 Published by Blackwell Publishers Ltd, 108 Cowley Road, Oxford OX4 1JF, UK and 350 Main Street, Malden, MA 02148, USA Vol 26: 47–59, 1999

Beta–Bernstein Smoothing for Regression Curves with Compact Support

BRUCE M. BROWN

University of Tasmania

SONG XI CHEN

La Trobe University

ABSTRACT. The problem of boundary bias is associated with kernel estimation for regression curves with compact support. This paper proposes a simple and unified approach for remedying boundary bias in non-parametric regression, without dividing the compact support into interior and boundary areas and without applying explicitly different smoothing treatments separately. The approach uses the beta family of density functions as kernels. The shapes of the kernels vary according to the position where the curve estimate is made. They are symmetric at the middle of the support interval, and become more and more asymmetric nearer the boundary points. The kernels never put any weight outside the data support interval, and thus avoid boundary bias. The method is a generalization of classical Bernstein polynomials, one of the earliest methods of statistical smoothing. The proposed estimator has optimal mean integrated squared error at an order of magnitude $n^{-4/5}$, equivalent to that of standard kernel estimators when the curve has an unbounded support.

Key words: bandwidth, Bernstein polynomials, beta kernels, boundary bias, hypergeometric distribution, mean integrated square error, non-parametric regression

1. Introduction

This paper proposes a variable kernel method of non-parametric curve estimation, for functions having compact support, which does not suffer from the problem of boundary bias. The method is closely connected to classical Bernstein polynomials.

There are two components in kernel smoothing methods. One is a kernel function which determines the shape of a local weight function; another is a smoothing bandwidth which controls the amount of smoothing used. It is well known that the performance of the kernel method depends largely on the smoothing bandwidth, and depends very little on the form of the kernel. Almost all kernels used are symmetric. Once chosen, the same kernel is used throughout the entire data domain. This is certainly an efficient way for smoothing data with unbounded support. However, it is not so for estimating curves which have compact support. For curves of this type, a fixed form of kernel leads to boundary bias. For instance, if a symmetric kernel is used, the estimate for the curve at the boundary points has an expected value only half the true curve value as half of the kernel weight is outside the data support.

Several authors have suggested ways for removing boundary bias in non-parametric regression. Among them, Gasser & Müller (1979) and Müller (1991) proposed using boundary kernels; Rice (1984) used Richardson extrapolation to combine two kernel estimates with different bandwidths; Hall & Wehrly (1991) suggested a hybrid method. All the existing methods have one thing in common, that is a different smoothing treatment from the one used in the interior has to be used in the boundary region. Gasser & Müller (1979) used different kernels while Rice (1984) used different bandwidths in the boundary areas.

The present paper proposes a unified smoothing approach for remedying boundary bias in

non-parametric density estimation, without dividing the compact support into interior and boundary areas. The approach uses the beta family of density functions as kernels. The shape of the kernels varies according to the position where the curve estimate is made. It is symmetric at the middle of the support interval, and becomes more and more asymmetric nearer the boundary points. The beta kernels put no weight outside the data support, and avoid boundary bias. In recent years, it has been shown by Fan & Gijbels (1992) and Fan (1993) that the local polynomial smoother is free of boundary bias and achieves the optimal rate of convergence for the mean integrated squared error. It is interesting to note that even a local polynomial smoother uses a fixed kernel in its initial form, the local least-squares regression it conducts leads to the use of different kernels at different places implicitly, and that seems to be the reason why the local polynomial smoother is free of boundary bias and has the optimal rate of convergence.

Because observations are made at discrete points in the support interval, the beta kernels have to be split in some way to apportion different weights to different observations. It is proposed to do this with binomial probabilities, and the resulting scheme has weights related to hypergeometric probabilities. The corresponding weights for Bernstein polynomials are binomial probabilities, so the present scheme is an extension of Bernstein polynomial smoothing.

Section 2 introduces the curve estimation model and the form of weights in the extended Bernstein smoothing scheme. All results are stated in section 3. The main result is that if the underlying curve has bounded continuous second derivatives, then the optimal mean integrated squared error has convergence rate $O(n^{-4/5})$, equivalent to the optimal rate in standard kernel estimation when the curve has an unbounded support. Section 4 presents results of a simulation study, and all the proofs are contained in section 5.

2. A beta kernel estimator

Suppose we have $n+1$ observations $y_0, \ldots, y_n$ which are responses at $n+1$ fixed equispaced design points $\{j/n;\ j = 0, 1, 2, \ldots, n\}$ in [0, 1] according to the regression model

$$y_j = m(j/n) + \epsilon_j, \quad j = 0, 1, 2, \ldots, n, \tag{1}$$

where m is an unknown smooth function on [0, 1], and the errors $\{\epsilon_j\}$ are independent, with zero mean and constant variance σ^2.

An early estimator of m, using variable kernel shapes, is one based on the Bernstein polynomials. The nth Bernstein polynomial of a continuous function g is

$$B_n g(x) = \sum_{j=0}^{n} g(j/n) \binom{n}{j} x^j (1-x)^{n-j}. \tag{2}$$

A corresponding estimator of m, now called the *Bernstein smoother*, is $\hat{m}_b$ where

$$\hat{m}_b(x) = \sum_{j=0}^{n} y_j \binom{n}{j} x^j (1-x)^{n-\mathrm{j}}. \tag{3}$$

Note that $E\{\hat{m}_b(x)\} = B_n m(x)$. The Bernstein polynomial $B_n m$ converges to m with a rate of convergence of n^{-1}, uniformly in [0, 1] if m has bounded second derivative. Therefore, the Bernstein smoother is free of boundary bias. Despite this good property, the estimator has been given little attention by statisticians. In fact it was mentioned in Priestley & Chao (1972), one of the pioneer papers in non-parametric regression. However, the authors did not develop it there, but went on to propose the well-known Priestly–Chao estimator with a fixed symmetric kernel. This is understandable as boundary bias was not then a concern. Another reason for a lack of interest in the Bernstein smoother is that it undersmooths, with a very small smoothing

bandwidth of $O(n^{-1/2})$. It may be shown, for example by lemma 4 of this paper, that $\mathrm{var}\{\hat{m}_b(x)\} = O(n^{-1/2})$ which means that the mean integrated squared error is $O(n^{-1/2})$.

