Ann. Inst. Statist. Math.
Vol. 54, No. 2, 312–323 (2002)
©2002 The Institute of Statistical Mathematics

LOCAL LINEAR SMOOTHERS USING ASYMMETRIC KERNELS

SONG XI CHEN

Department of Statistics and Applied Probability, National University of Singapore, Singapore 117543, Singapore, e-mail: stacsx@nus.edu.sg

(Received January 4, 2000; revised January 16, 2001)

Abstract. This paper considers using asymmetric kernels in local linear smoothing to estimate a regression curve with bounded support. The asymmetric kernels are either beta kernels if the curve has a compact support or gamma kernels if the curve is bounded from one end only. While possessing the standard benefits of local linear smoothing, the local linear smoother using the beta or gamma kernels offers some extra advantages in aspects of having finite variance and resistance to sparse design. These are due to their flexible kernel shape and the support of the kernel matching the support of the regression curve.

Key words and phrases: Beta kernels, gamma kernels, local linear smoother, nonparametric regression, sparse region.

1. Introduction

In recent years, local polynomial smoothing (Stone (1977), Cleveland (1979), Cleveland and Delvin (1988)) has been shown by Fan (1993), Fan and Gijbels (1992), Hastie and Loader (1993), Ruppert and Wand (1994) and others to be an effective smoothing method in nonparametric regression. It has the advantages of achieving full asymptotic minimax efficiency and automatically correcting for boundary bias. A review of local polynomial smoothing is given in Fan and Gijbels (1996). The standard application of the local polynomial smoothing has been focused on employing symmetric kernels, among them the compact kernels are popular choices. However, local polynomial smoothers using a compact kernel have a problem as the variance is unbounded in finite sample as revealed in Seifert and Gasser (1996*a*). Seifert and Gasser (1996*a*, 1996*b*) have proposed two methods to reduce the variability of the local linear estimator. One is to increase the bandwidth in areas of sparse design in order to include more design points. Another is to introduce a ridge parameter in the local linear estimator based on an empirical selection of the ridge parameter. Hall and Turlach (1997) used interpolation methods to add extra design points in sparse areas. In this paper we propose replacing the compact kernel in the local polynomial smoother by kernels whose support matches that of the curve in the local linear estimator in order to increase the effective sample size so as to reduce the variance of local polynomial smoothers. Seifert and Gasser (1996*a*) found that using the Gaussian kernel has attractive variance behaviour. This is because the Gaussian kernel has unbounded support that leads to a key matrix in local polynomial smoothing being non-singular, thus making the finite sample variance bounded.

When the curve under consideration has a bounded support, kernels whose support matches the support of the curve should have the same attractive variance properties as the Gaussian kernel. Recently, beta kernels, which are densities of some beta distributions, have been proposed to smooth the Bernstein polynomials in Brown and Chen (1999) and to construct Gasser-Müller estimator in Chen (2000*a*). Related gamma kernels are proposed for density estimation in Chen (2000*b*). These beta and gamma smoothers have the same order of bias throughout the support of the curve, achieve $n^{-4/5}$-order convergence for their mean integrated square errors and have attractive finite sample properties. It is also found, however, that their asymptotic mean square errors are of a larger order near the boundary. Even though this happens only in a small boundary area so small that it registers no effect on the mean integrated square error, it does pose an "asymptotic" hiccup for this nice idea of smoothing.

In this paper we consider local linear smoothing using the beta and gamma kernels. It turns out that local linear beta or gamma kernel smoothers remove the problem of an increased mean square error near the boundary, while maintaining the usual good properties of standard local linear smoothing with a fixed symmetric kernel. Using the beta or gamma kernels offers some extra benefits. Firstly, it is a kind of adaptive smoothing as both the beta and the gamma kernels have varying shapes and varying amounts of smoothness. Secondly, it has finite sample variance as long as there are two different design points not on the boundary of the support. The third is that the effective sample size is increased and thus the finite sample variance of the curve estimates can be reduced. These features are due to the fact that the support of the kernels matches the support of the regression curve. And the latter can make the local linear smoother having smaller variance when the curve has sparse areas. It has been shown that a local linear smoother using a symmetric kernel implicitly uses asymmetric kernels as well. However, the support of the kernels does not match that of the curve. This can produce problem in the variance when the design is sparse.

The paper is structured as follows. Section 2 introduces the local linear smoother using either beta or gamma kernels. The general properties using beta kernels are studied in Section 3, whereas those using gamma kernels are given in Section 4. Section 5 considers finite sample variance properties. Section 6 presents some simulation results. Some derivations are given in the Appendix.

