ELSEVIER Computational Statistics & Data Analysis 31 (1999) 131–145

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Beta kernel estimators for density functions

Song Xi Chen

Department of Statistical Science, La Trobe University, Bundoora, VIC 3083, Australia

Received 1 May 1998; received in revised form 1 February 1999

Abstract

Kernel estimators using non-negative kernels are considered to estimate probability density functions with compact supports. The kernels are chosen from a family of beta densities. The beta kernel estimators are free of boundary bias, non-negative and achieve the optimal rate of convergence for the mean integrated squared error. The proposed beta kernel estimators have two features. One is that the different amount of smoothing is allocated by naturally varying kernel shape without explicitly changing the value of the smoothing bandwidth. Another feature is that the support of the beta kernels can match the support of the density function; this leads to larger effective sample sizes used in the density estimation and can produce density estimates that have smaller finite-sample variance than some other estimators. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Beta kernels; Boundary bias; Density estimation; Local linear estimators; Variable kernels

1. Introduction

Let $X_1,\dots,X_n$ be a random sample from a distribution with an unknown probability density function f. A standard kernel density estimator for f is

$$\hat{f}(x) = (nh)^{-1} \sum K\{h^{-1}(x - X_i)\}, \tag{1.1}$$

where K and h are the kernel function and the smoothing bandwidth, respectively. Comprehensive reviews of the kernel smoothing method are available in Silverman (1986), Scott (1992) and Wand and Jones (1995). The standard kernel estimator (1.1) was developed primarily for densities with unbounded supports. The kernel function K is usually symmetric and is regarded as less important than the smoothing bandwidth. While using a symmetric kernel is appropriate for fitting densities with

0167-9473/99/$ - see front matter
PII: S 0167-9473(99)00010-9

unbounded supports, it is not adequate for densities with compact supports as it causes boundary bias.

Removing boundary bias has been an active area of research. Data reflection was proposed by Schuster (1985), but is only effective when the density functions have a zero derivative at boundary points. Boundary kernels were suggested by Müller (1991). Cowling and Hall (1996) proposed generating pseudodata beyond the boundary points by linear interpolation of the order statistics. Marron and Ruppert (1994) considered using empirical transformations. In recent years, the local polynomial regression method of Cleveland (1979) has been applied to density estimation by Lejeune and Sarda (1992) and Jones (1993). Another version of the local linear estimator was considered by Cheng et al. (1996) by data binning and a local polynomial fitting on the bin counts. As the local linear estimators can have negative values (so too the boundary kernel estimators), Jones and Foster (1996) proposed a non-negative estimator by combining a local linear estimator and a re-normalizing kernel estimator.

In the simple spirit of the standard kernel estimator (1.1), this paper considers estimators using a beta family of density functions as kernels to estimate density functions with compact supports. The idea of smoothing using the beta kernels has been discussed in Chen (1999) in the context of non-parametric regression. This paper concentrates on density estimation and finite-sample comparisons with the local linear estimator of Lejeune and Sarda (1992) and Jones (1993) and its non-negative modification proposed by Jones and Foster (1996). The beta kernel method is easy both in concept and in its implementation, and produces estimators which are free of boundary bias, always non-negative and which have mean integrated squared errors of $O(n^{-4/5})$. There are two special features about the beta kernels. One is that the shape of the beta kernels varies naturally, which leads to the amount of smoothing being changed according to the position where the density estimation is made without explicitly changing the bandwidth; this implies that the beta kernel estimators are adaptive density estimators. The other feature is that the support of the beta kernels matches the support of the density function; this leads to larger effective sample size used in the density estimation and can produce density estimates that have smaller variance.

The paper is structured as follows. In Section 2 beta kernel estimators are introduced. Their bias and variance properties are studied in Section 3. In Section 4, the mean integrated squared errors and the optimal smoothing bandwidths are derived. Section 5 compares the variance of the beta kernel estimators with that of a local linear estimator. The beta kernel estimators are used in Section 6 to analyze a tuna data set as an example. Section 7 presents results from a simulation study.

2. Beta kernel estimators

The idea of beta kernel smoothing was motivated by the Bernstein theorem in mathematical function analysis, as explained in Brown and Chen (1999) when considering estimation of regression curves with equally spaced fixed design points by combining the beta kernels and Bernstein polynomials. The Bernstein theorem states that for any function which is continuous and has a bounded support its Bernstein polynomials converge to the function uniformly. The uniform convergence means the Bernstein polynomials are free of boundary bias. The Bernstein polynomials use a special type of beta kernels: the binomial kernels which contribute to a shortcoming of undersmoothing, as a smoothing bandwidth of $n^{-1/2}$ is implicitly used. However, the unique feature exhibited in the binomial kernels is valuable and motivating. Chen (1999) uses only the beta kernels without the Bernstein polynomials to smooth regression curves with arbitrary design points. We consider the use of beta kernel for probability density estimation in this paper and compare with the two local linear density estimators. A referee has pointed out that the idea of beta kernel smoothing was considered before by Stine and Bloomfield (1978) in an unpublished research report. The author has since located another paper by Harrell and Davis (1982) in using the beta kernels for quantile smoothing. However, these schemes of using beta kernels had a problem of being lack of an explicit smoothing parameter, and was not promising for general use in curve estimation.

Let $X_1,\dots,X_n$ be a random sample from a distribution with an unknown probability density function f which has a compact support. We assume that the compact support is known and, without loss of generality, is $[0,1]$ and f has continuous second derivative. Let $K_{p,q}$ be the density function of a $Beta(p,\ q)$ random variable.