However, there are interesting features of the Bernstein smoother which make it worth exploring further. One is that it uses both variable kernel shape and variable amount of smoothing according to the position where the smoothing is made. The other is that it is free of boundary bias. There have been interesting proposals in Stadtmüller (1986) and Tenbusch (1997) to modify Bernstein polynomials for curve estimation. In the present paper, an approach linking Bernstein polynomials with a family of beta probability density functions is considered.

Let $K_{\alpha,\beta}$ denote the density function of a Beta(α, β) random variable. Our aim is to use the density functions K as naturally varying kernels to create an estimate $\hat{m}$ of the function m, putting no weight outside the data support interval [0, 1]. Specifically, we propose

$$\hat{m}(x) = \sum_{j=0}^{n} y_j w_j(x) \tag{4}$$

with weights $\{w_j(x)\}$ formed by splitting the beta p.d.f. $K_{\lambda x+1,\lambda(1-x)+1}$ in some way which allocates "pieces" to observations $\{y_j;\ j = 0, 1, 2, \ldots, n\}$. That is, set

$$w_j(x) = \int_0^1 K_{\lambda x+1,\lambda(1-x)+1}(t) f_j(t)\,dt, \tag{5}$$

where $\sum_{j=0}^{n} f_j(t) = 1$ for all $t \in [0, 1]$.

Making $\sum_j f_j(t) = 1$ ensures that $\sum_0^n w_j(x) = 1$, all x, and the estimation of m by $\hat{m}$ is location-shift invariant. The constant λ is a smoothing parameter, and converges to infinity as n tends to infinity, but with $\lambda/n \to 0$. The shapes of the beta kernels vary according to the position where the curve estimate is made. They are symmetric at the middle of the support interval, and become more and more asymmetric nearer the boundary points. One important feature is that the beta kernels put no weight outside [0, 1]. But since the variances of beta$\{\lambda x + 1,\ \lambda(1 - x) + 1\}$ distributions are approximately $(\lambda x + 1)\{\lambda(1 - x) + 1\}/\lambda^3$ and depend on x, the estimate $\hat{m}$ will be like a kernel estimate with both varying kernels and varying bandwidths, having smaller bandwidth for x near the boundaries at 0 and 1. Figure 1 plots the weight functions $w_j(x)$ with $f_j(t) = \binom{n}{j} t^j (1 - t)^{n-j}$ and those for the boundary kernel estimator using kernels in Müller (1991) modified from the quartic kernel for selected x values. The bandwidths used are $\lambda = 66.66$ and $h = 0.21$ for the beta and boundary kernels respectively. We see the weight functions of the two estimators are very close to each other except at $x = 0$. At $x = 0$, the boundary kernel weight function has negative values, whereas the Beta-Bernstein weight function remains positive.

We now propose the choice

$$f_j(t) = \binom{n}{j} t^j (1 - t)^{n-j},$$

thus treating j as a Bi(n, t) random variable, and establishing links with classical Bernstein polynomials. Clearly $\sum_0^n f_j(t) = 1$ for all t, as required. For convenience, assume λ to be an integer. Then

$$\begin{aligned} w_j(x) &= \int_0^1 \frac{t^{\lambda x}(1 - t)^{\lambda(1-x)}}{B\{\lambda x + 1,\ \lambda(1 - x) + 1\}} \binom{n}{j} t^j (1 - t)^{n-j}\,dt \\ &= \frac{\Gamma(\lambda x + j + 1)}{\Gamma(\lambda x + 1)} \frac{\Gamma\{\lambda(1 - x) + n - j + 1\}}{\Gamma\{\lambda(1 - x) + 1\}} \frac{(\lambda + 1)!\,n!}{(\lambda + n + 1)!\,j!(n - j)!} \end{aligned} \tag{6}$$

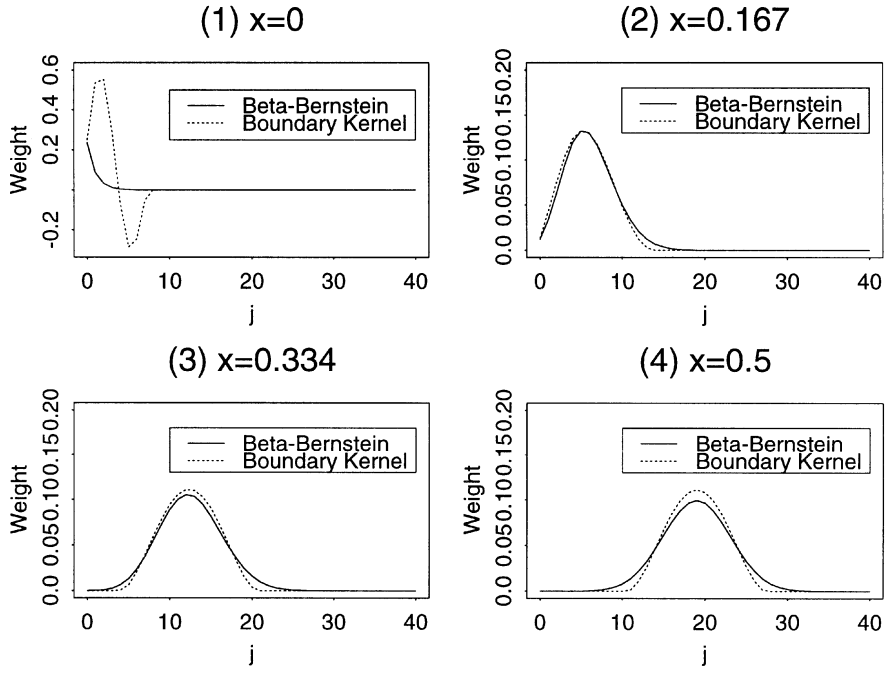

Fig. 1. The beta kernel weight function $w_j(x)$ and its boundary kernel counterpart for $n = 40$.

and this form is easy to deal with numerically as long as a gamma function routine is available.