2. Beta/gamma kernel based local linear smoothers

Let $Y_1, \ldots, Y_n$ be the responses of n design points $X_1, \ldots, X_n$ from a regression model

$$Y_i = m(X_i) + \epsilon_i \qquad i = 1, \ldots, n, \tag{2.1}$$

where $m(\cdot)$ is an unknown regression function with support $\mathcal{S}$ and the residual ϵ_i are uncorrelated random variables with zero mean and variance $\sigma^2(X_i)$. We consider in this paper that $\mathcal{S}$ is either $[0,1]$, which is bounded from both ends, or $[0,\infty)$, bounded from one end only.

The kernel we use to smooth at x is either

$$K_{x,b}(t) = \frac{t^{x/b}(1-t)^{(1-x)/b}}{B\{x/b+1, (1-x)/b+1\}} I(t \in \mathcal{S}) \quad \text{if} \quad \mathcal{S} = [0,1]$$

or

$$K_{x,b}(t) = \Gamma^{-1}(x/b+1) b^{-x/b-1} t^{x/b} e^{-t/b} I(t \in \mathcal{S}) \quad \text{if} \quad \mathcal{S} = [0,\infty)$$

where B and Γ are the beta and gamma functions respectively, and b is a smoothing parameter. Clearly, the kernel is the density of Beta$\{x/b+1,(1-x)/b+1\}$ or Gamma$\{x/b+1,b\}$ distribution. Both beta and gamma kernels have varying kernel shapes and become more asymmetric as x moves towards the boundary. Both beta and gamma kernels have $\mathcal{S}$ as the support, matching that of the regression curve.

The local linear smoother for m using either the beta or the gamma kernel $K_{x,b}$ is obtained by finding a and b that minimize

$$\sum_{j=1}^{n}\{Y_j-a-b(x-X_j)\}^2K_{x,b}(X_j).$$

Let $S_l(x)=n^{-1}\sum_{j=1}^{n}(x-X_j)K_{x,b}(X_j)$ for $l=0,1,2$. The local linear smoother $\hat{m}(x)=\hat{a}$, the solution of a to the above optimization, has a form

$$\hat{m}(x)=\sum_{j=1}^{n}w_j(x)Y_j/\sum_{j=1}^{n}w_j(x) \tag{2.2}$$

where the local linear weight

$$w_j(x)=n^{-1}\{S_2(x)-S_1(x)(x-X_j)\}K_{x,b}(X_j).$$

Let f be the density function of the design points. We assume throughout the paper that, for some positive constants f_c and σ_c^2,

$$\begin{cases}\text{(i) } m^{(2)}\in C(\mathcal{S}), f(\cdot) \text{ and } \sigma^2(\cdot) \text{ obey a first order Lipschitz condition in } \mathcal{S};\\ \text{(ii) } f(x)\geq f_c>0 \text{ and } \sigma^2(x)\leq\sigma_c^2 \text{ for all } x\in\mathcal{S};\\ \text{(iii) } b\to 0 \text{ and } nb^2\to\infty \text{ as } n\to\infty;\end{cases} \tag{2.3}$$

Let ξ_x be the beta or gamma random variable with $K_{x,b}$ as its density, $p_l(x)=E(\xi_x-x)^l$ for $l=0,1,2,\ldots$ and

$$A_b(x)=\begin{cases}\dfrac{B\{2x/b+1,2(1-x)/b+1\}}{B^2\{x/b+1,(1-x)/b+1\}} & \text{if } \mathcal{S}=[0,1];\\[2ex] \dfrac{\Gamma(2x/b+1)}{b2^{2x/b+1}\Gamma^2(x/b+1)} & \text{if } \mathcal{S}=[0,\infty).\end{cases}$$

We first present the following general theorem, whose proof is given in the Appendix.

THEOREM 1. *Assume the conditions given in* (2.3). *Then for any* $x\in\mathcal{S}$

$$\text{Bias}\{\hat{m}(x)\}=\frac{1}{2}m^{(2)}(x)p_2(x)+o\{p_2(x)\}+O\{n^{-1}A_b(x)\} \quad \textit{and} \tag{2.4}$$

$$\text{Var}\{\hat{m}(x)\}=n^{-1}A_b(x)\sigma^2(x)f^{-1}(x)+o\{n^{-1}A_b(x)\}. \tag{2.5}$$

Remark. The power of the local linear (polynomial) smoothing is to make the leading bias term free of the first derivatives of m and f. This power is maintained in beta or gamma kernels. It is this missing first derivatives in the bias that removes the increased mean square error near the boundary which was associated with the earlier beta/gamma kernel smoothers.

3. Asymptotic properties using beta kernels

We study the asymptotic properties of the local linear beta smoother $\hat{m}(x)$ assuming $S = [0,1]$ in this section. To simplify notations, we define $x \in S$ to be a

"interior x" if "x/b and $(1-x)/b \to \infty$" and
"boundary x" if "x/b or $(1-x)/b \to \kappa$"

where κ is a nonnegative constant. Clearly any fixed $x \in (0,1)$ which is free of n is an interior point.