We consider two beta kernel estimators for f. The first-beta kernel estimator uses $K_{x/b+1,(1-x)/b+1}$ as the kernel at $x\in[0,1]$, where b is a smoothing parameter satisfying the condition that $b\to 0$ as $n\to\infty$, and is defined as, for $x\in[0,1]$,

$$\hat f_1(x)=n^{-1}\sum_{i=1}^{n}K_{x/b+1,(1-x)/b+1}(X_i).$$

It is similar to the standard kernel estimator (1.1), only replaces a fixed kernel with the beta kernels.

In order to reduce the bias of $\hat f_1$ as outlined in Section 4, another beta kernel estimator is

$$\hat f_2(x)=n^{-1}\sum_{I=1}^{n}K^{\star}_{x,b}(X_i),$$

where $K^{\star}_{x,b}$ are boundary beta kernels defined as

$$K^{\star}_{x,b}(t)=\begin{cases}K_{x/b,(1-x)/b}(t) & \text{if } x\in[2b,1-2b],\\ K_{\rho(x),(1-x)/b}(t) & \text{if } x\in[0,2b),\\ K_{x/b,\rho(1-x)}(t) & \text{if } x\in(1-2b,1],\end{cases}$$

where $\rho(x,b)=2b^2+2.5-\sqrt{4b^4+6b^2+2.25-x^2-x/b}$. Note that for each fixed b $\rho(x,b)$ is a monotonic increasing function of x between $[0,2b]$. In particular, it has $y=x/b$ as its tangent line at $x=2b$ and $\rho(0,b)=1$.

An eminent feature of the beta kernels is that the kernel shape changes according to the value of x, as shown by a set of beta kernels displayed in Fig. 1 (all the figures in the paper are produced using S-plus). This varying kernel shape in fact changes the amount of smoothing applied by the beta kernel estimators. To appreciate this

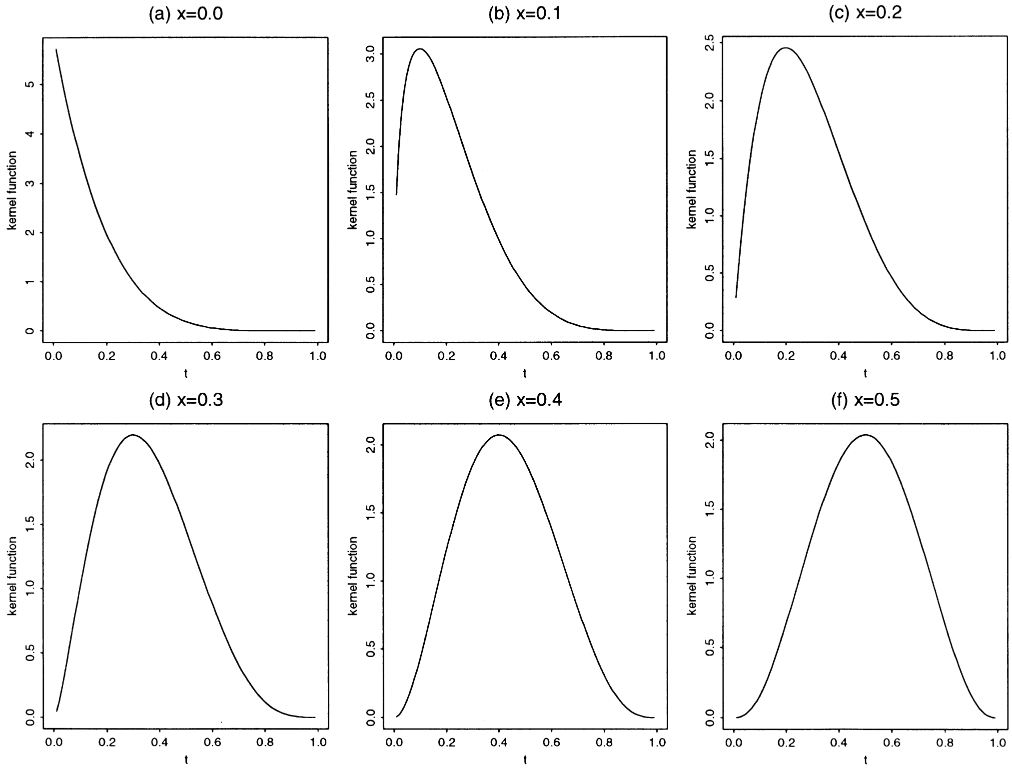

Fig. 1. Beta kernels $K_{x/b+1,(1-x)/b+1}(t)$ for $b = 0.2$.

point, we notice that the variance of a $Beta\{x/b+1,\ (1-x)/b+1\}$ random variable is

$$(b^{-1}x+1)\{b^{-1}(1-x)+1\}(b^{-1}+2)^{-2}(b^{-1}+3)^{-1} = bx(1-x) + \mathrm{O}(b^2).$$

This beta kernel adaptive scheme is in contrast with the existing adaptive scheme where a different amount of smoothing is achieved by changing the value of smoothing bandwidth. Also the beta kernels are non-negative which implies that the beta kernel estimators are non-negative as well.

3. Bias and variance

Note that

$$E\{\hat{f}_1(x)\} = \int_0^1 K_{x/b+1,(1-x)/b+1}(y) f(y)\, \mathrm{d}y = E\{f(\xi_x)\},$$

where ξ_x is a $Beta\{x/b+1, (1-x)/b+1\}$ random variable. Using the same derivation for the bias of the beta kernel regression estimator given in Chen (1999), it can be shown that the bias of the beta estimator is

$$Bias\{\hat{f}_1(x)\} = \{(1-2x)f'(x) + \tfrac{1}{2}x(1-x)f''(x)\}b + \mathrm{o}(b), \tag{3.1}$$

where the remainder term is uniformly $\mathrm{o}(b)$ for $x \in [0,1]$. The bias is of $\mathrm{O}(b)$ throughout $[0,1]$, indicating that $\hat{f}_1$ is free of boundary bias.

