Ann Inst Stat Math (2012) 64:919–943
DOI 10.1007/s10463-011-0339-4

On estimating distribution functions using Bernstein polynomials

Alexandre Leblanc

Received: 9 May 2008 / Revised: 25 July 2011 / Published online: 20 November 2011

Abstract It is a known fact that some estimators of smooth distribution functions can outperform the empirical distribution function in terms of asymptotic (integrated) mean-squared error. In this paper, we show that this is also true of Bernstein polynomial estimators of distribution functions associated with densities that are supported on a closed interval. Specifically, we introduce a higher order expansion for the asymptotic (integrated) mean-squared error of Bernstein estimators of distribution functions and examine the relative deficiency of the empirical distribution function with respect to these estimators. Finally, we also establish the (pointwise) asymptotic normality of these estimators and show that they have highly advantageous boundary properties, including the absence of boundary bias.

Keywords Bernstein polynomials · Distribution function estimation · Mean integrated squared error · Mean squared error · Asymptotic properties · Efficiency · Deficiency

1 Introduction

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables having a common unknown distribution function F with associated density f supported on a closed interval. Without loss of generality, we take that interval to be [0, 1]. Now, when F is known to be continuous, it is natural to consider the estimation of F by using smooth functions rather then the empirical distribution function, which is not continuous. One way of doing this, in the case where f is supported on the unit interval, is to make use of the famous Bernstein polynomial approximations. This is particularly appealing since

A. Leblanc (✉)
Department of Statistics, University of Manitoba, Winnipeg, MB R3T 2N2, Canada
e-mail: alex_leblanc@umanitoba.ca

Bernstein polynomials are known to yield very smooth estimates that typically have acceptable behaviour at the boundaries. This is the approach that will be considered in this paper.

Specifically, following Babu et al. (2002), the Bernstein estimator of order $m > 0$ of the distribution F is defined as

$$\hat{F}_{m,n}(x) = \sum_{k=0}^{m} F_n(k/m) P_{k,m}(x), \tag{1}$$

where $P_{k,m}(x) = \binom{m}{k} x^k (1-x)^{m-k}$ are binomial probabilities and F_n denotes the empirical distribution function obtained from a random sample of size n. Throughout this paper, we assume that $m = m_n$ depends on n. The suffix n will however be omitted for the sake of clarity. Note that $\hat{F}_{m,n}$ is a polynomial of degree m with coefficients depending on the data and, thus, leads to very smooth estimates. Note also that taking the derivative of $\hat{F}_{m,n}$ with respect to x leads to

$$\hat{f}_{m,n}(x) = \frac{\mathrm{d}}{\mathrm{d}x} \hat{F}_{m,n}(x) = m \sum_{k=0}^{m-1} \left[F_n\big([k+1]/m\big) - F_n(k/m) \right] P_{k,m-1}(x), \tag{2}$$

which is the Bernstein density estimator of order m, as it is defined by Babu et al. (2002) and many others.

Now, let B_m denote the Bernstein polynomial of order m of F according to

$$B_m(x) = \sum_{k=0}^{m} F(k/m) P_{k,m}(x).$$

A quick inspection of (1) and the previous equality makes it clear that $\mathbb{E}\big[\hat{F}_{m,n}(x)\big] = B_m(x)$ for all $x \in [0, 1]$ and all $n \geq 1$. Note also that B_m is a genuine distribution function and that $\hat{F}_{m,n}$ yields, with probability one and for any value of m, estimates that are genuine distribution functions. To see this, notice that

$$\hat{F}_{m,n}(0) = 0 = F(0) = B_m(0) \qquad \text{and} \qquad \hat{F}_{m,n}(1) = 1 = F(1) = B_m(1), \tag{3}$$

with probability one for all values of m, and that both functions have a nonnegative first derivative over the unit interval. See Babu et al. (2002) and Lorentz (1986, Section 1.7) for more details.

Bernstein polynomial estimators of density functions have become quite popular and recently attracted a lot of attention. See, for instance, the original work of Vitale (1975) and recent extensions/generalizations by Tenbusch (1994), Babu et al. (2002), Kakizawa (2004), Rao (2005), Babu and Chaubey (2006) and Leblanc (2010). Working from a completely different perspective, Petrone (1999) introduced a fully Bayesian approach to nonparametric density estimation on a compact interval through the use of Bernstein polynomials. This approach was further studied by Ghosal (2001) and Petrone and Wasserman (2002). Bernstein-based or related approaches to other

problems of nonparametric function estimation have also been developed by different authors. For example, Tenbusch (1997) and Brown and Chen (1999) have suggested different regression methods, Choudhuri et al. (2004) have developed a Bayesian approach to spectral density estimation and Chang et al. (2005) have developed a Bayesian approach to the estimation of cumulative hazard functions.

In light of this, it is surprising that the estimator defined in (1) has not attracted more attention in the literature. Babu et al. (2002) have shown it to be uniformly strongly consistent when $m, n \to \infty$. Leblanc (2009) has shown it to have the Chung–Smirnov property, which quantifies its extreme fluctuations (about F) as $n \to \infty$. Specifically, he showed that, under fairly general conditions on m and F, we have

$$\limsup_{n\to\infty} (2n/\log\log n)^{1/2} \sup_{x\in[0,1]} \left|\hat{F}_{m,n}(x) - F(x)\right| \leq 1, \qquad \text{almost surely,}$$

and that the equality actually holds under slightly more restrictive conditions. Finally, Babu and Chaubey (2006) considered the problem of estimating a multivariate distribution function by using Bernstein polynomials in multiple dimensions.

In an attempt to partly fill the gap in the literature related to the estimator defined in (1), we show that it outperforms the empirical distribution function in terms of asymptotic mean-squared error (MSE) and mean-integrated squared error (MISE). We also establish the (pointwise) asymptotic normality of this estimator. Kernel estimators of distribution functions are known to have these properties. See, for instance, the work of Azzalini (1981) and Jones (1990) for the asymptotic MSE and MISE properties and Watson and Leadbetter (1964) for the asymptotic normality of these estimators.

Specifically, in Sect. 2 we derive the MSE properties of $\hat{F}_{m,n}$. In Sect. 3, we show that the estimator is asymptotically normal for appropriate choices of m. In Sect. 4, we obtain the MISE properties of the estimator. In Sect. 5, we specifically address the issue of asymptotic efficiency and the notion of deficiency to conclude that the Bernstein estimator $\hat{F}_{m,n}$ asymptotically outperforms the empirical distribution function locally, in terms of MSE, and globally, in terms of MISE, for certain choices of the order m of the estimator. In Sect. 6, we present a brief numerical example that highlights some of the theoretical results obtained in the paper. Finally, in Sect. 7, we present a simulation study that compares the performance of the Bernstein estimator $\hat{F}_{m,n}$ with the empirical distribution function and with a standard Gaussian kernel estimator.

2 Some basic results

We start by considering some basic properties of the family of estimators defined in (1). Specifically, we focus on establishing the bias, variance and mean-squared error properties of the Bernstein estimator $\hat{F}_{m,n}$. First, note that (3) implies that the estimator $\hat{F}_{m,n}$ has very advantageous behaviour at the boundary points. Indeed, this estimator is unbiased and has zero variance at $x = 0, 1$. To eventually obtain the behaviour of

the estimator inside the unit interval, we make the assumption that

$$F \text{ is continuous and admits two continuous and bounded derivatives on } [0, 1], \tag{4}$$

and start by giving a result that can be found in Lorentz (1986, Section 1.6).

Lemma 1 *Under assumption* (4)*, we have for* $x \in (0, 1)$ *that*

$$B_m(x) = F(x) + m^{-1}b(x) + o(m^{-1}),$$

where $b(x) = x(1 - x)f'(x)/2$*. Finally, for the trivial case where* f *is the uniform density* (*and only in that case*)*, we have that* $B_m(x) = F(x) = x$ *for all* $m \geq 1$ *and* $x \in [0, 1]$. □

It should be noted that here, and throughout this paper, we use o and O in the usual way to denote a uniform bound (with respect to x) on an error of approximation. A pointwise bound in x will be emphasized by using o_x and O_x, as nonuniform error bounds have important implications in the derivations of some of our results. Note that the proofs of all of our main results can be found in the Appendix. We are now ready to state the basic properties of the Bernstein estimator $\hat{F}_{m,n}$.