From the basic properties of beta random variables, $E(\xi_x) = (x+b)/(1+2b)$ and $\mathrm{Var}(\xi_x) = b(x+b)(1-x+b)/\{(1+2b)^2(1+3b)\}$. Thus,

$$p_2(x) = \frac{bx(1-x) + b^2\{2-4x(1-x)\} + b^3\{4-12x(1-x)\}}{(1+2b)^2(1+3b)} = \begin{cases} bx(1-x) + O(b^2) & \text{for interior } x; \\ (2+\kappa)b^2 + O(b^3) & \text{for boundary } x. \end{cases} \tag{3.1}$$

Chen (2000a) shows that for small b

$$A_b(x) = \begin{cases} b^{-1/2}\dfrac{1}{\sqrt{4\pi}\sqrt{x(1-x)}} + o(b^{-1/2}) & \text{for interior } x; \\ b^{-1}\dfrac{\Gamma(2\kappa+1)}{2^{2\kappa+1}\Gamma^2(\kappa+1)} + o(b^{-1}) & \text{for boundary } x. \end{cases} \tag{3.2}$$

From the Theorem 1, (3.2) and (3.2), we have

$$\mathrm{Bias}\{\hat{m}(x)\} = \begin{cases} \dfrac{1}{2}x(1-x)m^{(2)}(x)b + O(b^2) & \text{for interior } x; \\ \dfrac{1}{2}(2+\kappa)m^{(2)}(x)b^2 + O(b^3) & \text{for boundary } x \end{cases} \tag{3.3}$$

and

$$\mathrm{Var}\{\hat{m}(x)\} = \begin{cases} n^{-1}b^{-1/2}\dfrac{\sigma^2(x)}{\sqrt{4\pi}\sqrt{x(1-x)}f(x)} + o(n^{-1}b^{-1/2}) & \text{for interior } x; \\ n^{-1}b^{-1}\dfrac{\sigma^2(x)\Gamma(2\kappa+1)}{2^{2\kappa+1}\Gamma^2(\kappa+1)f(x)} + o(n^{-1}b^{-1}) & \text{for boundary } x. \end{cases} \tag{3.4}$$

In the boundary areas, in terms of the order of magnitude of b the bias is of a smaller order whereas the variance is of larger order than those in the interior. However, b does not represent the total amount of smoothing used; rather, $p_2(x)$ is the real amount of smoothing used at x. The trade-off between the bias and the variance is directly due to $p_2(x)$ having different orders between the boundary and the interior as shown in (3.1).

By adjusting b so that $p_2(x)$ is of the same order within $[0,1]$, the mean square error can be made of order $n^{-4/5}$ everywhere within $[0,1]$. To appreciate this, notice that

$$MSE\{\hat{m}(x)\} = \begin{cases} n^{-1}b^{-1/2}\dfrac{\sigma^2(x)}{\sqrt{4\pi}\sqrt{x(1-x)}f(x)} + \dfrac{1}{4}x^2(1-x)^2\{m^{(2)}(x)\}^2b^2 \text{ or} \\ n^{-1}b^{-1}\dfrac{\Gamma(2\kappa+1)\sigma^2(x)}{2^{2\kappa+1}\Gamma^2(\kappa+1)f(x)} + \dfrac{1}{4}(2+\kappa)^2\{m^{(2)}(x)\}^2b^4 \end{cases}$$

respectively in the interior with error terms of $o(n^{-1}b^{-1/2}+b^2)$ or in the boundary with error terms of $o(n^{-1}b^{-1}+b^4)$. The optimal bandwidth is

$$(3.5)\quad b^\star(x)=\begin{cases}\left[\dfrac{\sigma^2(x)}{\sqrt{4\pi}x^{5/2}(1-x)^{5/2}f(x)\{m^{(2)}(x)\}^2}\right]^{2/5}n^{-2/5} & \text{for interior } x;\\[2ex] \left[\dfrac{\Gamma(2\kappa+1)\sigma^2(x)}{2^{2\kappa+1}\Gamma^2(\kappa+1)f(x)(2+\kappa)^2\{m^{(2)}(x)\}^2}\right]^{1/5}n^{-1/5} & \text{for boundary } x.\end{cases}$$