The involvement of f' in the bias is due to the fact that x is not the mean of the $Beta\{x/b+1,(1-x)/b+1\}$ distribution; rather, it is the mode. The use of the boundary beta kernels $K^{\star}_{x,b}$ largely eliminates the involvement of f' in the bias of $\hat{f}_2$. Indeed using the same method that derives (3.1), it may be shown that

$$Bias\{\hat{f}_2(x)\} = \begin{cases} \tfrac{1}{2}x(1-x)f''(x)b + \mathrm{O}(b^2) & \text{if } x \in [b, 1-2b], \\ \xi(x)bf'(x) + \mathrm{o}(b) & \text{if } x \in [0, 2b), \\ -\xi(1-x)bf'(x) + \mathrm{o}(b) & \text{if } x \in (1-2b, 1], \end{cases}$$

where $\xi(x) = (1-x)\{\rho(x) - x/b\}/\{1 + b\rho(x) - x\}$. Now f' is removed from the bias in the interior, and is only present in small areas near the boundaries but is compensated for by the disappearance of f''. The integrated squared bias is

$$IB^2\{\hat{f}_2(x)\} = \int_0^1 Bias^2\{\hat{f}_2(x)\}\,\mathrm{d}x = \tfrac{1}{4}b^2 \int_0^1 \{x(1-x)f''(x)\}^2\,\mathrm{d}x + \mathrm{o}(b^2) \tag{3.2}$$

which does not involve f'.

The variance of $\hat{f}_1$ is

$$\begin{aligned} Var\{\hat{f}_1(x)\} &= n^{-1} Var\{K_{x/b+1,(1-x)/b+1}(X_i)\} \\ &= n^{-1}\left[E\{K_{x/b+1,(1-x)/b+1}(X_i)\}^2 - [E\{\hat{f}_b(x)\}]^2\right]. \end{aligned}$$

It is easy to show that

$$E\{K_{x/b+1,(1-x)/b+1}(X_i)\}^2 = A_b(x)E\{f(\gamma_x)\}, \tag{3.3}$$

where γ_x is a $Beta\{2x/b+1, 2(1-x)/b+1\}$ random variable and

$$A_b(x) = \frac{B\{2x/b+1, 2(1-x)/b+1\}}{B^2\{x/b+1,(1-x)/b+1\}}. \tag{3.4}$$

We need the following lemma whose proof is given in Chen (1999) to determine the order of magnitude of $A_b(x)$.

Lemma. *For b small enough,*

$$A_b(x) \le \frac{1}{2\sqrt{\pi}}\{x(1-x)\}^{-1/2} b (b^{-1}+1)^{3/2} \quad \textit{for any } x \in [0,1]$$

and also

$$A_b(x) \sim \begin{cases} \frac{1}{2\sqrt{\pi}}\{x(1-x)\}^{-1/2} b^{-1/2} & \textit{if } x/b \textit{ and } (1-x)/b \to \infty, \\ \frac{\Gamma(2\kappa+1)}{2^{1+2\kappa}\Gamma^2(\kappa+1)} b^{-1} & \textit{if } x/b \to \kappa \textit{ or } (1-x)/b \to \kappa \end{cases}$$

for a positive constant κ. As a result of the lemma and (3.3),

$$Var\{\hat{f}_1(x)\} = \begin{cases} \frac{1}{2\sqrt{\pi}} \frac{n^{-1}b^{-1/2}}{\{x(1-x)\}^{1/2}} \{f(x) + O(n^{-1})\} & \text{if } x/b \text{ and } (1-x)/b \to \infty, \\ \frac{\Gamma(2\kappa+1)}{2^{1+2\kappa}\Gamma^2(\kappa+1)} n^{-1}b^{-1}\{f(x) + O(n^{-1})\} & \text{if } x/b \to \kappa \text{ or } (1-x)/b \to \kappa. \end{cases} \tag{3.5}$$

The variance of $\hat{f}_2$ is similar. The only difference is that the multiplier in front of $n^{-1}b^{-1}$ in the case of x/b or $(1-x)/b \to \kappa$ has a slightly different form.

While the beta estimators are both free of boundary bias, their asymptotic variances are of a larger order $(n^{-1}b^{-1})$ near the boundaries than those $(n^{-1}b^{-1/2})$ in the interior. However, as will be seen in the next two sections that (i) the impact of the increased variance near the boundary on the mean integrated squared error is negligible; and (ii) the two beta kernel estimator tend to have smaller finite-sample variance than the two local linear estimators due to the fact that the support of the beta kernel matches that of the density f. Therefore, the result in (3.5) is asymptotic indeed.

4. Global properties

Let $\delta = b^{1-\varepsilon}$ where $0 < \varepsilon < 1$. From (3.5), for $i = 1$ or 2,

$$\begin{aligned} \int_0^1 Var\{\hat{f}_i(x)\}\,dx &= \int_0^{\delta} + \int_{\delta}^{1-\delta} + \int_{1-\delta}^{1} Var\{\hat{f}_1(x)\}\,dx \\ &= \int_{\delta}^{1-\delta} \frac{1}{2\sqrt{\pi}} \{x(1-x)\}^{-1/2} n^{-1}b^{-1/2} f(x)\,dx + O(n^{-1}b^{-\varepsilon}) \\ &= \frac{1}{2\sqrt{\pi}} n^{-1}b^{-1/2} \int_0^1 \{x(1-x)\}^{-1/2} f(x)\,dx + o(n^{-1}b^{-1/2}) \end{aligned} \tag{4.1}$$

by choosing ε properly.