To summarize, our proposal is to estimate m by $\hat{m}$ in (4), with weights $\{w_j\}$ given by (6).

3. Main results

To assess the performance of $\hat{m}$, we consider the integrated mean squared error

$$\begin{aligned}\mathrm{MISE}(\hat{m}) &= \int_0^1 E\{\hat{m}(x) - m(x)\}^2\, dx \\ &= \int_0^1 \{[\mathrm{bias}\{\hat{m}(x)\}]^2 + \mathrm{var}\{\hat{m}(x)\}\}\, dx. \end{aligned} \tag{7}$$

Our main result, stated below as theorem 4, shows that under a smoothness condition on m, taking $\lambda = O(n^{2/5})$ yields a minimum MISE of $O(n^{-4/5})$. This rate coincides with the optimal rate in kernel estimation for data with unbounded support, and for a bounded support interval, is achieved by the standard kernel methods only in the interior of the data support interval. In addition, the present estimate does not suffer from boundary bias. There is a price paid for such good bias property of the Beta-Bernstein estimator. The order of magnitude of the variance of this estimator increases in and near the boundary points, while the variance of the boundary kernel estimator is at the same order throughout [0, 1]. However, as the size of the area where the Beta-Bernstein estimator has an increased variance is one order of magnitude smaller than the boundary bias area, it has no effect on the overall convergence rate of the MISE.

The results in theorem 4 is built up from results in theorem 1 for bias and theorems 2 and 3 for variance. In what follows, K always denotes a generic positive and finite constant.

Theorem 1
Let $m'' \in C[0, 1]$. *Then* $\text{bias}\{\hat{m}(x)\}$ *is* $O(\lambda^{-1})$, *uniformly in* $0 \leqslant x \leqslant 1$, *and in particular,*

$$\text{bias}\{\hat{m}(x)\} = \{(1-2x)m'(x) + \tfrac{1}{2}x(1-x)m''(x)\}\lambda^{-1} + o(\lambda^{-1}) + O(n^{-1})$$

where $o(\lambda^{-1})$ *and* $O(n^{-1})$ *are uniformly so in* $0 \leqslant x \leqslant 1$.

The bound on $\text{bias}\{\hat{m}(x)\}$ established in theorem 1 enables the bias term on the right hand side of (7) to be dealt with. The other term is integrated variance,

$$\text{IV} = \int_0^1 \text{var}\{\hat{m}(x)\}\, dx = \sigma^2 \int_0^1 \sum_{j=0}^{n} w_j^2(x)\, dx,$$

from (1) and (4). Theorem 2 given below establishes a bound on IV.

Theorem 2
For all n, λ *large enough and* $\lambda = o(n^{1/2})$ *as n tends to infinity*

$$\text{IV} = O(n^{-1}\lambda^{1/2}) \quad \text{as} \quad \lambda,\, n \to \infty.$$

While theorem 2 finds the bound on IV, the following theorem 3 estimates the rate constant.

Theorem 3
For all, n, λ *large enough and* $\lambda = o(n^{1/2})$ *as n tends to infinity,*

$$\text{IV} = \sigma^2 \frac{\sqrt{\pi}}{2} \frac{\sqrt{\lambda}}{(\lambda + n)} + o(n^{-1}\lambda^{1/2}).$$

Putting theorems 2 and 3 together yields

Theorem 4
Let $m'' \in C[0, 1]$. *Then as* λ *and n tend to infinity with* $\lambda = o(n)$, $\text{MISE}(\hat{m}) = O(\lambda^{-2} + n^{-1}\lambda^{1/2})$. *The optimal rate* $\lambda = O(n^{2/5})$ *yields* $\text{MISE}(\hat{m}) = O(n^{-4/5})$, *and in particular*

$$\text{MISE}(\hat{m}) = \lambda^{-2} \int_0^1 \{(1-2x)m'(x) + \tfrac{1}{2}x(1-x)m''(x)\}^2\, dx + \frac{\sqrt{\pi}}{2}\sigma^2 n^{-1}\lambda^{1/2} + o(\lambda^{-2} + n^{-1}\lambda^{1/2}). \tag{8}$$

The optimal bandwidth which minimizes the leading order terms in (8) is

$$\lambda^* = (\tfrac{1}{8}\sqrt{\pi}\sigma^2)^{-2/5} \left[\int_0^1 \{(1-2x)m'(x) + \tfrac{1}{2}x(1-x)m''(x)\}^2\, dx\right]^{2/5} n^{2/5}. \tag{9}$$

Substituting the optimal bandwidth in (9) into (8), we have the optimal mean integrated squared error

$$\text{MISE}^* = 5(\tfrac{1}{8}\sqrt{\pi}\sigma^2)^{4/5}\left[\int_0^1 \{(1-2x)m'(x) + \tfrac{1}{2}x(1-x)m''(x)\}^2\,dx\right]^{1/5} n^{-4/5}. \tag{10}$$

The optimal mean integrated squared error for the boundary kernel estimator as given in Müller (1988) is

$$5\{R(K)\sigma_K^2\sigma^2\}^{4/5}\left[\int_0^1 \{m''(x)\}^2\,dx\right]^{1/5} n^{-4/5}. \tag{11}$$

where $\sigma_k^2 = \int u^2 K(u)\,du$ and $R(K) = \int K^2(u)\,du$.