Theorem 1 *Under assumption* (4)*, we have for* $x \in (0, 1)$ *that*

$$\mathbb{B}\text{ias}\big[\hat{F}_{m,n}(x)\big] = \mathbb{E}\big[\hat{F}_{m,n}(x)\big] - F(x) = m^{-1}b(x) + o(m^{-1}),$$

where $b(x)$ *is defined as in Lemma* 1*. Also, we have*

$$\mathbb{V}\text{ar}\big[\hat{F}_{m,n}(x)\big] = n^{-1}\sigma^2(x) - m^{-1/2}n^{-1}V(x) + o_x(m^{-1/2}n^{-1}),$$

where

$$\sigma^2(x) = F(x)\big[1 - F(x)\big] \quad \text{and} \quad V(x) = f(x)\big[2x(1 - x)/\pi\big]^{1/2},$$

as both $m, n \to \infty$. □

Notice that the previous result implies $\hat{F}_{m,n}$ has uniform bias inside the unit interval in addition to being unbiased at the boundary. Obviously, this estimator is then free of boundary bias. On the other hand, from its definition given in (1), it seems natural to consider $h = 1/m$ as the "bandwidth" of the Bernstein estimator. Doing so, Lemma 1 suggests that the bias of $\hat{F}_{m,n}$ is $O(m^{-1}) = O(h)$, which is more than the bias typically obtained using kernel estimators generally having a bias at least as small as $O(h^2)$ (except possibly near the boundaries).

Another consequence of the previous result is that $\hat{F}_{m,n}$ can asymptotically outperform the empirical distribution function at every $x \in (0, 1)$ in terms of MSE.

(Both estimators achieve an MSE of zero at $x = 0, 1$.) Indeed, from Theorem 1, we have that

$$\begin{aligned}\mathrm{MSE}\big[\hat{F}_{m,n}(x)\big] &= n^{-1}\sigma^2(x) - m^{-1/2}n^{-1}V(x) + m^{-2}b^2(x) \\ &\quad + o(m^{-2}) + o_x(m^{-1/2}n^{-1}).\end{aligned} \tag{5}$$

On the other hand, it is well known that

$$\mathrm{MSE}[F_n(x)] = \mathbb{V}\mathrm{ar}[F_n(x)] = n^{-1}\sigma^2(x),$$

so that $\hat{F}_{m,n}$ and F_n are equivalent in MSE up to the first-order. However, when considering also higher order terms, it turns out that $\hat{F}_{m,n}$ asymptotically dominates F_n in terms of MSE when m is chosen carefully. This comes from the fact that the second term on the right-hand side of (5) is always negative, and is formally established in the next corollary. A thorough investigation of the conditions under which Bernstein estimators outperform the empirical distribution function F_n is postponed until Sect. 5.

Corollary 1 *Assuming (4), $f(x) \neq 0$ and $f'(x) \neq 0$ all hold, the asymptotically optimal choice of m for estimating $F(x)$, with respect to MSE, is*

$$m_{\mathrm{opt}} = n^{2/3}\left[\frac{4b^2(x)}{V(x)}\right]^{2/3},$$

in which case

$$\mathrm{MSE}\big[\hat{F}_{m_{\mathrm{opt}},n}(x)\big] = n^{-1}\sigma^2(x) - n^{-4/3}\,\frac{3}{4}\left[\frac{V^4(x)}{4b^2(x)}\right]^{1/3} + o_x(n^{-4/3}),$$

for $x \in (0, 1)$, where $\sigma^2(x)$, $b(x)$ and $V(x)$ are defined as in Theorem 1. □

We note that other results similar to this have been obtained for different estimators of smooth distribution functions. For example, see Read (1972) for an estimator based on linear interpolation and Azzalini (1981) for the case of kernel estimators.

Before we move on to study the global properties of the Bernstein estimator, we next complete our study of the local first-order properties of the Bernstein estimator by focusing on the limiting distribution of $\hat{F}_{m,n}(x)$ for given values of $x \in (0, 1)$.

3 Asymptotic normality

In this section, we establish the asymptotic normality of the Bernstein estimator $\hat{F}_{m,n}$ at every x inside the unit interval. In essence, we will establish that when the order m of the Bernstein estimator is chosen large enough (so that bias becomes negligible), the asymptotic distribution of $\hat{F}_{m,n}$ is the same as that of the empirical distribution F_n. Note that this is a property that kernel estimators of distribution functions are known to have. This will be addressed again shortly. We first state a general result

that establishes the asymptotic normality of the Bernstein estimator for any choice of $m \to \infty$ when $n \to \infty$.

Theorem 2 *Assume* (4) *holds and* $m, n \to \infty$. *For* $x \in (0, 1)$ *such that* $0 < F(x) < 1$, *we have that*

$$n^{1/2}\big(\hat{F}_{m,n}(x) - B_m(x)\big) \xrightarrow{\mathcal{D}} N\big(0, \sigma^2(x)\big),$$

where $\sigma^2(x)$ *is defined as in Theorem* 1 *and "*$\xrightarrow{\mathcal{D}}$*" denotes convergence in distribution.* □

Notice how the previous result contrasts with that obtained in the case of density estimation, where asymptotic normality holds for m values that are large enough, but not too large. Indeed, Babu et al. (2002, Proposition 1) showed that in the density estimation setting, for their asymptotic normality result to hold, we need that $mn^{-2/3} \to \infty$, and also that $mn^{-1} \to 0$. This is not the case here as asymptotic normality holds for any m such that $m \to \infty$, with no restriction whatsoever on the rate at which m increases. Note also that, under an appropriate choice of bandwidth, a result similar to Theorem 2 has been obtained by Watson and Leadbetter (1964) for general kernel estimators of distribution functions.

Now, as interest is mainly in how $\hat{F}_{m,n}(x)$ behaves with respect to $F(x)$, we note that, from Lemma 1, we have

$$n^{1/2}\big(\hat{F}_{m,n}(x) - F(x)\big) = n^{1/2}\big(\hat{F}_{m,n}(x) - B_m(x)\big) + m^{-1}n^{1/2}b(x) + o(m^{-1}n^{1/2}),$$

with $b(x)$ (defined as in Lemma 1) being bounded over the unit interval. This leads to the following result.

Corollary 2 *Assume* (4) *holds and* $m, n \to \infty$. *Then, for* $x \in (0, 1)$ *such that* $0 < F(x) < 1$,

(i) *if* $mn^{-1/2} \to \infty$,

$$n^{1/2}\big(\hat{F}_{m,n}(x) - F(x)\big) \xrightarrow{\mathcal{D}} N\big(0, \sigma^2(x)\big),$$

(ii) *if* $mn^{-1/2} \to c$ *for some constant* $c > 0$,

$$n^{1/2}\big(\hat{F}_{m,n}(x) - F(x)\big) \xrightarrow{\mathcal{D}} N\big(c^{-1}b(x), \sigma^2(x)\big),$$

where $\sigma^2(x)$ *and* $b(x)$ *are defined as in Theorem* 1. □

Note that (i) can be derived without using Theorem 2 by relying instead on Theorem 4 of Leblanc (2009) (which basically states that, under a smoothness assumption on F and an appropriate choice for the order of the Bernstein estimator, the distance between $\hat{F}_{m,n}(x)$ and $F_n(x)$ is "small" enough with probability one) and the fact that $F_n(x)$ has itself a well-known asymptotic normal distribution.

We now point out that the phenomenon observed by Hjort and Walker (2001) in the context of kernel estimation is also observed with Bernstein polynomial estimators. Specifically, Hjort and Walker (2001) proved that MISE optimal bandwidths for density estimation, when using kernel estimators, lead to density estimates for which the associated estimate of the distribution function F has the property of lying outside of reasonable confidence bands for F (based on the empirical distribution function F_n), with probability tending to one. This phenomenon is linked to the fact that the MISE optimal bandwidths for density estimation, in that context, satisfy $hn^{1/5} \to c$ for some finite constant $c > 0$, while it is necessary that $hn^{1/4} \to 0$ for the kernel estimator of the distribution function to have a limiting distribution centred at $F(x)$ when properly rescaled.

As was mentioned above, this phenomenon is also observed with Bernstein polynomial estimators. Indeed, the MISE optimal choice of the order m of Bernstein estimators, in the context of density estimation, satisfies $mn^{-2/5} \to c$ for some constant $c > 0$. See, for instance, Babu et al. (2002) and Leblanc (2010). However, it is not difficult to see that if $mn^{-1/2} \to 0$ and $f'(x) \neq 0$, then

$$\mathbb{P}\Big[n^{1/2}\big|\hat{F}_{m,n}(x) - F(x)\big| > \varepsilon\Big] \longrightarrow 1,$$

for all $\varepsilon > 0$. According to this, if the MISE optimal choice of m is used for density estimation, the estimator $\hat{F}_{m,n}$ of the distribution function associated with the density estimator does not converge in distribution (for this choice of m) to a limiting distribution centred at $F(x)$ when properly rescaled. It is not difficult to see that, as a result, $\hat{F}_{m,n}$ will also lie outside of confidence bands based on F_n with probability tending to one.