Combining (3.1), (3.2) and (4.1), the mean integrated squared errors for $\hat{f}_1$ and $\hat{f}_2$ are, respectively,

$$\begin{aligned} MISE(\hat{f}_1) = b^2 \int_0^1 \{(1-2x)f'(x) + \tfrac{1}{2}x(1-x)f''(x)\}^2\,dx \\ + \frac{1}{2\sqrt{\pi}} n^{-1}b^{-1/2} \int_0^1 \{x(1-x)\}^{-1/2} f(x)\,dx + o(n^{-1}b^{-1/2} + b^2) \end{aligned} \tag{4.2}$$

and

$$MISE(\hat{f}_2) = \tfrac{1}{4}b^2 \int_0^1 \{x(1-x)f''(x)\}^2 \, \mathrm{d}x$$
$$+ \frac{1}{2\sqrt{\pi}} n^{-1} b^{-1/2} \int_0^1 \{x(1-x)\}^{-1/2} f(x) \, \mathrm{d}x + \mathrm{o}(n^{-1}b^{-1/2} + b^2). \tag{4.3}$$

The optimal bandwidths which minimize the dominant order terms in (4.2) and (4.3) are

$$b_1^\star = \frac{[\frac{1}{2\sqrt{\pi}} \int_0^1 \{x(1-x)\}^{-1/2} f(x) \, \mathrm{d}x]^{2/5}}{4^{2/5} \left[\int_0^1 \{(1-2x)f'(x) + \frac{1}{2}x(1-x)f''(x)\}^2 \, \mathrm{d}x\right]^{2/5}} n^{-2/5}$$

and

$$b_2^\star = \frac{\left[\frac{1}{2\sqrt{\pi}} \int_0^1 \{x(1-x)\}^{-1/2} f(x) \, \mathrm{d}x\right]^{2/5}}{\left[\int_0^1 \{x(1-x)f''(x)\}^2 \, \mathrm{d}x\right]^{2/5}} n^{-2/5}.$$

So, the optimal bandwidths are $\mathrm{O}(n^{-2/5})$ as compared with $h = \mathrm{O}(n^{-1/5})$ for the other kernel estimators. Substituting the above optimal bandwidths, we have the optimal mean integrated squared errors

$$MISE^\star(\hat{f}_1) = \frac{5}{4^{4/5}} \left[\frac{1}{2\sqrt{\pi}} \int_0^1 \frac{f(x)}{\{x(1-x)\}^{1/2}} \, \mathrm{d}x\right]^{4/5}$$
$$\times \left[\int_0^1 \{(1-2x)f'(x) + \tfrac{1}{2}x(1-x)f''(x)\}^2 \, \mathrm{d}x\right]^{1/5} n^{-4/5}$$

and

$$MISE^\star(\hat{f}_2) = \frac{5}{4^{3/5}} \left[\frac{1}{2\sqrt{\pi}} \int_0^1 \{x(1-x)\}^{-1/2} f(x) \, \mathrm{d}x\right]^{4/5}$$
$$\left[\int_0^1 \{x(1-x)f''(x)\}^2 \, \mathrm{d}x\right]^{1/5} n^{-4/5}.$$

Thus, both $\hat{f}_1$ and $\hat{f}_2$ achieve the optimal rate of convergence for the mean integrated squared errors.

It may be shown that for any density function f, if both $\int_0^1 \{f'(x)\}^2 \, \mathrm{d}x$ and $\int_0^1 \{f''(x)\}^2 \, \mathrm{d}x$ are finite, then

$$\int_0^1 \left\{(1-2x)f'(x) + \frac{1}{2}x(1-x)f''(x)\right\}^2 \mathrm{d}x \geq \int_0^1 \left\{\frac{1}{2}x(1-x)f''(x)\right\}^2 \mathrm{d}x.$$

This means that $MISE^\star(\hat{f}_1) \geq MISE^\star(\hat{f}_2)$ and $b_1^\star \geq b_2^\star$. Therefore, $\hat{f}_2$ should have a better global performance and use a smaller bandwidth than $\hat{f}_1$, and is thus recommended.

5. Variance comparison

Even though the variance behaviour of the beta kernel estimators near the boundaries has little effect on the mean integrated squared error, one may still be concerned about it. To allay this concern, in this section we compare the variance of the beta kernel estimators with that of the local linear estimator of Lejeune and Sarda (1992) and Jones (1993) and the non-negative estimator of Jones and Foster (1996).

For non-negative integers s and m and any symmetric kernel K with compact support $[-1,1]$, let define

$$a_{sm}(x,h)=\begin{cases}\displaystyle\int_{-1}^{x/h} t^s K^m(t)\,\mathrm{d}t & \text{if } x\in[0,1-h],\\ \displaystyle\int_{-(1-x)/h}^{1} t^s K^m(t)\,\mathrm{d}t & \text{if } x\in(1-h,1].\end{cases}$$

The local linear estimator of Lejeune and Sarda (1992) and Jones (1993) is

$$\hat f_l(x)=(nh)^{-1}\sum_1^n K_l\left(x,h,\frac{x-X_i^1}{h}\right),$$

where

$$K_l(x,h,t)=\frac{a_{21}(x,h)-a_{11}(x,h)t}{a_{01}(x,h)a_{21}(x,h)-a_{11}^2(x,h)}K(t)$$

is the local linear kernel used at x. Clearly, $K_l(x,h,t)=K(t)$ if $x>h$ implying that the original kernel K is used in the interior ($h\le x\le 1-h$). In the boundary areas ($x<h$ or $x>1-h$), $K_l(t)$ is a linear combination of $K(t)$ and $tK(t)$.