Both estimators achieve the optimal rate of convergence for the mean integrated squared error. The appearance of m' in (10) is due to the fact that x is not the mean of the beta($\lambda x + 1, \lambda(1-x) + 1$) distribution, rather it is the mode. However, m' can be removed from the mean integrated squared error if $K_{\lambda x, \lambda(1-x)}$ are used as kernels for x in the interior of [0, 1] and some modified beta kernels are used in the boundary areas. However, we do not consider this modification in this paper.

As $\lambda = O(n^{2/5})$ and converges to ∞ as n is large, it is convenient to use $b = \lambda^{-1}$ as the smoothing parameter instead in numerical calculation. However, in practice we cannot use (9) to choose λ or b because of the unknown derivatives of m, but the penalizing function approach can be used to choose the bandwidth. We only give a brief description of the penalizing function procedure; details are available in Härdle (1990). The penalizing function score for using a particular bandwidth level b is

$$PF(b) = n^{-1}\sum_{i=0}^{n}\{Y_i - \hat{m}(x_i)\}^2\eta\{W_j(x_i)\}$$

where $\eta(u) = 1 + 2u$ is a penalizing function proposed by Shibata (1981). The score function $PF(b)$ is computed over a grid of b values. The b value which minimizes $PF(b)$ is used as the bandwidth.

All the proofs of results in this section are deferred to section 5.

4. Empirical results

In this section we present some empirical results designed to investigate the performance of the proposed beta kernel estimator for a regression function. We consider the following quadratic regression model:

$$y_i = (x_i - 0.5)^2 + \epsilon_i \quad i = 1, \ldots, n, \tag{12}$$

where the fixed design points x_i are taken at equally spaced points in [0, 1] and ϵ_i are uncorrelated normal random variables with zero mean and standard deviation $\sigma = 0.05$ and 0.1. The normal random variables were generated using the routines given in Press *et al.* (1992). The kernel estimator of Gasser & Müller (1979) type but with no boundary correction and Müller (1991)'s boundary kernel estimators are also considered for comparison. A quartic kernel

$$K(u) = \tfrac{15}{16}(1-u^2)^2 I(|u| < 1)$$

and boundary kernels modified from it, as given in Table 1 of Müller (1991), are used.

We start with a simulated data set ($n = 40$) generated under model (12). Figure 2 displays a scatter plot of the data and three fitted regression curves using the three estimators. Four levels of the Beta-Bernstein smoothing bandwidths λ and the kernel bandwidth h were used re-

Table 1. *Simulated optimal average integrated squared errors and their standard errors for the beta and the boundary kernel estimates with regression curve* $y = (x - 0.5)^2$ *with* $\sigma = 0.05$ *and* 0.1 *respectively. The columns headed "Predic.", "PF" and "Direct" give* 10^3 *times the optimal mean or average integrated squared errors based on the theoretical expansion, penalizing function and the direct mean square error calculation in finding the smoothing bandwidths. Their standard errors multiplied by* 10^3 *are given inside the parentheses*

		(1) $\sigma = 0.05$			
		Beta-Bernstein		Boundary kernels	
n	Predic.	Direct	PF	Direct	PF
20	0.789	0.490 (0.346)	0.574 (0.367)	0.598 (0.386)	0.978 (0.668)
40	0.453	0.299 (0.185)	0.367 (0.189)	0.352 (0.205)	0.571 (0.507)
60	0.328	0.213 (0.125)	0.279 (0.131)	0.240 (0.143)	0.392 (0.428)
80	0.260	0.179 (0.091)	0.235 (0.100)	0.191 (0.108)	0.291 (0.338)
100	0.218	0.159 (0.079)	0.208 (0.090)	0.164 (0.089)	0.237 (0.285)
120	0.188	0.143 (0.071)	0.185 (0.078)	0.139 (0.079)	0.188 (0.193)
140	0.166	0.127 (0.061)	0.170 (0.069)	0.123 (0.067)	0.170 (0.183)
160	0.150	0.116 (0.055)	0.152 (0.064)	0.109 (0.061)	0.139 (0.131)
180	0.136	0.106 (0.048)	0.143 (0.059)	0.097 (0.051)	0.124 (0.115)
200	0.125	0.100 (0.046)	0.134 (0.056)	0.091 (0.048)	0.119 (0.129)
		(2) $\sigma = 0.1$			
		Beta-Bernstein		Boundary kernels	
n	Predic.	Direct	PF	Direct	PF
20	2.393	1.554 (1.15)	2.072 (1.25)	2.111 (1.42)	3.357 (2.59)
40	1.374	0.929 (0.60)	1.389 (0.73)	1.144 (0.75)	1.661 (1.53)
60	0.994	0.696 (0.45)	1.058 (0.53)	0.778 (0.514)	1.183 (0.95)
80	0.789	0.583 (0.349)	0.901 (0.43)	0.640 (0.422)	0.948 (0.76)
100	0.660	0.494 (0.286)	0.767 (0.365)	0.526 (0.334)	0.715 (0.74)
120	0.571	0.445 (0.252)	0.676 (0.328)	0.448 (0.266)	0.580 (0.472)
140	0.504	0.411 (0.225)	0.620 (0.298)	0.401 (0.239)	0.538 (0.579)
160	0.453	0.361 (0.189)	0.542 (0.269)	0.354 (0.219)	0.467 (0.546)
180	0.431	0.329 (0.177)	0.502 (0.272)	0.319 (0.198)	0.401 (0.241)
200	0.379	0.312 (0.275)	0.455 (0.236)	0.287 (0.174)	0.360 (0.218)