4 MISE of the Bernstein estimator

We now obtain the mean-integrated squared error (MISE) of the Bernstein estimator as given by (1). It is important to note that this result is not obtained through integrating the expression for $\text{MSE}\big[\hat{F}_{m,n}(x)\big]$ obtained in (5), even though intuitively one might think of it in that way. This is because of the nonuniformity (with respect to x) of the error term in the asymptotic expression for the variance of the Bernstein estimator obtained in Theorem 1.

We here define the MISE of an estimator $\hat{F}$ of the distribution function F defined on the unit interval as

$$\text{MISE}\big[\hat{F}\big] = \mathbb{E}\left[\int_0^1 \big[\hat{F}(x) - F(x)\big]^2 \mathrm{d}x\right], \tag{6}$$

and turn our attention to $\text{MISE}\big[\hat{F}_{m,n}\big]$. Following Altman and Léger (1995) and many others, it would have also been possible to define the MISE of an estimator $\hat{F}$ by

$$\text{MISE}\big[\hat{F}\big] = \mathbb{E}\left[\int_0^1 \big[\hat{F}(x) - F(x)\big]^2 W(x) f(x)\, \mathrm{d}x\right],$$

where W is a nonnegative weighting function. Given it is assumed that X is supported on the unit interval, there is no obvious benefit to using this second definition, and so we work with the slightly simpler definition provided by (6). Note, however, that our next result could easily be adapted to account for such a modification.

Theorem 3 *Under assumption (4), we have that*

$$\mathrm{MISE}\big[\hat{F}_{m,n}\big] = n^{-1}C_1 - m^{-1/2}n^{-1}C_2 + m^{-2}C_3 + o(m^{-1/2}n^{-1}) + o(m^{-2}),$$

where

$$C_1 = \int_0^1 \sigma^2(x)\,\mathrm{d}x, \qquad C_2 = \int_0^1 V(x)\,\mathrm{d}x, \qquad \textit{and} \qquad C_3 = \int_0^1 b^2(x)\,\mathrm{d}x,$$

and $\sigma^2(x)$, $b(x)$ and $V(x)$ are defined as in Theorem 1. □

Note that the constants C_1, C_2 and C_3 are all strictly positive, except in the trivial case where f is the uniform density, in which case $C_3 = 0$. The following result is a direct consequence of the previous theorem and identifies the asymptotically optimal order m of the Bernstein estimator with respect to MISE. It also establishes the fact that, for a carefully chosen value of m, $\hat{F}_{m,n}$ asymptotically dominates F_n in terms of their MISE performance.

Corollary 3 *If assumption (4) holds and if $C_3 > 0$ (see above), the asymptotically optimal choice of m for estimating F, with respect to MISE, is*

$$m_{\mathrm{opt}} = n^{2/3}\left[\frac{4C_3}{C_2}\right]^{2/3},$$

in which case

$$\mathrm{MISE}\big[\hat{F}_{m_{\mathrm{opt}},n}\big] = n^{-1}C_1 - n^{-4/3}\frac{3}{4}\left[\frac{{C_2}^4}{4C_3}\right]^{1/3} + o(n^{-4/3}),$$

where the constants C_1, C_2 and C_3 are defined as in Theorem 3. □

Results similar to this have been obtained for general kernel estimators by Jones (1990), among others. The selection of m for specific data sets, although an interesting problem, will not be addressed here. Notice, however, that the plug-in approach suggested by Altman and Léger (1995) and the cross-validation method of Bowman et al. (1998) for estimating smooth distribution functions using kernel estimators could certainly be adapted to the current context.

5 Deficiency of the empirical distribution function

In this section, we focus on the relative deficiency of the empirical distribution function with respect to the Bernstein estimator $\hat{F}_{m,n}$. In doing this, our goal is to better

appreciate the performance of the two estimators and better understand the differences between the two. Indeed, as was pointed out in Sect. 2, the first-order properties of the two estimators are the same, so that the second-order properties have to be considered if one is to really compare these estimators.

Following the work of Hodges and Lehman (1970), we define $i_L(n, x)$ to be the sample size required for the empirical distribution function to have the same (or smaller) MSE as $\hat{F}_{m,n}$ at the point x, that is

$$i_L(n, x) = \min\left\{k \in \mathbb{N} : \mathrm{MSE}[F_k(x)] \leq \mathrm{MSE}[\hat{F}_{m,n}(x)]\right\}.$$

A local comparison of the two estimators can now be made, at the point x, by comparing $i_L(n, x)$ with n. Indeed, the usual notion of asymptotic relative efficiency is now simply the limiting behaviour of the ratio $i_L(n, x)/n$. Obviously, when two estimators share the same first-order properties, one should find that this ratio converges to one. What is of interest, in those cases, is the limiting behaviour of the difference $i_L(n, x) - n$, known as (local) *asymptotic deficiency*. In the current context, this corresponds to the number of additional observations required for the empirical distribution function to perform at least as well as the Bernstein estimator, in terms of MSE, at the point x.

To compare the global performance of F_n with that of the Bernstein estimator $\hat{F}_{m,n}$, one can instead focus on the deficiency in MISE. For this, we define

$$i_G(n) = \min\left\{k \in \mathbb{N} : \mathrm{MISE}[F_k] \leq \mathrm{MISE}[\hat{F}_{m,n}]\right\},$$

and consider the limiting behaviour of the ratio $i_G(n)/n$ and of the difference $i_G(n) - n$. The following result establishes conditions under which F_n is asymptotically efficient (to the first order), but asymptotically deficient (locally in MSE and globally in MISE) with respect to $\hat{F}_{m,n}$. It also gives this asymptotic deficiency in closed form.

Theorem 4 *Assume that* (4) *holds,* $x \in (0, 1)$ *and that* $m, n \to \infty$. *Then, if* $mn^{-1/2} \to \infty$, *we have that*

$$i_L(n, x) = n\big[1 + o_x(1)\big] \quad \text{and} \quad i_G(n) = n\big[1 + o(1)\big].$$

In addition,

(i) *if* $mn^{-2/3} \to \infty$ *and* $mn^{-2} \to 0$, *then*

$$i_L(n, x) - n = m^{-1/2} n\big[\theta(x) + o_x(1)\big],$$
$$\text{and} \quad i_G(n) - n = m^{-1/2} n\big[C_2/C_1 + o(1)\big],$$

(ii) *if* $mn^{-2/3} \to c$ *for some constant* $c > 0$,

$$i_L(n, x) - n = n^{2/3}\big[c^{-1/2}\,\theta(x) - c^{-2}\,\gamma(x) + o_x(1)\big],$$

and

$$i_G(n) - n = n^{2/3}\big[c^{-1/2}\, C_2/C_1 - c^{-2}\, C_3/C_1 + o(1)\big],$$

where

$$\theta(x) = V(x)/\sigma^2(x) \quad \text{and} \quad \gamma(x) = b^2(x)/\sigma^2(x),$$

where $V(x)$, $\sigma^2(x)$ and $b(x)$ are defined as in Theorem 1, and C_1, C_2 and C_3 are defined as in Theorem 3. □

Note that the case of local deficiency in MSE where $x = 0$ or 1 is not covered in the previous result. Actually, in that case, it is trivial to show that

$$i_L(n, 0) = i_L(n, 1) = n,$$

this being true for any choice of $m > 0$.

Also, Theorem 4 should be interpreted as an indicator of when Bernstein estimators outperform (locally, in MSE, and globally, in MISE) the empirical distribution function in a significant way. Indeed, we point out the fact that (i) and (ii) identify setups where the asymptotic deficiency of F_n grows to infinity with n. This observation gives a different view of the seemingly small difference in MSE and MISE between the estimators considered here. Indeed, as n increases and even if the difference in MSE (or MISE) seems relatively small, one needs increasingly many more observations to see a reduction in MSE (or MISE), using the empirical distribution function, of the same order as that which would be obtained by instead using a Bernstein estimator of a carefully selected order m without increasing the sample size.

To our knowledge, Aggarwal (1995) was the first to exhibit an estimator of distribution functions that dominates the empirical estimator F_n in terms of MISE. Similarly, Read (1972) was the first to exhibit a continuous estimator of smooth distribution functions that dominates the empirical estimator F_n in terms of MSE. Note that both of these authors did not discuss deficiency in their work. We also mention that the deficiency in MSE of the empirical distribution function with respect to kernel estimators has been first established by Reiss (1981) and later obtained in a form similar to our Theorem 4 by Falk (1983).

As a final comment, we point out that the selection of an optimal order m of the Bernstein estimator could be made based on deficiency. Indeed, it seems reasonable to consider choosing m in such a way as to maximize the deficiency of the empirical distribution with respect to the Bernstein estimator, thus making sure the former is outperformed by the latter as much as possible, for example, in terms of MISE. Obviously, doing this is justified only if one thinks of the empirical distribution function as a reference or standard that should be outperformed. It should probably come with no surprise, however, that this leads to the same choice of the optimal order m_{opt} as identified in Corollary 3, as can be seen from the following simple argument.

