*, **21**, 501–504.
- von Mises, R. (1947). On the asymptotic distributions of differentiable statistical functions, *Ann. Math. Statist.*, **18**, 309–348.
- Wang, S. (1989). On the bootstrap and smoothed bootstrap, *Comm. Statist. Theory Methods*, **18**(11), 3949–3962.
- Wu, T. J. (1995). Adaptive root n estimates of integrated squared density derivatives, *Ann. Statist.*, **23**, 1474–1495.
- Young, G. A. (1990). Alternative smoothed bootstraps, *J. Roy. Statist. Soc. Ser. B*, **52**(3), 477–484.