The above estimator can take negative values. To ensure non-negativity, Jones and Foster (1996) proposes a modification to $\hat f_l$, denoted as $\hat f_{\mathrm{nl}}$,

$$\hat f_{\mathrm{nl}}(x)=\bar f(x)\exp\{\hat f_l(x)/\bar f(x)-1\},$$

where $\bar f(x)=\hat f(x)/a_{01}(x,h)$ is the re-normalization of the standard kernel estimator (1.1) and is non-negative itself.

The bias and variance of $\hat f_l(x)$ and $\hat f_{\mathrm{nl}}(x)$ have been given in Jones (1993) and Jones and Foster (1996). Define

$$V(x,h)=\frac{a_{21}^2(x,h)a_{02}(x,h)-2a_{21}(x,h)a_{11}(x,h)a_{12}(x,h)+a_{11}^2(x,h)a_{22}(x,h)}{a_{21}(x,h)a_{01}(x,h)-a_{11}^2(x,h)}. \tag{5.1}$$

Note that for $p\ge 1$ $V(x,h)=R(K)=\int_{-1}^1 K^2(t)\,\mathrm{d}t$. Then, for $\hat f_E$ being either $\hat f_l$ or $\hat f_{\mathrm{nl}}$,

$$Var\{\hat f_E(x)\}=(nh)^{-1}V(x,h)f(x)+\mathrm{o}\{(nh)^{-1}\}. \tag{5.2}$$

As the two beta kernel estimators and the two local linear estimators basically have the same variance, respectively, a comparison is made only between $\hat f_1$ and $\hat f_l$ regarding the variance coefficient function. The variance coefficient function is the multiplier of $(nh)^{-1}f(x)$ or $(nb^{1/2})^{-1}f(x)$ in the expansion for the variance. To

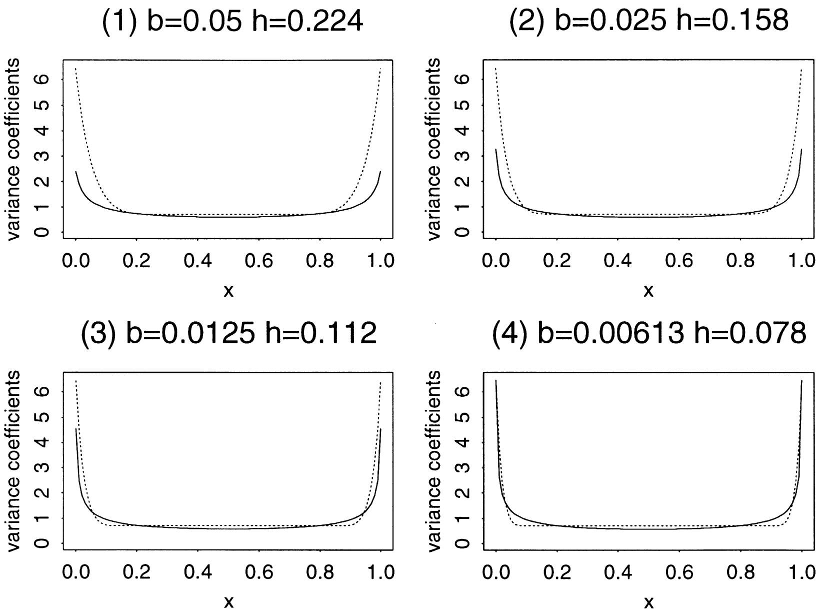

Fig. 2. Variance coefficient functions of the beta and the local linear estimators.

make the amount of smoothing used by the two smoothers in the same scale, we let $h=\sqrt{b}$ based on the facts that the optimal order for global h and b are of $n^{-1/5}$ and $n^{-2/5}$, respectively. The variance coefficient function for $\hat{f}_1$ is just $\sqrt{b}A_b(x)$, and that for the local linear smoother is $V(x,h)$ given in (5.1).

In Fig. 2 we plot the two variance coefficient functions for four levels of bandwidths. We see that the two coefficient functions are almost identical in the interior of $[0,1]$, and only differ in the boundaries. When the bandwidth is at a larger level in plots (1)–(3), the beta estimator has in fact a smaller variance coefficients near the boundaries. It is only when b is less than 0.00613 that the beta estimator will have a larger variance coefficient near the boundaries. Generally speaking, 0.00613 is a very small bandwidth value and requires a very large sample size to reach it. The exact sample size corresponding to the bandwidth value depends on the underlying density f. According to a formula for the optimal bandwidth given in the next section, if $f(x)$ is the density function of a truncated $\mathrm{N}(0,0.5^2)$ distribution in $[0,1]$, whose definition is given in Section 7, the sample size required to reach $b=0.00613$ is about 90000 for $\hat{f}_1$. If $f(x)=\alpha^{-1}\exp(-x/\alpha)/\{1.0-\exp(-1.0/\alpha)\}$, which is the density of the truncated $\exp(\alpha)$ distribution in $[0,1]$, the sample size is around 37 000 for $\alpha=0.5$. Therefore, generally speaking the boundary variance behaviour should not be a concern when the sample size is not very large.