First, our goal is to maximize the deficiency of the empirical estimator F_n; note that when $mn^{-2/3} \to c$, the asymptotic deficiency of F_n is positive only when

$$c > \left[C_3/C_2\right]^{2/3} = c^*.$$

In this case, the asymptotic deficiency of F_n is of the order of $n^{2/3}$, the largest it actually can be. This suggests choosing m so that $mn^{-2/3} \to c$, where $c > c^*$ is chosen to maximize

$$g(c) = c^{-1/2} C_2/C_1 - c^{-2}\, C_3/C_1.$$

Elementary calculations lead to the previous expression being maximized when

$$c = c_{\text{opt}} = \left[4C_3/C_2\right]^{2/3} = 2^{4/3} c^*,$$

leading, in turn, to the deficiency-based optimal order of the Bernstein estimator satisfying $m_{\text{opt}} n^{-2/3} \to c_{\text{opt}}$, or

$$m_{\text{opt}} = n^{2/3}\left[c_{\text{opt}} + o(1)\right].$$

This is in agreement with the result obtained earlier in Corollary 3 based on minimizing the MISE of the Bernstein estimator.

6 Numerical example

We consider an example that highlights the features of the Bernstein estimator $\hat{F}_{m,n}$. Specifically, we look at the so-called *suicide data* given in Table 2.1 of Silverman (1986). These data consist of durations (in days) of psychiatric treatment for 86 patients used as controls in a study of suicide risks. They are an example of data leading to problematic behaviour of typical density estimators close to a boundary (e.g. see Leblanc 2010). It is clear in this setup that the distribution function to be estimated is defined only for $x > 0$. For convenience, we also assume that the maximum treatment duration is 800 days (the data are such that $\min_i(x_i) = 1$ and $\max_i(x_i) = 737$) and analyse the original data rescaled to the unit interval. Of utmost interest is the behaviour of estimators near $x = 0$.

In Fig. 1, we display different Bernstein estimators of the underlying density f of treatment durations along with a histogram of the data. Specifically, we graphed the estimator $\hat{f}_{m,n}$ introduced in (2) for $m = 5$, 10, 19 and 60. Note that $m = 19$ is the data-driven optimal choice of m based on least-squares cross-validation for the density estimation problem (cf. Leblanc 2010). It is obvious here that the choices of $m = 5$ and 10 lead to considerable oversmoothing. On the other hand, the choice of $m = 60$ leads to an undesirable feature at $x = 0$ and is actually undersmoothing.

Different Bernstein estimators $\hat{F}_{m,n}$ of the underlying distribution function F of treatment durations are pictured in Fig. 2. Also shown on this graph is the empirical distribution function constructed from the data. The oversmoothing, in the cases of

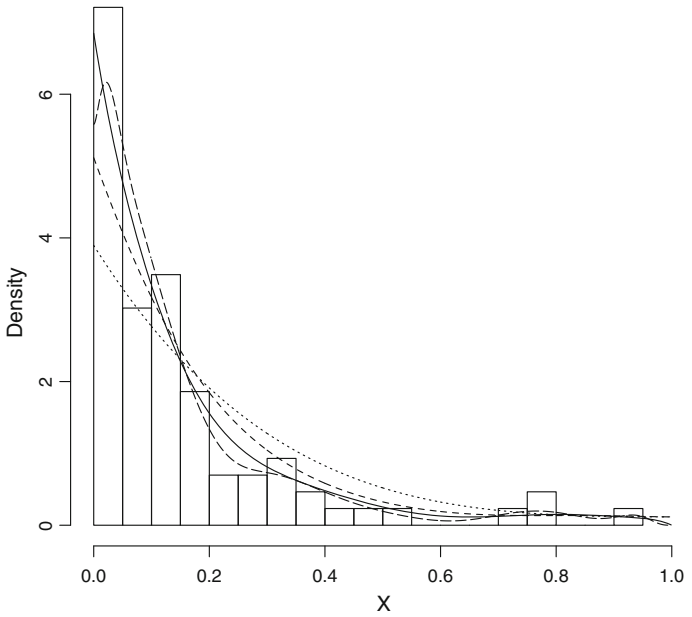

Fig. 1 Bernstein density estimates $\hat{f}_{m,n}$ obtained with $m = 5$ (*dotted line*), $m = 10$ (*short-dashed line*), $m = 19$ (*full line*) and $m = 60$ (*long-dashed line*)

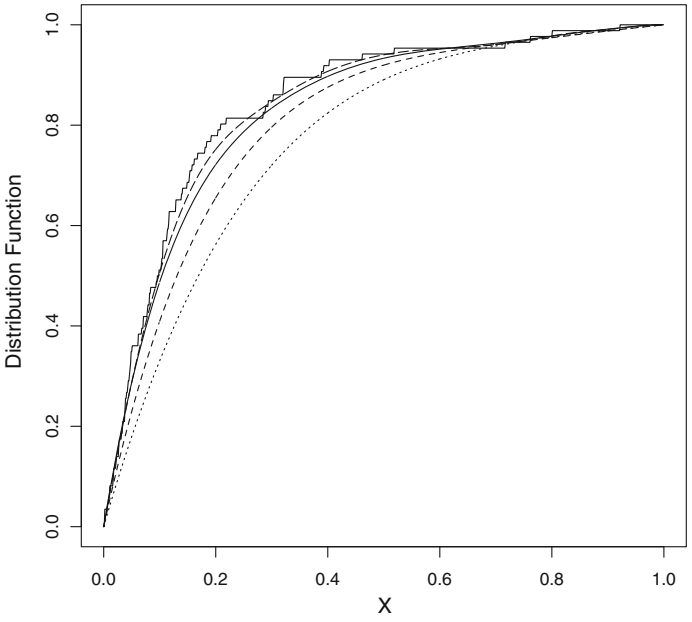

Fig. 2 Bernstein estimates $\hat{F}_{m,n}$ of the distribution function calculated with $m = 5, 10, 19$ and 60 as above

$m = 5$ and 10, is again quite apparent. Note, however, that the choice of $m = 19$ also leads to oversmoothing in this case, and that $m = 60$ seems to be the most appropriate choice for m among the values considered here.

This last point is particularly interesting. Indeed, from Theorem 3 of Leblanc (2010), we know that $m_{\text{opt}} \propto n^{2/5}$ in the density estimation setup. On the other hand, Corollary 3 establishes that $m_{\text{opt}} \propto n^{2/3}$ when estimating a distribution function. In other words, the asymptotically optimal order of the Bernstein density estimator is much smaller than the optimal order used for estimating a distribution function. Hence, what we see here is in agreement with these two asymptotic results: it seems that optimal smoothing for density estimation leads to oversmoothing when considering the distribution function. This is linked to the earlier discussion presented after Corollary 2 in Sect. 3, and can also be observed with kernel estimators of density and distribution functions (cf. Hjort and Walker 2001).

7 Simulation study

Babu et al. (2002) presented a short simulation study looking at the behaviour of the Bernstein estimator $\hat{F}_{m,n}$ (and at the density estimator $\hat{f}_{m,n}$). However, they have not considered comparing its performance with that of the empirical distribution F_n or any kernel estimator. This is what we do in this section.

Specifically, we study the performance of the Bernstein estimator in estimating different distributions by comparing it to the performances of F_n and of the Gaussian kernel estimator

$$\hat{K}_{h,n}(x) = \frac{1}{n} \sum_{i=1} \Phi\Big(\frac{x - X_i}{h}\Big),$$

where Φ denotes the standard normal distribution function and h is the bandwidth of the estimator. See, for instance, Altman and Léger (1995) and Bowman et al. (1998) for more discussion on kernel estimators of distribution functions and, in particular, on bandwidth selection. We also refer the reader to the papers of Swanepoel and Van Graan (2005), Liu and Yang (2008) and Chacón and Rodríguez-Casal (2010) for recent work on kernel estimators of distribution functions.