spectively, corresponding to plots (a), (b), (c) and (d). The same bandwidth h was used by both the boundary kernel and the standard kernel estimates in each of the plots. The bandwidth of the beta-Bernstein estimator λ was chosen as h^{-2} to bring the amount of smoothing on the same scale. The h-value used in (b) was close to the average optimal h value according to the simulation results reported in Table 2 below. The boundary bias associated with the standard kernel estimator was obvious. Both the Beta-Bernstein and the boundary kernel estimators were quite close to each other and did not have the boundary bias. The plots for the two estimators in (a) and (b) were quite reasonable, whereas those in (c) and (d) showed signs of undersmooth, indicating the bandwidths used were too small.

We then conducted a simulation study under model (12). The aims were (i) to verify the theoretical expansion for the mean integrated squared error and the formula for optimal λ for the Beta-Bernstein kernel estimator given in (10) and (9) respectively; (ii) compare the mean integrated squared errors of the Beta-Bernstein and the boundary kernel estimators. We did not consider the standard kernel estimator as it was clear that the boundary bias would cause the mean integrated squared error to be much larger.

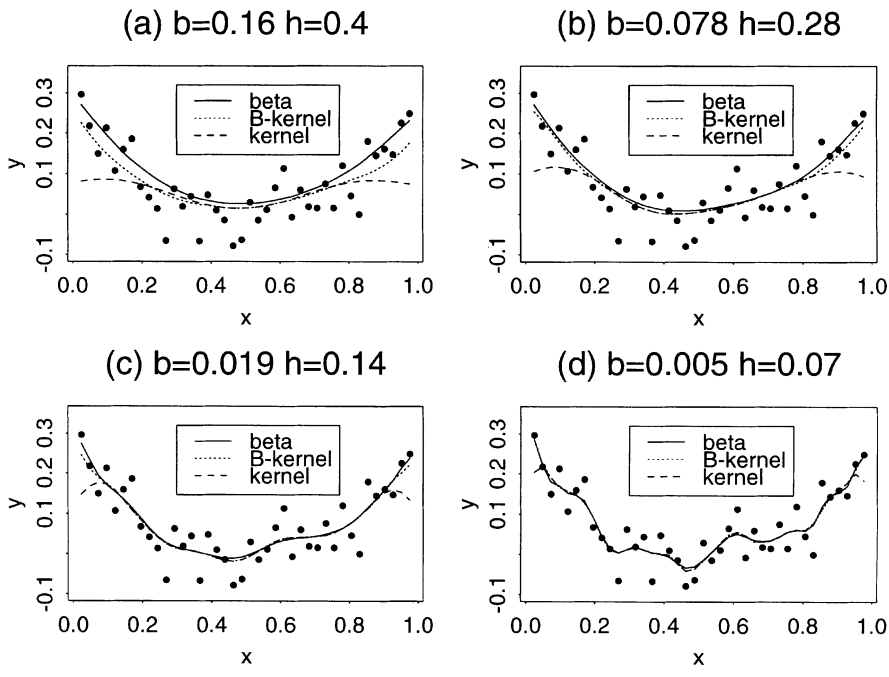

Fig. 2. Estimated regression curves by the beta, the boundary kernel and the standard kernel methods for a simulated data set under the model: $Y_i = (x_i - 0.5)^2 + N(0, 0.05^2)$.

Two methods were used for choosing the smoothing bandwidths. One is the penalizing function approach described in section 4. And the other is a direct method which uses the knowledge that $m(x) = (x - 0.5)^2$. For the Beta-Bernstein estimator, this direct method defined a score function

$$R(b) = \int_0^1 \{\hat{m}(t) - m(t)\}^2 \, dt$$

where $b = \lambda^{-1}$. A similar score function was defined for the boundary kernel estimator. The two score functions were minimized respectively, using the Golden Section Search algorithm given in Press *et al.* (1992). For the boundary kernel estimator, the same bandwidth h is used everywhere in [0, 1] without bandwidth variation, as recommended by Müller (1991).

The average integrated squared errors and their standard errors of the Beta-Bernstein and boundary kernel estimates from 1000 simulations are shown in Table 1 whereas the corresponding average optimal bandwidths used to compute the average integrated squared errors given in Table 1 are shown in Table 2. To confirm the theory developed for the proposed estimator, we also list in Table 1 values of the leading term in the optimal mean integrated squared error expansions in (10), whereas the optimal b-values from (9) are given in Table 2, both are under the heading "Predic.".

The results in Tables 1 and 2 can be summarized as follows. Both the theoretical mean integrated squared errors and optimal b values for the Beta-Bernstein estimates were close to their simulated counterparts when n was large, thus confirming the theoretical expansions developed for the mean integrated squared errors. We find the Beta-Bernstein estimator peformed better for small to medium sample sizes regardless of what method was used to choose the smoothing bandwidth. When the sample size is large, the boundary kernel estimator

Table 2. *Average optimal smoothing bandwidths and their standard errors used in Table 1. The columns headed "Predic.", "PF" and "Direct" give the optimal bandwidth values based on the theoretical expansion, the penalizing function and the direct mean square error calculation respectively. Their standard errors multiplied by 10^2 are given inside parentheses. For the Beta-Bernstein smoother the smoothing bandwidth is $b = \lambda^{-1}$*