It may be argued that Fig. 2 presents a theoretical picture only. To shed more light on the issue, in Fig. 3 we present simulated variance of the density estimates at $x=0$ using the two beta kernel, the local linear and Jones and Foster's non-negative estimators. The simulation results are part of a comprehensive simulation study whose details are described in Section 7. The compact Biweight kernel is used by the two

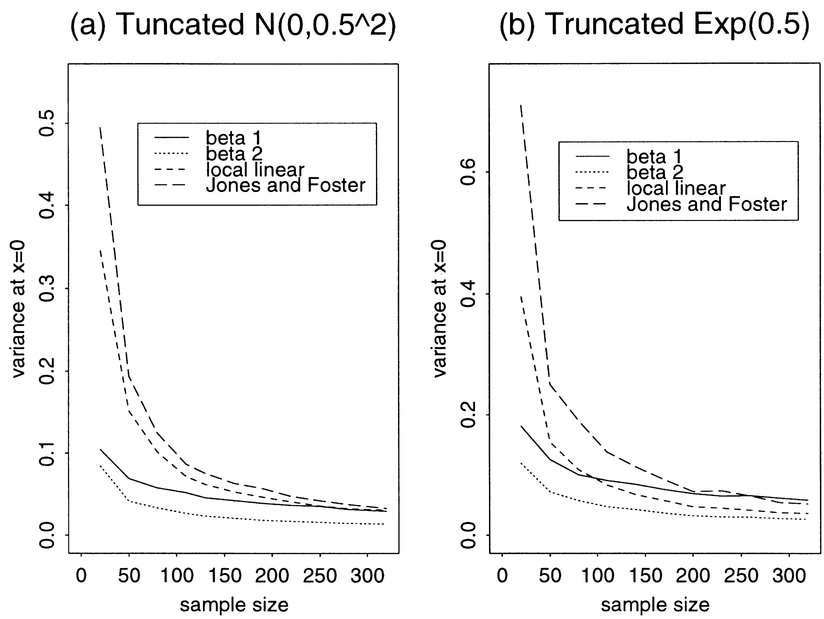

Fig. 3. Variance of four density estimates at $x = 0$.

non-beta kernel estimators. The sample size ranges from 20 to 300, and the underlying densities are, respectively, those of the truncated $N(0, 0.5^2)$ and the truncated $\exp(0.5)$ distributions. We observe that the modified beta kernel estimator $\hat{f}_2$ had the smallest variance among the estimators for all the cases considered. The first-beta kernel estimator had larger variance than the other two non-beta kernel estimators only in the case of exponential distribution when the sample size was large. It is interesting to note that the non-negative estimators of Jones and Foster had the largest variance for almost all the cases considered.

The smaller variance we just observed for the beta kernel estimators is probably due to the fact that the support of the beta kernels matches the support of the density. Thus, at any fixed x each data value contributes to the density estimate. This is in contrast to the local linear estimator with a compact kernel, where only data within a small area near x are utilised. So, the beta kernel estimators should have larger effective sample sizes than the local linear estimator and thus can reduce the variance.

The smaller effective sample size employed by a local linear regression estimator was revealed in the context of non-parametric regression by Seifert and Gasser (1996), though the concept of effective sample size was not mentioned. They noticed that a local linear regression estimator using a compact kernel had erratic behaviour, and that using the Gaussian kernel can stabilize the variance. Indeed, using the Gaussian kernel increases the effective sample size as it has support in $(-\infty, \infty)$. However, it is not quite appropriate to use a kernel with unbounded support to estimate densities with compact supports, as the entire density domain will be covered by boundary areas.

6. An example

We first apply the proposed beta kernel smoothers to a tuna data set, given in Chen (1996), which were collected from an aerial line transect survey to estimate the abundance of Southern Bluefin Tuna over the Great Australian Bight. The data are the perpendicular sighting distances (in miles) of detected tuna schools to the transect lines flown by a light airplane with tuna spotters on board. The tuna abundance estimation relies on the estimation of f, the probability density function of the perpendicular sighting distances. As only absolute distances are recorded, the sighting distances are confined in $[0, w]$ where w is the maximum sighting distance and is 20 miles. The closer a tuna school is to the transect line, the larger the probability of being detected. Therefore, there are relatively more X_i near 0. This means that the density f should be a non-increasing function and has a compact support $[0, 20]$.

For comparison, we also applied the local linear estimator of Jones (1993) using the Biweight kernel. We did not include the non-negative estimator of Jones and Foster (1996) as the local linear estimator respected non-negativity for this particular data set. So far, the beta kernel smoothers have been proposed for a density with $[0, 1]$ as its support. An extension to any support is easy. For example, for the tuna data

$$\hat{f}_2(x) = (nw)^{-1} \sum K^{\star}_{x/w,b}(X_i/w).$$

The cross validation method was used to choose the smoothing bandwidths by the three kernel smoothers, and prescribed $b_1 = 0.095$ for $\hat{f}_1$, $b_2 = 0.097$ for $\hat{f}_2$, $h = 0.884$ for the local linear estimator. As the cross validation method may sometimes not give good bandwidth prescription, other bandwidth values were also tested by visual inspections. It was found that the bandwidths given by the cross validation for the two beta estimators were quite adequate. But the one for the local linear estimator was too large, and $h = 0.6$ was used instead. The corresponding density estimates using the bandwidth are plotted in Fig. 4.

As the real underlying density function is unknown, it is hard to judge among the three density estimates. However, for line transect survey data a widely believed feature among practitioners for the density function is that the density should have a "shoulder" near $x = 0$ as the detection of animal schools near the transect line remains high in an area near the transect line. We find the two beta kernel estimates have, basically, "shoulders" near $x = 0$, and are quite close to each other near $x = 0$ and in the tails. The local linear estimate does not have a "shoulder" near $x = 0$. This lack of "shoulder" was also observed for other values of bandwidth tried.