As a measure of performance, we use the MISE of each of the mentioned estimators, as defined in (6). In the case of the Bernstein and kernel estimators, the MISE value depends, respectively, on the order m and the bandwidth h that are considered. Specifically, let

$$\text{ISE}\big[\hat{F}\big] = \int_0^1 \big[\hat{F}(x) - F(x)\big]^2 \mathrm{d}x, \tag{7}$$

and note that, from M pseudo-random samples of size n,

$$\text{MISE}\big[\hat{F}\big] \simeq \frac{1}{M} \sum_{i=1}^{M} \text{ISE}_i\big[\hat{F}\big]$$

Table 1 Summary of simulation study, all approximated MISE values $\times 10^{-3}$

	n	F_n	$\hat{F}_{m,n}$		$\hat{K}_{h,n}$	
		MISE	MISE	m_{opt}	MISE	h_{opt}
Beta(2,1)	20	6.70	3.84	7	4.42	0.143
	50	2.69	1.78	11	2.02	0.100
	100	1.34	0.96	16	1.08	0.075
Beta(10,10)	20	3.11	2.06	61	2.10	0.058
	50	1.25	0.92	114	0.93	0.044
	100	0.62	0.48	174	0.48	0.035
Truncated $N(1/2, 1/4)$	20	9.14	7.06	11	7.49	0.115
	50	3.69	3.11	27	3.28	0.062
	100	1.85	1.63	43	1.71	0.043
$1/2\,\text{Beta}(2.5,6) + 1/2\,\text{Beta}(9,1)$	20	6.20	3.68	10	4.11	0.124
	50	2.51	1.71	17	1.84	0.093
	100	1.25	0.92	26	0.97	0.075

is a Monte Carlo approximation of $\text{MISE}\big[\hat{F}\big]$, where $\text{ISE}_i\big[\hat{F}\big]$ denotes the value of ISE calculated from the ith randomly generated sample from F and obtained from (7).

We ran simulations using four different underlying distribution functions on the unit interval: the Beta(2,1) (with linear density), the Beta(10,10) (with density concentrated around 1/2), the N(1/2, 1/4) truncated to the unit interval (smooth, but positive density at the boundaries) and the mixture $1/2\,\text{Beta}(2.5,6) + 1/2\,\text{Beta}(9,1)$ (asymmetric density, bimodal with a mode at a boundary). In each case, we approximated the MISE of F_n, $\hat{F}_{m,n}$ (for integers $2 \leq m \leq 200$) and $\hat{K}_{h,n}$ (for $h = i/1000$ with integers $1 \leq i \leq 200$) using $M = 10{,}000$ pseudo-random samples of sizes $n = 20$, 50 and 100. We next summarize our findings.

First, we observe that, in all cases presented in Table 1, both smooth estimators do better than the empirical distribution function F_n for appropriate choices of the smoothing parameters. Indeed, we see that for $n = 20$, the potential reduction in MISE ranges between 23 and 43% for the Bernstein estimator, and between 18 and 34% for the Gaussian kernel estimator, when compared with F_n. For $n = 50$, this reduction is between 16 and 34% for the Bernstein estimator and between 11 and 27% for the kernel estimator. Finally, for $n = 100$, the reduction is between 12 and 28% for the former, and between 8 and 23% for the latter. These results are in line with our Corollary 3 and with the comments of Swanepoel and Van Graan (2005) and others suggesting the benefits of smoothing in the case of distribution function estimation.

Our second observation is that, from the previous perspective, the Bernstein estimator does better than its kernel counterpart in all the presented cases. Obviously, there might be other kernel estimators that do better than the Gaussian kernel estimator used here, but this suggests that there could be interesting gains in MISE reduction when considering using the Bernstein estimator $\hat{F}_{m,n}$ over simple standard kernel estimators like $\hat{K}_{h,n}$.

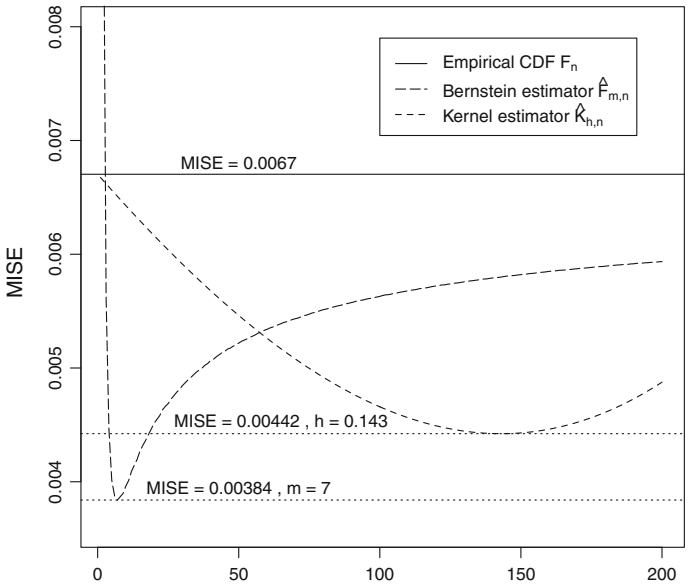

Fig. 3 Approximated MISE of F_n, of the Bernstein estimator and of the Gaussian kernel estimator for the Beta(2,1) distribution and $n = 20$. The x-axis displays values of m, for the Bernstein estimator, and of $h \times 10^3$ for the kernel estimator

To further investigate this, we plotted, in Fig. 3, the MISE of both smooth estimators, as functions of their respective smoothing parameter m and h, and added the MISE of the empirical distribution for the case where the true underlying distribution is Beta(2,1) and $n = 20$. This highlights once again that smoothing is beneficial when estimating a distribution function. Indeed, for almost all the considered values of the smoothing parameters m and h, both smooth estimators have a reduced MISE compared to F_n. As mentioned above, this reduction is quite significant in the best cases. Going back to the comparison between $\hat{F}_{m,n}$ and $\hat{K}_{h,n}$, it is interesting to see that the MISE of the Bernstein estimator is smaller than the MISE of the optimal Gaussian kernel estimator (with $h = 0.143$) for values of m between 5 and 18 inclusive. Figure 4 tells a similar story for the case where $n = 100$, but the domination of $\hat{F}_{m,n}$ over the optimal kernel estimator (with $h = 0.075$) is now for m values between 10 and 61 inclusive.

Going back to Table 1, we see that the case of the Beta(10,10) distribution is the one where the optimal performance of the smooth estimators is most similar. This makes sense as the density of the Beta(10,10) is exactly zero at both boundaries and practically zero for $x < 0.1$ and $x > 0.9$, implying boundary issues should not play a big role for the Gaussian kernel estimator in this case. Note also that optimal smoothing is done here with much larger order m for the Bernstein estimator and much smaller bandwidth h for the kernel estimator. This was expected because the Beta(10,10) density is much more concentrated then the other three considered in the current study, suggesting that less smoothing is better in this case.

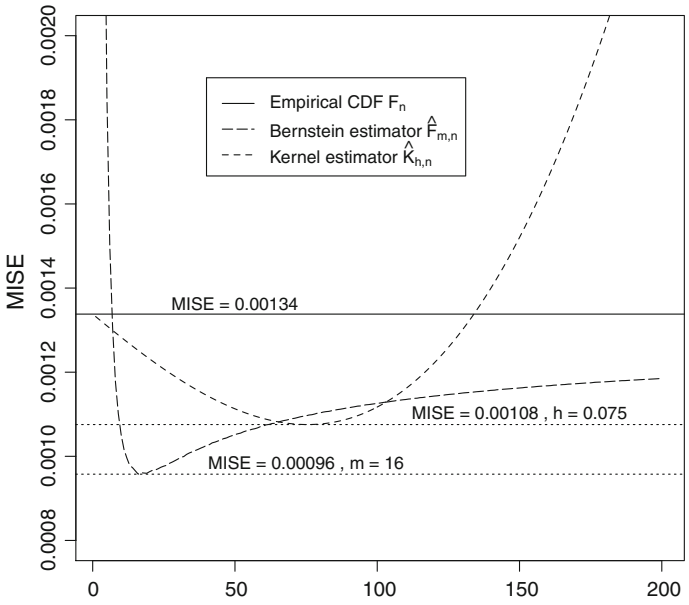

Fig. 4 Approximated MISE of F_n, of the Bernstein estimator and of the Gaussian kernel estimator for the Beta(2,1) distribution and $n = 100$. The x-axis displays values of m, for the Bernstein estimator, and of $h \times 10^3$ for the kernel estimator

Graphs similar to Figs. 3 and 4 were also obtained for the other combinations of underlying distributions and sample sizes given in Table 1 but are not shown here, since they highlight similar patterns.

8 Conclusion

The literature on Bernstein estimators of distribution functions is surprisingly sparse given the recent interest in using Bernstein polynomials for function estimation in different areas of statistics. In this article, we have shown that Bernstein estimators of distribution functions have very good boundary properties, including the absence of boundary bias. We also showed that a few important properties that contributed to the popularity of kernel estimators of distribution functions are also satisfied by Bernstein estimators. Mainly, we have shown that Bernstein estimators of distribution functions are asymptotically normal and first-order efficient. Using the concept of asymptotic deficiency, we also established that they asymptotically dominate the empirical distribution in terms of both MSE and MISE when the order m of the estimator is selected appropriately.

Through a simple real-life example and a small simulation study, we have shown how Bernstein estimators can lead to very satisfactory estimates of the underlying distribution. Finally, our simulations also suggest that the Bernstein estimator studied here behaves quite well when compared with both the empirical distribution F_n and the simple Gaussian kernel estimator.