			(1) $\sigma = 0.05$		
		Beta-Bernstein		Boundary kernels	
n	Predic.	Direct	PF	Direct	PF
20	0.031	0.018 (0.020)	0.016 (0.027)	0.245 (0.74)	0.189 (1.19)
40	0.023	0.015 (0.011)	0.014 (0.020)	0.211 (0.45)	0.187 (0.68)
60	0.020	0.014 (0.007)	0.013 (0.016)	0.197 (0.39)	0.177 (0.48)
80	0.018	0.011 (0.003)	0.010 (0.010)	0.179 (0.35)	0.172 (0.36)
100	0.016	0.010 (0.002)	0.009 (0.008)	0.172 (0.29)	0.165 (0.32)
120	0.015	0.009 (0.001)	0.008 (0.006)	0.167 (0.28)	0.164 (0.26)
140	0.014	0.009 (0.001)	0.007 (0.005)	0.164 (0.26)	0.157 (0.26)
160	0.013	0.009 (0.001)	0.007 (0.004)	0.157 (0.24)	0.156 (0.22)
180	0.013	0.009 (0.001)	0.007 (0.003)	0.159 (0.22)	0.152 (0.19)
200	0.012	0.008 (0.0008)	0.006 (0.003)	0.154 (0.20)	0.150 (0.18)
			(2) $\sigma = 0.1$		
		Beta-Bernstein		Boundary kernels	
n	Predic.	Direct	PF	Direct	PF
20	0.054	0.041 (0.103)	0.043 (0.029)	0.299 (1.18)	0.278 (2.060)
40	0.041	0.035 (0.047)	0.037 (0.021)	0.270 (0.94)	0.265 (1.090)
60	0.035	0.030 (0.030)	0.031 (0.017)	0.246 (0.76)	0.253 (0.849)
80	0.031	0.027 (0.021)	0.029 (0.015)	0.234 (0.69)	0.238 (0.659)
100	0.028	0.025 (0.017)	0.026 (0.013)	0.227 (0.63)	0.227 (0.615)
120	0.026	0.023 (0.010)	0.025 (0.012)	0.218 (0.57)	0.222 (0.504)
140	0.025	0.021 (0.010)	0.023 (0.012)	0.209 (0.51)	0.217 (0.495)
160	0.023	0.020 (0.007)	0.022 (0.011)	0.203 (0.45)	0.212 (0.443)
180	0.022	0.019 (0.009)	0.021 (0.011)	0.200 (0.44)	0.206 (0.388)
200	0.021	0.017 (0.004)	0.019 (0.010)	0.196 (0.42)	0.203 (0.369)

performed better; however, both were very close. This is not surprising as direct computation of (10) and (11) reveals that the coefficient of the boundary kernel estimator is smaller than that of the Beta-Bernstein estimator for $m(x) = (x - 0.5)^2$. The lead of the boundary kernel estimator at large sample sizes is reduced when σ^2 increases from 0.05 to 0.1. Also there was larger variation in the average integrated squared errors of the boundary kernel estimator for all the cases. For both estimators the optimal bandwidth and the average integrated squared errors from the penalizing function become close to those from the direct average integrated squared errors minimization when the sample size becomes large.

5. Proofs

This section contains the proofs of the theorems given in section 3.

The nth Bernstein polynomial of a continuous function g is defined in (2). For some of the many properties of these polynomials, see for example Davis (1963), Lorentz (1953) or Brown *et al.* (1987). The form of definition clearly involves a binomial distribution, and using that fact yields easy derivations of an elementary lemma which will be needed. Note that we can write

$$B_n g(x) = E\{g(X_x/n)\}, \tag{13}$$

where $X_x \sim Bi(n, x)$.

Lemma 1
Let $|g''(x)| \leqslant K$ *for* $0 \leqslant x \leqslant 1$. *Then*

$$|B_n g(x) - g(x)| \leqslant \tfrac{1}{2} n^{-1} x(1-x) K.$$

Proof. If $U = n^{-1}X_x - x$, then by (13) and a Taylor expansion

$$B_n g(x) - g(x) = E\{Ug'(x) + \tfrac{1}{2}U^2 g''(x_1)\}$$

for some $x_1 \in [0, 1]$. But $|g''(x_1)| \leqslant K$, while $E(U) = 0$, $\mathrm{var}(U) = n^{-1}x(1-x)$, completing the proof.

Lemma 2
Let $m'' \in C[0, 1]$. *Then, as* h *tends to 0*

$$m(x+h) - m(x) = hm'(x) + \tfrac{1}{2}h^2 m''(x) + o(h^2)$$

uniformly in x, $x + h \in [0, 1]$.

Proof. For any x, $x + h \in [0, 1]$, with $h > 0$,

$$m(x+h) - m(x) - hm'(x) - \tfrac{1}{2}h^2 m''(x) = \int_0^h (h-y)\{m''(x+y) - m''(x)\}\, dy$$

$$\leqslant \epsilon_h \int_0^h (h-y)\, dy = \epsilon_h h^2/2,$$

where $\lim_{h\to 0} \epsilon_h = 0$, by uniform continuity of m'' on $[0, 1]$. The case for $h < 0$ is similar.

Proof of theorem 1. Observe from the definition (5) of $w_j(x)$ that

$$E\{\hat{m}(x)\} = \sum_{j=0}^{n} w_j(x) m(j/n) = E\{B_n m(X)\},$$

where $X \sim \mathrm{beta}\{\lambda x + 1, \lambda(1-x) + 1\}$, which has mean $E(X) = (\lambda x + 1)/(\lambda + 2)$ and variance $\omega^2 = \mathrm{var}(X) = (\lambda x + 1)\{\lambda(1-x) + 1\}/\{(\lambda+2)^2(\lambda+3)\}$.