7. Simulation results

In this section we report results of a simulation study designed to investigate performance of the proposed beta kernel estimators $\hat{f}_1$ and $\hat{f}_2$. For comparison purposes, the local linear estimator of Jones (1993) and the non-negative estimator of Jones and Foster (1996) were also evaluated. It should be pointed out that a comparison

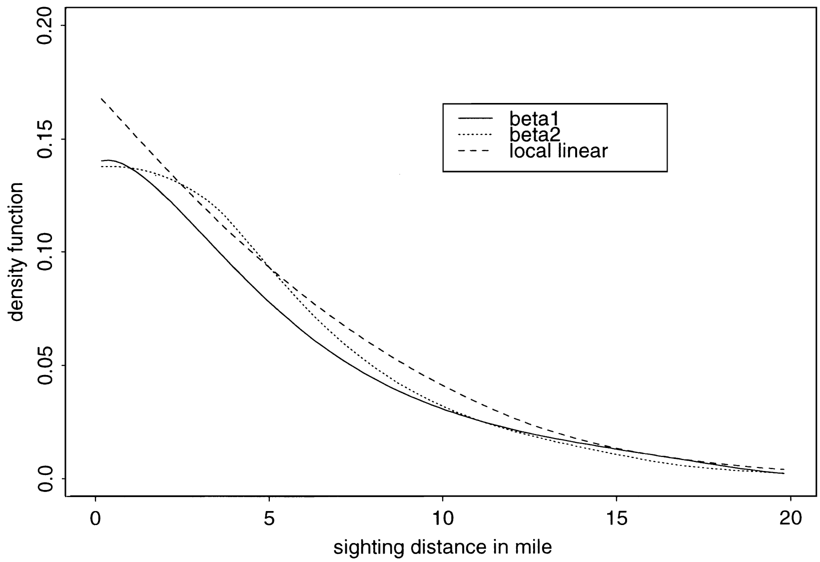

Fig. 4. Estimated density curves of the beta, the boundary kernel and the local linear kernel density estimates for a tuna data set.

or competition among various estimators is, at times, hardly fair. For instance, in the current comparison, the non-negative estimator is constructed using two density estimators, and the local linear estimator is free to take negative values.

Two density functions which have compact support $[0,1]$ were considered. One is

$$f(x) = \frac{2}{\sigma\sqrt{2\pi}} \exp\{-x^2/(2\sigma^2)\}\{2\Phi(1/\sigma) - 1\}^{-1}$$

which is the density of $|Y_i|$ where Y_i are the truncated $\mathrm{N}(0,\sigma^2)$ random variables in $[-1,1]$. Here Φ is the the standard normal distribution function. Another is

$$f(x) = \alpha^{-1}\exp(-x/\alpha)\{1.0 - \exp(-1/\alpha)\}^{-1}$$

which is the density of the truncated exponential distribution $\exp(\alpha)$ random variable in $[0,1]$.

Random samples were generated from the two distributions. The simulation program is written in C. We chose $\sigma = 0.25$ and 0.5 for the truncated normal distribution and $\alpha = 0.25$ and 0.5 for the truncated exponential distribution. The sample sizes used in the simulation ranged from 20 to 320. For each sample size, 1000 random samples were generated by modifying routines in Press et al. (1992). For each simulated sample and each estimator, the smoothing bandwidth was chosen by directly minimizing the integrated squared error (ISE). For the two beta kernel estimators,

$$ISE(b) = \int_0^1 \{\hat{f}_i(x) - f(x)\}^2 \, \mathrm{d}x$$

for $i = 1$ and 2, respectively, and similar expressions are available for the two non-beta kernel estimators. The minimization of the integrated squared error with respect to

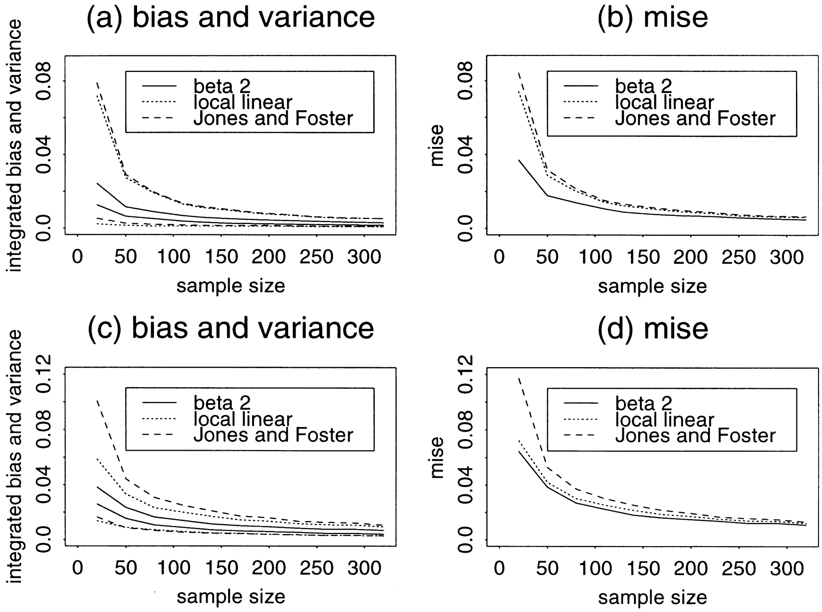

Fig. 5. Bias, variance and the *MISE* for the truncated normal $N(0,\sigma^2)$ density: $\sigma = 0.5$ for (a) and (b); $\sigma = 0.25$ for (c) and (d).

the smoothing bandwidth was carried out by the golden search algorithm given in Press et al. (1992).

By substituting the optimal bandwidths, the average integrated squared bias, the integrated variance and the integrated squared errors were calculated as measures of performance for each of the estimators. The results are summarized in Figs. 5 and 6. As expected from the result given at the end of Section 5, the simulation showed that the first-beta kernel estimator $\hat{f}_1$ always had a slightly larger bias and variance than $\hat{f}_2$. To minimize overcrowding, we do not include the simulation results of $\hat{f}_1$ in the figures.