Appendix

In this Appendix, we present proofs for selected results presented in the paper. However, we first present a series of results linked to different sums of binomial probabilities defined by

$$S_m(x) = \sum_{k=0}^{m} P_{k,m}^2(x),$$

and, for $j = 0$, 1 and 2,

$$R_{j,m}(x) = m^{-j} \sum\sum_{0 \le k < l \le m} (k - mx)^j P_{k,m}(x) P_{l,m}(x),$$

where $P_{k,m}(x) = \binom{m}{k} x^k (1-x)^{m-k}$ are the binomial probabilities. These results are given in the following lemma.

Lemma 2 *Let* $\psi_1(x) = \big[4\pi x(1-x)\big]^{-1/2}$ *and* $\psi_2(x) = \big[x(1-x)/(2\pi)\big]^{1/2}$. *Then the following results hold:*

(i) $0 \le S_m(x) \le 1$ *for* $x \in [0, 1]$,
(ii) $S_m(x) = m^{-1/2}\big[\psi_1(x) + o_x(1)\big]$ *for* $x \in (0, 1)$,
(iii) $S_m(0) = S_m(1) = 1$,
(iv) $R_{1,m}(x) = m^{-1/2}\big[-\psi_2(x) + o_x(1)\big]$ *for* $x \in (0, 1)$,
(v) $0 \le R_{2,m}(x) \le (4m)^{-1}$ *for* $x \in (0, 1)$,
(vi) $R_{j,m}(0) = R_{j,m}(1) = 0$ *for* $j = 0, 1, 2$.

Proof First note that (i), (iii) and (vi) trivially hold. The proof of (ii) is due to Babu et al. (2002, Lemma 3.1). We now turn to the proofs of (iv) and (v).

To prove (iv), we rely on Theorem 1 of Cressie (1978). Indeed, this result allows us to write

$$\sum_{l=k}^{m} P_{l,m}(x) = 1 - \Phi\big(\delta_k - G_x(\delta_{k-1/2})\big) + O_x(m^{-1}), \tag{8}$$

where the error term is independent of k, Φ stands for the normal distribution function,

$$\delta_k = (k - mx)\big[mx(1-x)\big]^{-1/2},$$

and

$$G_x(t) = \left[\frac{1}{2} + \frac{1}{6}(1-2x)(t^2-1)\right]\big[mx(1-x)\big]^{-1/2}.$$

Note that the correction factor $G_x(\delta_{k-1/2})$ is the reason why the normal approximation given in (8) is of order m^{-1}. This is crucial, as the uncorrected normal approximation

to the binomial tail probabilities is of order $m^{-1/2}$, which is not precise enough in the current context. Now, a Taylor series expansion of $\Phi(t)$ about $t = 0$ leads to

$$\Phi(t) = \frac{1}{2} + \frac{t}{\sqrt{2\pi}} + o(|t|),$$

so that we can actually write

$$\sum_{l=k+1}^{m} P_{l,m}(x) = \frac{1}{2} - \frac{\delta_{k+1} - G_x(\delta_{k+1/2})}{\sqrt{2\pi}} + o_x\big(|\delta_{k+1} - G_x(\delta_{k+1/2})|\big) + O_x(m^{-1}),$$

the last error term being independent of k. Although still fairly crude, we will see that this approximation will allow the derivation of an asymptotic expression for $R_{1,m}(x)$. Indeed, it can be shown that

$$\begin{aligned}\delta_{k+1} - G_x(\delta_{k+1/2}) =& \frac{1}{3}(2-x)\big[mx(1-x)\big]^{-1/2} \\ &+ \left[1 - \frac{1}{6}(1-2x)\big[mx(1-x)\big]^{-1}\right]\delta_k \\ &- \frac{1}{6}(1-2x)\big[mx(1-x)\big]^{-1/2}\delta_k^2 + O_x(m^{-5/2}),\end{aligned}$$

so that

$$\begin{aligned}R_{1,m}(x) =& m^{-1}\sum_{k=0}^{m}(k-mx)P_{k,m}(x)\left[\sum_{l=k+1}^{m} P_{l,m}(x)\right] \\ =& \left[\frac{1}{2} - \frac{1}{3}(2-x)\big[2\pi mx(1-x)\big]^{-1/2}\right]m^{-1}T_{1,m}(x) \\ &- \big[2\pi mx(1-x)\big]^{-1/2}m^{-1}T_{2,m}(x) \\ &+ o_x\big(m^{-3/2}H_{1,m}(x)\big) + o_x\big(m^{-3/2}H_{2,m}(x)\big) + O_x\big(m^{-5/2}H_{3,m}(x)\big),\end{aligned} \tag{9}$$

where

$$T_{j,m}(x) = \sum_{k=0}^{m}(k-mx)^j P_{k,m}(x), \qquad \text{and} \qquad H_{j,m}(x) = \sum_{k=0}^{m}|k-mx|^j P_{k,m}(x), \tag{10}$$

with $H_{j,m}(x) = T_{j,m}(x)$ for even values of j. Note that (cf. Lorentz 1986, Section 1.5) it is easy to obtain

$$\begin{aligned}&T_{1,m}(x) = 0, \quad T_{2,m}(x) = mx(1-x), \quad T_{3,m}(x) = mx(1-x)(1-2x), \\ &T_{4,m}(x) = 3m(m-2)x^2(1-x)^2 + mx(1-x).\end{aligned}$$

Hence, we have that

$$R_{1,m}(x) = -x(1-x)\big[2\pi mx(1-x)\big]^{-1/2} + o_x(m^{-1/2}) + o_x\big(m^{-3/2}H_{1,m}(x)\big) + O_x\big(m^{-5/2}H_{3,m}(x)\big). \tag{11}$$

However, note that the Cauchy–Schwartz inequality implies that

$$\begin{aligned} m^{-3/2}H_{1,m}(x) &\le m^{-3/2}\Big[T_{2,m}(x)\Big]^{1/2} \\ &= m^{-3/2}\big[mx(1-x)\big]^{1/2} \le (2m)^{-1} = O(m^{-1}), \end{aligned} \tag{12}$$

and that

$$m^{-5/2}H_{3,m}(x) \le m^{-5/2}\Big[T_{2,m}(x)T_{4,m}(x)\Big]^{1/2} = O(m^{-1}).$$

Substituting these two results into (11) leads to (iv).

Finally, (v) is easily proved since $R_{2,m}$ is clearly a nonnegative function and

$$R_{2,m}(x) \le m^{-2}\sum_{k=0}^{m}\sum_{l=0}^{m}(k-mx)^2 P_{k,m}(x)P_{l,m}(x) = m^{-2}T_{2,m}(x) = m^{-1}x(1-x),$$

so that $0 \le R_{2,m}(x) \le (4m)^{-1}$. □

Lemma 3 *Let g be any continuous function on* $[0, 1]$. *Then,*

(i) $m^{1/2}\displaystyle\int_0^1 S_m(x)\mathrm{d}x = \int_0^1 \psi_1(x)\mathrm{d}x + O(m^{-1}) = \sqrt{\pi}/2 + O(m^{-1}),$

(ii) $m^{1/2}\displaystyle\int_0^1 g(x)\,R_{1,m}(x)\mathrm{d}x = -\int_0^1 g(x)\psi_2(x)\mathrm{d}x + o(1),$

where ψ_1 and ψ_2 are defined as in Lemma 2.

Proof The proof of (i) can be found in Leblanc (2010, Lemma 4). We now prove (ii) using an approach similar to what was used there.