Therefore we can write

$$E\{\hat{m}(x)\} = E\left\{ B_n m\left(\frac{\lambda x + 1}{\lambda + 2} + \omega U\right)\right\}, \tag{14}$$

where $E(U) = 0$, $\mathrm{var}\,(U) = 1$. This expression enables bias $\{\hat{m}(x)\}$ to be expressed as

$$E\{\hat{m}(x)\} - m(x) = A + B + C, \tag{15}$$

where

$$A = E\left\{ B_n m\left(\frac{\lambda x + 1}{\lambda + 2} + \omega U\right) - m\left(\frac{\lambda x + 1}{\lambda + 2} + \omega U\right)\right\},$$

$$B = E\left\{ m\left(\frac{\lambda x + 1}{\lambda + 2} + \omega U\right)\right\} - m\left(\frac{\lambda x + 1}{\lambda + 2}\right),$$

and

$$C = m\left(\frac{\lambda x + 1}{\lambda + 2}\right) - m(x).$$

Now $|A| \leqslant \frac{1}{2}n^{-1}x(1-x)K$ by lemma 1. To bound B and C, apply lemma 2. For B, replace x, h by $(\lambda x + 1)/(\lambda + 2)$ and ωU, and note that $E(U) = 0$. For C, replace x, h by x and $(1 - 2x)/(\lambda + 2)$. The results are

$$B = \tfrac{1}{2}\omega^2 m''\left(\frac{\lambda x + 1}{\lambda + 2}\right) + o(\lambda^{-1}),$$

$$C = \frac{1 - 2x}{\lambda + 2} m'(x) + O(\lambda^{-2}),$$

and already we have $A = O(n^{-1})$, with all O terms being uniformly in $0 \leqslant x \leqslant 1$. Using these expressions for A, B and C in (15) yields the expansion for bias$\{\hat{m}(x)\}$ stated in the theorem. The uniform $O(\lambda^{-1})$ bound for bias $\{\hat{m}(x)\}$ follows since both m'' and m' bounded on [0, 1].

The following lemma regarding the Stirling's formula is well-known.

Lemma 3
Let $S(x) = \sqrt{2\pi}x^{x+1/2}\exp(-x)$, $x > 0$ *and let* $R(x) = S(x)/\Gamma(1 + x)$. *Then* $R(x) < 1$ *for all* $x > 0$, $R(x) \to 1$ *as* x *tends to infinity.*

Proof of theorem 2. Let $w_{j_x^*}(x) = \max_{0 \leqslant j \leqslant n} w_j(x)$. By considering ratios w_j/w_{j+1}, it is easy to show that $j_x^* =$ integer part of $(n + 1)x$.

First restrict x to $\delta < x < 1/2$ where $\delta = \lambda^{-3/2}$, so that $n\delta$ tends to infinity since $\lambda = o(n^{1/2})$, but $\lambda\delta \to 0$. From (6)

$$w_{j_x^*}(x) = \frac{\Gamma(\lambda x + j_x^* + 1)}{\Gamma(\lambda x + 1)} \frac{\Gamma\{\lambda(1-x) + n - j_x^* + 1\}}{\Gamma\{\lambda(1-x) + 1\}} \frac{(\lambda + 1)!n!}{(\lambda + n + 1)!j_x^*!(n - j_x^*)!}.$$

In the above expression for $w_{j_x^*}(x)$, all the arguments of the gamma functions tend to infinity except possibly in the term $\Gamma(\lambda x + 1)$. But from lemma 3,

$$\{\Gamma(\lambda x + 1)\}^{-1} \leqslant \{S(x)\}^{-1},$$

and applying Stirling's formula carefully yields

$$w_{j_x^*}(x) \leqslant (\lambda + 1)(\lambda + n + 1)^{-1} O[\{(\lambda + n)^{1/2}\{n\lambda x(1-x)\}^{-1/2}], \tag{16}$$

uniformly in $\delta < x < 1/2$. Exactly similar reasoning gives the same bound for $1/2 < x < 1 - \delta$.

But $\{w_j(x), j = 0, 1, \ldots, n\}$ is a probability distribution for every x, so

$$\sum_{j=0}^{n} w_j^2(x) \leqslant \max_{0 \leqslant j \leqslant n} w_j(x) = w_{j_x^*}(x),$$

and from (16)

$$\int_\delta^{1-\delta} \sum_{j=0}^{n} w_j^2(x)\, dx \leqslant \int_\delta^{1-\delta} w_{j_x^*}(x)\, dx = O(n^{-1}\lambda^{1/2}) \tag{17}$$

as λ and n tend to infinity.

For the case $x < \delta$ and $x > 1 - \delta$, we need the following upper bound for w_j:

$$w_j(x) \leqslant \frac{\lambda+1}{\lambda+n+1} \tag{18}$$

for all $0 \leqslant x \leqslant 1$ and $0 \leqslant j \leqslant n$. To appreciate (18), we also notice that as j increases from 0 to $n/2$ (for even n) the maximum function value of w_j decreases, and in fact the upper bound is true at w_0.

Thus, noting that $\delta = \lambda^{-3/2}$,

$$\int_0^{\delta} \sum_{j=0}^{n} w_j^2(x)\,dx = \int_{1-\delta}^{1} \sum_{j=0}^{n} w_j^2(x)\,dx = O\left\{\frac{(\lambda+1)^2}{(\lambda+n+1)^2}\, n\lambda^{-3/2}\right\} = O(n^{-1}\lambda^{1/2}) \tag{19}$$

as λ and n tend to infinity. From (17) and (19) it follows that

$$\mathrm{IV} = \sigma^2 \int_0^1 \sum_{j=0}^{n} w_j^2(x)\,dx = O(n^{-1}\lambda^{1/2}) \quad \text{as } \lambda,\ n \to \infty.$$