The bias, the variance and the integrated squared error were all getting smaller when the sample size increased. And all three measures seemed to converge as n increased. We observe in panels (a) and (c) of Figs. 5 and 6 that the beta kernel estimator $\hat{f}_2$ had the smallest integrated variance for almost all the cases considered, except for the truncated exponential distribution with $\alpha = 0.25$ in (c) of Fig. 6 where the variance of $\hat{f}_2$ is only marginally larger than that of the local linear estimator. The non-negative estimator of Jones and Foster had the largest integrated variance for all the cases considered. The above results re-enforce the results given in Fig. 3 and the view that the beta estimators use larger effective sample sizes in their density estimation. Indeed, the larger sample size used by the beta kernel estimator contributed to a larger integrated bias as shown in the panels (a) and (c) of both figures.

Combining the bias and variance given in panels (a) and (c), we obtained the average minimized integrated errors shown in panels (b) and (d). Fig. 5 shows

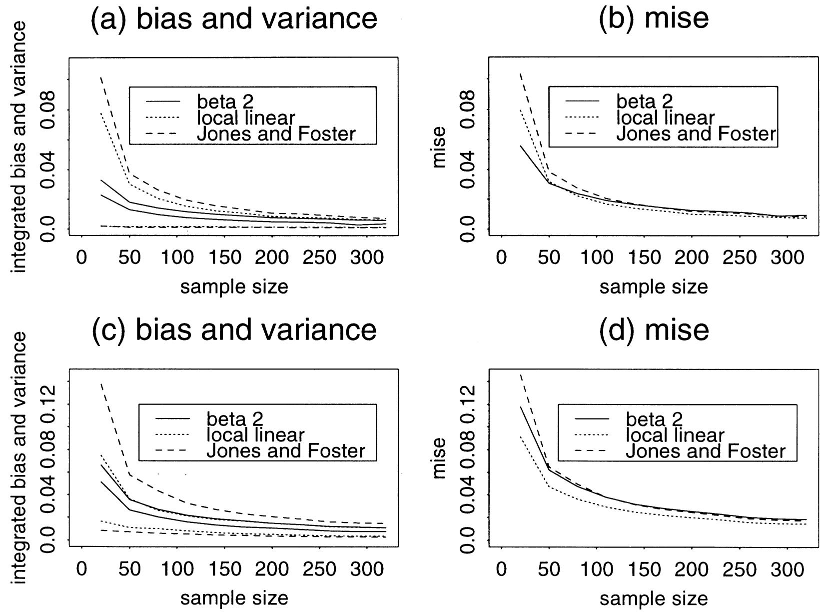

Fig. 6. Bias, variance and the *MISE* for the truncated normal exp(α) density: $\alpha = 0.5$ for (a) and (b); $\alpha = 0.25$ for (c) and (d).

that the beta kernel estimator had the smallest averaged integrated squared error for the truncated normal density which is sparse near $x = 1$. The local linear smoother performed better in Fig. 6 as the sparseness of the density was improved in the exponential density. Even in this case, the beta kernel estimator performed better when the sample size was small as observed in (b) of Fig. 6. The Jones–Foster estimator had the largest integrated error when n was less than 130 for both densities, and had almost the same integrated squared errors with the beta kernel estimator when n was large in Fig. 6.

In summary, the beta kernel estimator $\hat{f}_2$ is a serious competitor with the existing density estimators, and tends to produce density estimates which have smaller variance due to its larger effective sample sizes.

Acknowledgements

The author thanks Professor B.M. Brown for stimulating discussions, Professor M.C. Jones for constructive comments on an early version of the paper and a referee for bringing up some previous works on beta kernel smoothing into his attention.

References

Brown, B.M., Chen, S.X., 1999. Beta-Bernstein smoothing for regression curves with compact supports. Scand. J. Statist., 26, 47–59.

Chen, S.X., 1996. Empirical likelihood confidence intervals for nonparametric density estimation. Biometrika 83 (2), 329–341.

Chen, S.X., 1999. Beta kernel smoothers for regression curves. Statist. Sinica 9, to appear.

Cheng, M.Y., Fan, J., Marron, J.S., 1996. On automatic boundary corrections. Ann. Statist. 25, 1691–1708.

Cleveland, W.S., 1979. Robust locally weighted regression and smoothing scatterplots. J. Am. Statist. Assoc. 74, 829–836.

Cowling, A., Hall, P., 1996. On pseudodata methods for removing boundary effects in kernel density estimation. J. Roy. Statist. Soc. B 58, 551–563.

Harrell, F.E., Davis, C.E., 1982. A new distribution-free quantile estimator. Biometrika 69, 635–640.

Jones, M.C., 1993. Simple boundary correction for kernel density estimation. Statist. Comp. 3, 135–146.

Jones, M.C., Foster, P.J., 1996. A simple nonnegative boundary correction method for kernel density estimation. Statist. Sin. 6, 1005–1013.

Lejeune, M., Sarda, P., 1992. Smooth estimators of distribution and density functions. Comput. Statist. Data. Anal. 14, 457–471.

Marron, J.S., Ruppert, D., 1994. Transformation to reduce boundary bias in kernel density estimation. J. Roy. Statist. Soc. B 56, 653–671.

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

Press, W.H., Flannery. F., Teukolsky, S.A., Vettering, W.T., 1992. Numerical Recipes: the Art of Scientific Computing. Cambridge University Press, Cambridge.

Schuster, E.F., 1985. Incorporating support constraints into nonparametric estimators of densities. Commun. Statist. A 14, 1123–1136.

Scott, D.W., 1992. Multivariate Density Estimation. Wiley, New York.

Seifert, B., Gasser, Th., 1996. Finite sample variance of local polynomials: analysis and solutions. J. Am. Statist. Assoc. 91, 267–275.

Silverman, B.W., 1986. Density Estimation. Chapman & Hall, London.

Stine, R., Bloomfield, P., 1978. A descriptive analysis of hourly temperature variation, Department of Statistics, Princeton University, unpublished research report.

Wand, M., Jones, M.C., 1995. Kernel Smoothing. Chapman & Hall, London.