First, let $G_m(x) = m^{1/2}R_{1,m}(x)$ and $G(x) = -\psi_2(x)$ and note that Lemma 2 (iv) implies that G_m converges almost everywhere to G on the unit interval. On the other hand, from (9), (10) and (12), we have

$$\big|G_m(x)\big| \le m^{-1/2}H_{1,m}(x) \le 1/2,$$

for all m and $x \in [0, 1]$. Thus, the sequence is uniformly bounded on the unit interval and, hence, is also uniformly integrable. Now, the almost everywhere convergence and uniform integrability of G_m together imply that (cf. Theorem 16.14 and its Corollary of Billingsley 1995, pp. 217–218)

$$\int_0^1 \big|G_m(x) - G(x)\big|\mathrm{d}x = o(1),$$

i.e. the sequence also converges in L^1. This result also implies that

$$\left|\int_0^1 g(x)G_m(x)\,\mathrm{d}x - \int_0^1 g(x)G(x)\,\mathrm{d}x\right| \leq \sup_{x\in[0,1]}|g(x)|\int_0^1 \big|G_m(x) - G(x)\big|\mathrm{d}x = o(1),$$

which proves (ii) is also satisfied. □

Proof of Theorem 1 It is clear that

$$\mathbb{E}\big[\hat{F}_{m,n}(x)\big] = B_m(x), \tag{13}$$

for all $x \in [0, 1]$, so that the expression for the bias of $\hat{F}_{m,n}$ just follows from Lemma 1. Let us now focus on calculating the variance of our estimator. For this, we define for any $x \in [0, 1]$,

$$\Delta_i(x) = \mathbb{I}(X_i \leq x) - F(x),$$

where $\mathbb{I}(A)$ denotes the indicator function of the event A, so that $\Delta_1(x), \ldots, \Delta_n(x)$ are i.i.d. with mean zero. Note that

$$\hat{F}_{m,n}(x) - B_m(x) = \sum_{k=0}^{m}\big[F_n(k/m) - F(k/m)\big]P_{k,m}(x) = \frac{1}{n}\sum_{i=1}^{n} Y_{i,m},$$

where

$$Y_{i,m} = \sum_{k=0}^{m}\Delta_i(k/m)\,P_{k,m}(x).$$

For given m, the random variables $Y_{1,m}, \ldots, Y_{n,m}$ are also i.i.d. with mean zero, so that

$$\mathbb{V}\mathrm{ar}\big[\hat{F}_{m,n}(x)\big] = \frac{1}{n}\mathbb{E}[Y_{1,m}^2]. \tag{14}$$

However, it is easy to verify that

$$\mathbb{E}\big[\Delta_1(x)\Delta_1(y)\big] = \min\big(F(x), F(y)\big) - F(x)F(y),$$

for any $x, y \in [0, 1]$, so that

$$\begin{aligned}\mathbb{E}[Y_{1,m}^2] &= \sum_{k=0}^{m}\sum_{l=0}^{m}\mathbb{E}\big[\Delta_1(k/m)\Delta_1(l/m)\big]P_{k,m}(x)P_{l,m}(x)\\ &= \sum_{k=0}^{m}F(k/m)P_{k,m}^2(x) + 2\sum_{0\leq k<l\leq m}\sum F(k/m)P_{k,m}(x)P_{l,m}(x) - B_m^2(x).\end{aligned} \tag{15}$$

It is now a matter of obtaining an asymptotic expression for (15). For this, we use arguments similar to those used by Babu et al. (2002, Lemma 3.2) and Leblanc (2010, Proposition 1). For this, we first expand $F(k/m)$ about x to write

$$F(k/m) = F(x) + O\big(|k/m - x|\big),$$

which holds for all $0 \leq k \leq m$. This allows us to write the first term of (15) as

$$\sum_{k=0}^{m} F(k/m) P_{k,m}^2(x) = F(x) S_m(x) + O\big(I_m(x)\big), \tag{16}$$

where

$$I_m(x) = \sum_{k=0}^{m} |k/m - x| P_{k,m}^2(x).$$

For the second term of (15), we instead write $F(k/m)$ as

$$F(k/m) = F(x) + (k/m - x) f(x) + O\big((k/m - x)^2\big), \tag{17}$$

and note that

$$1 = \sum_{k=0}^{m} \sum_{l=0}^{m} P_{k,m}(x) P_{l,m}(x) = 2R_{0,m}(x) + S_m(x),$$

so that

$$R_{0,m}(x) = \frac{1}{2}\big[1 - S_m(x)\big].$$

This last result, along with (17) and Lemma 2 (v), leads to

$$\begin{aligned} \sum_{0 \leq k < l \leq m} \sum F(k/m) P_{k,m}(x) P_{l,m}(x) &= F(x) R_{0,m}(x) + f(x) R_{1,m}(x) + O\big(R_{2,m}(x)\big) \\ &= \frac{1}{2} F(x)\big[1 - S_m(x)\big] + f(x) R_{1,m}(x) + O(m^{-1}), \end{aligned} \tag{18}$$

so that, substituting (16) and (18) back into (15) and using Lemma 1, we get

$$\mathbb{E}[Y_{1,m}^2] = \sigma^2(x) + 2 f(x) R_{1,m}(x) + O(I_m(x)) + O(m^{-1}). \tag{19}$$

Now, using the Cauchy–Schwarz inequality and the fact that $0 \leq P_{k,m}(x) \leq 1$, we have

$$\begin{aligned} I_m(x) &\leq \left[\sum_{k=0}^{m}\left(\frac{k}{m}-x\right)^2 P_{k,m}(x)\right]^{1/2}\left[\sum_{k=0}^{m} P_{k,m}^3(x)\right]^{1/2} \\ &\leq \left[\frac{T_{2,m}(x)}{m^2}S_m(x)\right]^{1/2} \\ &\leq \left[\frac{S_m(x)}{4m}\right]^{1/2}, \end{aligned} \tag{20}$$

so that, after using Lemma 2 (ii), we get $I_m(x) = O_x(m^{-3/4})$. This fact and Lemma 2 (iv) allow us to rewrite (19) as

$$\mathbb{E}[Y_{1,m}^2] = \sigma^2(x) - m^{-1/2}V(x) + o_x(m^{-1/2}). \tag{21}$$

In light of (14), this leads to the required asymptotic expression for the variance of $\hat{F}_{m,n}(x)$. □

Proof of Theorem 2 We essentially follow the approach taken by Babu et al. (2002, Proposition 1). It has been established earlier that, for fixed m, $\hat{F}_{m,n}(x)$ is an average of i.i.d. random variables. Let $s_m^2 = \mathbb{E}[Y_{1,m}^2]$; then, making use of the central limit theorem for double arrays (cf. Serfling 1980, Section 1.9.3), the required result will hold if and only if the following Lindeberg condition is satisfied,

$$s_m^{-2}\,\mathbb{E}\Big[Y_{1,m}^2\,\mathbb{I}\big(|Y_{1,m}| > \varepsilon s_m n^{1/2}\big)\Big] \longrightarrow 0, \tag{22}$$

for every $\varepsilon > 0$ as $n \to \infty$. However, note that in light of (21), we have that $s_m \to \sigma(x)$ as $m \to \infty$, and that

$$|Y_{1,m}| \leq 2\sum_{k=0}^{m} P_{k,m}(x) = 2,$$

for all m. Obviously, then, (22) holds when $m, n \to \infty$. □

Proof of Theorem 3 The proof of this result follows along the lines of the proof of Theorem 3 of Leblanc (2010). We first note that (20), Lemma 3 (i) and Jensen's inequality together lead to

$$\begin{aligned} \int_0^1 I_m(x)\,\mathrm{d}x &\leq \left[\frac{1}{4m}\int_0^1 S_m(x)\,\mathrm{d}x\right]^{1/2} \\ &= \left[\frac{1}{4m^{3/2}}\big(\sqrt{\pi}/2 + O(m^{-1})\big)\right]^{1/2} = O(m^{-3/4}), \end{aligned}$$

the function $G(x) = \sqrt{x}$ being concave on [0, 1]. Combining this with (13), (14) and (19), we can write

$$\begin{aligned}\mathrm{MISE}\big[\hat{F}_{m,n}\big] &= \int_0^1 \Big(\mathbb{V}\mathrm{ar}\big[\hat{F}_{m,n}(x)\big] + \mathbb{B}\mathrm{ias}\big[\hat{F}_{m,n}(x)\big]^2\Big)\,\mathrm{d}x \\ &= n^{-1}\bigg[\int_0^1 \sigma^2(x)\,\mathrm{d}x + 2\int_0^1 f(x)R_{1,m}(x)\,\mathrm{d}x\bigg] + m^{-2}\int_0^1 b^2(x)\,\mathrm{d}x \\ &\quad + O(m^{-3/4}n^{-1}) + o(m^{-2}),\end{aligned}$$

since the $O(m^{-1})$ term of (19) is independent of x. It now suffices to use Lemma 3 (ii) and to notice that $2f(x)\psi_2(x) = V(x)$ to get

$$\mathrm{MISE}\big[\hat{F}_{m,n}\big] = n^{-1}C_1 - m^{-1/2}n^{-1}C_2 + m^{-2}C_3 + o(m^{-1/2}n^{-1}) + o(m^{-2}),$$

as was claimed. □

Proof of Theorem 4 We cover here only the case of local deficiency since, for all practical purposes, the case of global deficiency uses identical arguments. For conciseness, we use $i(n)$ in lieu of $i_L(n, x)$ in what follows.