Lemma 4

If $p_i = \binom{k}{i}\theta^i(1-\theta)^{k-i}$, *then*

$$\sum_{i=0}^{k} p_i^2 \to \tfrac{1}{2}\{k\pi\theta(1-\theta)\}^{-1/2} \quad \textit{as } k \textit{ tends to infinity}$$

Proof. In general if a discrete distribution $\{p_i\}$ has characteristic function ϕ, then

$$\sum_{i=0}^{k} p_i^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |\phi(t)|^2\,dt,$$

and applying to the present case when $\phi(t) = (1-\theta+\theta\exp(it)^k$ gives

$$\sum_{i=0}^{k} p_i^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} [1-2\theta(1-\theta)\{1-\cos(t)\}]^k\,dt \tag{20}$$

Then standard calculations show that as k tends to infinity,

$$k^{-1/2}\sum_{i=0}^{k} p_i^2 \to \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp\{-\theta(1-\theta)y^2\}dy = \frac{1}{2\pi}\left\{\frac{2\pi}{2\theta(1-\theta)}\right\}^{1/2} = \tfrac{1}{2}\{\pi\theta(1-\theta)\}^{-1/2}.$$

Proof of theorem 3. As a rigorous upper bound for IV has been established in theorem 3, we just give an informal proof of theorem 4 in the following as a formal one is very tedious.

First the integral in IV is approximated by the Riemann sum

$$\left(\frac{\mathrm{IV}}{\sigma^2}\right)_1 = \lambda^{-1}\sum_{i=0}^{\lambda}\sum_{j=0}^{n} w_j^2\left(\frac{i}{\lambda}\right) \tag{21}$$

Next, the weights $\{w_j\}$ are linked to certain hypergeometric probabilities, i.e. from (6)

$$w_j\left(\frac{i}{\lambda}\right) = \left(\frac{\lambda+1}{\lambda+n+1}\right)\frac{\binom{\lambda}{i}\binom{n}{a-i}}{\binom{\lambda+n}{a}},$$

where $a = i + j$. Therefore

$$\lambda^{-1}\sum_{i,j} w_j^2\left(\frac{i}{\lambda}\right) = \frac{(\lambda+1)^2}{\lambda(\lambda+n+1)^2}\sum_{i,a} P^2(Y_a = i), \tag{22}$$

where Y_a has the hypergeometric distribution $\mathrm{Hg}(\lambda+n;\lambda,a)$. Now for λ and n tends to infinity with λ/n tends to 0, this distribution is essentially $\mathrm{Bi}\{\lambda, a/(\lambda+n)\}$. But from lemma 4 the sum of squares of $\mathrm{Bi}(k,\theta)$ probabilities $\sim \frac{1}{2}\{\pi k\theta(1-\theta)\}^{-1/2}$ as k tends to infinity for θ fixed, so from (21) and (22) we have

$$\begin{aligned}\left(\frac{\mathrm{IV}}{\sigma^2}\right)_1 &\sim \frac{(\lambda+1)}{(\lambda+n)^2}\sum_a \frac{1}{2}\left\{\pi\lambda\frac{a}{(\lambda+n)}\left(1-\frac{a}{\lambda+n}\right)\right\}^{-1/2}\\ &\sim \frac{(\lambda+1)}{\sqrt{\lambda}(\lambda+n)}\frac{1}{2\sqrt{\pi}}\int_0^1 \frac{dx}{\sqrt{x(1-x)}}\\ &\sim \frac{\sqrt{\pi}}{2}\frac{\sqrt{\lambda}}{(\lambda+n)}, \quad \text{as } n, \lambda \text{ tend to infinity.}\end{aligned} \tag{23}$$

Acknowledgement

The authors thank two referees and the Associate Editor for helpful comments and suggestions.

References

Brown, B. M., Elliott, D. & Paget, D. F. (1987). Lipschitz constraints for the Bernstein polynomials of a Lipschitz continuous function. *J. Approx. Theory* **49**, 196–199.

Davis, P. J. (1963). *Interpolation and approximation*. Blaisdell, Waltham, MA.

Fan, J. Q. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.* **21**, 196–216.

Fan, J. Q. & Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *Ann. Statist.* **20**, 2008–2036.

Gasser, Th. & Müller, H.-G. (1979). Kernel estimation of regression functions. In *Smoothing techniques for curve estimation* (eds Th. Gasser & M. Rosenblatt), 23–68. Springer, Heidelberg.

Hall, P. & Wehrly (1991). A geometrical method for removing edge effects from kernel-type nonparametric regression estimators. *J. Amer. Statist. Assoc.* **86**, 665–672.

Härdle, W. (1990). *Applied nonparametric regression*. Cambridge University Press, Cambridge.

Lorentz, G. G. (1953). *Bernstein polynomials*. Math. Expo. **8**, University of Toronto Press, Toronto.

Müller, H.-G. (1988). *Nonparametric regression analysis of longitudinal data*. Springer, Berlin.

Müller, H.-G. (1991). Smooth optimum kernel estimators near endpoints. *Biometrika* **78**, 521–530.

Press, W. H., Flannery, B. F., Teukolsky, S. A. & Vetterling, W. T. (1992). *Numerical recipes: the art of scientific computing*. Cambridge University Press, Cambridge.

Priestley, M. B. & Chao, M. T. (1972). Non-parametric function fitting. *J. Roy. Statist. Soc. Ser. B* **34**, 385–392.

Rice, J. A. (1984). Boundary modification for kernel regression. *Comm. Statist. Theory Meth.* **13**, 893–900.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika* **68**, 45–54.

Stadtmüller, U. (1986). Asymptotic properties of nonparametric curve estimators. *Period. Math. Hungar.* **17**, 83–108.

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

Received March 1997, in final form February 1998

S. X. Chen, Department of Statistical Science, La Trobe University, VIC 3083, Australia.