Turning to the proof of the first result, we start by noting that, by definition, $i(n)$ satisfies

$$\mathrm{MSE}\big[F_{i(n)}(x)\big] \le \mathrm{MSE}\big[\hat{F}_{m,n}(x)\big] \le \mathrm{MSE}\big[F_{i(n)-1}(x)\big],$$

that is,

$$\begin{aligned}[i(n)]^{-1}\sigma^2(x) &\le n^{-1}\sigma^2(x) - m^{-1/2}n^{-1}V(x) + m^{-2}b^2(x) \\ &\quad + o_x(m^{-1/2}n^{-1}) + o(m^{-2}) \\ &\le [i(n)-1]^{-1}\sigma^2(x),\end{aligned} \tag{23}$$

and $\lim_{n\to\infty} i(n) = \infty$. From this, we can see that

$$1 \le \frac{i(n)}{n}\Big[1 - m^{-1/2}\theta(x) + m^{-2}n\,\gamma(x) + o_x(m^{-1/2}) + o_x(m^{-2}n)\Big] \le \frac{i(n)}{i(n)-1},$$

where $\theta(x) = V(x)/\sigma^2(x)$ and $\gamma(x) = b^2(x)/\sigma^2(x)$. Now, as long as $mn^{-1/2} \to \infty$ (so that $m^{-2}n \to 0$), taking the limit as $n \to \infty$ in the previous inequality leads to $i(n)/n \to 1$, so that the condition for first-order efficiency indeed holds. To see that (i) also holds, it suffices instead to rewrite (23) as

$$\begin{aligned}m^{-1/2}n^{-1}\theta(x) &\le A_{1,n} + m^{-2}\gamma(x) + o_x(m^{-1/2}n^{-1}) + o_x(m^{-2}) \\ &\le m^{-1/2}n^{-1}\theta(x) + A_{2,n},\end{aligned}$$

where

$$A_{1,n} = \frac{1}{n} - \frac{1}{i(n)}, \qquad \text{and} \qquad A_{2,n} = \frac{1}{i(n)-1} - \frac{1}{i(n)}.$$

This, in turn, can be rewritten as

$$\theta(x) \le m^{1/2} n A_{1,n} + m^{-3/2} n\, \gamma(x) + o_x(1) + o_x(m^{-3/2} n) \le \theta(x) + m^{1/2} n A_{2,n}. \tag{24}$$

Now, assuming that $mn^{-2/3} \to \infty$ and $mn^{-2} \to 0$ (so that $m^{-3/2}n \to 0$ and $m^{-1/2}n \to \infty$), we have that

$$\lim_{n\to\infty} m^{1/2} n A_{1,n} = \left(\lim_{n\to\infty} \frac{i(n)-n}{m^{-1/2}n}\right)\left(\lim_{n\to\infty} \frac{n}{i(n)}\right) = \lim_{n\to\infty} \frac{i(n)-n}{m^{-1/2}n},$$

and that

$$\lim_{n\to\infty} m^{1/2} n A_{2,n} = \left(\lim_{n\to\infty} m^{1/2} n^{-1}\right)\left(\lim_{n\to\infty} \frac{n}{i(n)}\right)\left(\lim_{n\to\infty} \frac{n}{i(n)-1}\right) = 0.$$

Hence, taking the limit in (24), it is clear that (i) does hold. Finally, for (ii), note that if $mn^{-2/3} \to c > 0$, a similar argument instead leads to

$$\lim_{n\to\infty} \frac{i(n)-n}{m^{-1/2}n} = \theta(x) - c^{-3/2}\, \gamma(x),$$

and, since

$$\begin{aligned} \lim_{n\to\infty} \frac{i(n)-n}{m^{-1/2}n} &= \left(\lim_{n\to\infty} \frac{i(n)-n}{n^{2/3}}\right)\left(\lim_{n\to\infty} m^{1/2} n^{-1/3}\right) \\ &= c^{1/2} \lim_{n\to\infty} \frac{i(n)-n}{n^{2/3}}, \end{aligned}$$

the required result easily follows. □

Acknowledgments This research was supported by the National Sciences and Engineering Research Council of Canada. In addition, the author wishes to thank Brad C. Johnson for his comments and suggestions and Dave Gabrielson for his help with the simulation study.

References

Aggarwal, O. P. (1995). Some minimax invariant procedures for estimating a cumulative distribution function. *Annals of Mathematical Statistics, 26*, 450–463.

Altman, N., Léger, C. (1995). Bandwidth selection for kernel distribution function estimation. *Journal of Statistical Planning and Inference, 46*, 195–214.

Azzalini, A. (1981). A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika, 68*, 326–328.

Babu, G. J., Chaubey, Y. P. (2006). Smooth estimation of a distribution function and density function on a hypercube using Bernstein polynomials for dependent random vectors. *Statistics and Probability Letters, 76*, 959–969.

Babu, G. J., Canty, A. J., Chaubey, Y. P. (2002). Application of Bernstein polynomials for smooth estimation of a distribution and density function. *Journal of Statistical Planning and Inference, 105*, 377–392.

Billingsley, P. (1995). *Probability and measure* (3rd ed.). New York: Wiley.

Bowman, A., Hall, P., Prvan, T. (1998). Bandwidth selection for the smoothing of distribution functions. *Biometrika, 85*, 799–808.

Brown, B. M., Chen, S. X. (1999). Beta-Bernstein smoothing for regression curves with compact support. *Scandinavian Journal of Statistics, 26*, 47–59.

Chacón, J. E., Rodríguez-Casal, A. (2010). A note on the universal consistency of the kernel distribution function estimator. *Statistics and Probability Letters, 80*, 1414–1419.

Chang, I.-S., Hsiung, C. A., Wu, Y.-J., Yang, C.-C. (2005). Bayesian survival analysis using Bernstein polynomials. *Scandinavian Journal of Statistics, 32*, 447–466.

Choudhuri, N., Ghosal, S., Roy, A. (2004). Bayesian estimation of the spectral density of a time series. *Journal of the American Statistical Association, 99*, 1050–1059.

Cressie, N. (1978). A finely tuned continuity correction. *Annals of the Institute of Statistical Mathematics, 30*, 435–442.

Falk, M. (1983). Relative efficiency and deficiency of kernel type estimators of smooth distribution functions. *Statistica Neerlandica, 37*, 73–83.

Ghosal, S. (2001). Convergence rates for density estimation with Bernstein polynomials. *Annals of Statistics, 29*, 1264–1280.

Hjort, L. H., Walker, S. G. (2001). A note on kernel density estimators with optimal bandwidths. *Satistics and Probability Letters, 54*, 153–159.

Hodges, J. L. Jr., Lehman, E. L. (1970). Deficiency. *Annals of Mathematical Statistics, 41*, 783–801.

Jones, M. C. (1990). The performance of kernel density functions in kernel distribution function estimation. *Satistics and Probability Letters, 9*, 129–132.

Kakizawa, Y. (2004). Bernstein polynomial probability density estimation. *Journal of Nonparametric Statistics, 16*, 709–729.

Leblanc, A. (2009). Chung–Smirnov property for Bernstein estimators of distribution functions. *Journal of Nonparametric Statistics, 21*, 133–142.

Leblanc, A. (2010). A bias-reduced approach to density estimation using Bernstein polynomials. *Journal of Nonparametric Statistics, 22*, 459–475.

Liu, R., Yang, L. (2008). Kernel estimation of multivariate cumulative distribution function. *Journal of Nonparametric Statistics, 20*, 661–677.

Lorentz, G. G. (1986). *Bernstein polynomials* (2nd ed.). New York: Chelsea Publishing.

Petrone, S. (1999). Bayesian density estimation using Bernstein polynomials. *Canadian Journal of Statistics, 27*, 105–126.

Petrone, S., Wasserman, L. (2002). Consistency of Bernstein polynomial posteriors. *Journal of the Royal Statistical Society, Series B, 64*, 79–100.

Rao, B. L. S. P. (2005). Estimation of distribution and density functions by generalized Bernstein polynomials. *Indian Journal of Pure and Applied Mathematics, 36*, 63–88.

Read, R. R. (1972). Asymptotic inadmissibility of the sample distribution function. *Annals of Mathematical Statistics, 43*, 89–95.

Reiss, R.-D. (1981). Nonparametric estimation of smooth distribution functions. *Scandinavian Journal of Statistics, 8*, 116–119.

Serfling R. J. (1980). *Approximation theorems of mathematical statistics.* New York: Wiley.

Silverman, B. W. (1986). *Density estimation*. Boca Raton: Chapman & Hall/CRC.

Swanepoel, J. W. H., Van Graan, F. C. (2005). A new kernel distribution function estimator based on a non-parametric transformation of the data. *Scandinavian Journal of Statistics, 32*, 551–562.

Tenbusch, A. (1994). Two-dimensional Bernstein polynomial density estimators. *Metrika, 41*, 233–253.

Tenbusch, A. (1997). Nonparametric curve estimation with Bernstein estimates. *Metrika, 45*, 1–30.

Vitale, R. A. (1975). A Bernstein polynomial approach to density function estimation. *Statistical Inference and Related Topics, 2*, 87–99.

Watson, G. S., Leadbetter, M. R. (1964). Hazard analysis II. *Sankhya A, 26*, 101–116.