The optimal mean square error, with an error term of $o(n^{-4/5})$, is

$$(3.6)\quad MSE^\star\{\hat{m}(x)\}=\begin{cases}\dfrac{5}{4}n^{-4/5}\left\{\dfrac{\sigma^2(x)m^{(2)}(x)}{\sqrt{4\pi}f(x)}\right\}^{4/5} & \text{for interior } x;\\[2ex] \dfrac{5}{4}n^{-4/5}(2+\kappa)^{1/5}\left\{\dfrac{\Gamma(2\kappa+1)\sigma^2(x)m^{(2)}(x)}{2^{2\kappa+1}\Gamma^2(\kappa+1)f(x)}\right\}^{4/5} & \\ & \text{for boundary } x.\end{cases}$$

Hence, the optimal mean square error is of order $n^{-4/5}$ throughout $[0,1]$. This improves the beta kernel estimator considered in Chen (2000*a*) whose mean square error is of order $n^{-2/3}$ in the boundary areas and is of order $n^{-4/5}$ in the interior. The optimal bandwidth given in (3.5) prescribes a larger b value in the boundary to offset a reduced value of $bx(1-x)$ as x approaches the boundaries. By doing so, the total amount of smoothness $p_2(x)=O(n^{-2/5})$ throughout $[0,1]$.

It is interesting to see that the optimal mean square error for interior x coincides with that of the local linear smoother using the Gaussian kernel. However, they are not asymptotically equivalent in the boundary areas and their finite sample handling of the smoothing may be quite different.

It may be shown in a manner similar to that given in Chen (2000*a*) that the bias and the variance in the boundary areas have negligible contribution to the mean integrated square error, that is with an error term of $o(n^{-1}b^{-1/2}+b^2)$

$$MISE(\hat{m})=n^{-1}b^{-1/2}\int_0^1\frac{\sigma^2(x)}{\sqrt{4\pi}\sqrt{x(1-x)}f(x)}dx+b^2\frac{1}{4}\int_0^1 x^2(1-x)^2\{m^{(2)}(x)\}^2dx.$$

The optimal global bandwidth

$$(3.7)\quad b^\star=\left\{\frac{1}{\sqrt{4\pi}}\int_0^1\frac{\sigma^2(x)}{\sqrt{x(1-x)}f(x)}dx\right\}^{2/5}\left\{\int_0^1\{x(1-x)m^{(2)}(x)\}^2dx\right\}^{-2/5}n^{-2/5},$$

and the optimal mean integrated square error is

$$(3.8)\quad MISE^\star(\hat{m})=\frac{4}{5}\left\{\frac{1}{\sqrt{4\pi}}\int_0^1\frac{\sigma^2(x)}{\sqrt{x(1-x)}f(x)}dx\right\}^{4/5}\left\{\int_0^1\{x(1-x)m^{(2)}(x)\}^2dx\right\}^{-4/5}n^{-4/5}.$$

4. Asymptotic properties using gamma kernels

In this section, we investigate the asymptotic properties of the local linear gamma smoother assuming $I = [0, \infty)$. We define $x \in S$ to be a

$$\text{"interior } x\text{" if "}x/b \to \infty\text{" or "boundary } x\text{" if "}x/b \to \kappa\text{"}. \tag{4.1}$$

As ξ_x is the Gamma$(x/b + 1, b)$ random variable, $E(\xi_x) = x + b$ and $\mathrm{Var}(\xi_x) = bx + b^2$. Thus,

$$p_2(x) = bx + 2b^2. \tag{4.2}$$

According to Chen (2000b) for small b, when $S = [0, \infty)$,

$$A_b(x) = \begin{cases} b^{-1/2}/\{\sqrt{4\pi}\sqrt{x}\} + o(b^{-1/2}) & \text{for interior } x; \\ b^{-1}\Gamma(2\kappa+1)/\{2^{2\kappa+1}\Gamma^2(\kappa+1)\} + o(b^{-1}) & \text{for boundary } x. \end{cases} \tag{4.3}$$

This is almost the same as that of the beta kernel given in (3.2) except that $\sqrt{1-x}$ does not appear when x is in the interior as $x = 1$ is no longer a boundary point.

Substituting (4.2) and (4.3) into Theorem 1, we will reproduce the formulae from (3.3) to (3.8) except that $1-x$ does not appear in these formulae when x is in the interior and the integration interval is $S = [0, \infty)$.

However, using the gamma kernels produces some unique and interesting properties. Notice that

$$\mathrm{Bias}\{\hat{m}(x)\} = \begin{cases} \frac{1}{2} x m^{(2)}(x) b + O(b^2) & \text{interior } x \\ \frac{1}{2}(2+\kappa) m^{(2)}(x) b^2 + O(b^3) & x \text{ in boundary} \end{cases} \tag{4.4}$$

and

$$\mathrm{Var}\{\hat{m}(x)\} = \begin{cases} n^{-1}b^{-1/2}\dfrac{\sigma^2(x)}{\sqrt{4\pi}\sqrt{x}f(x)} + o(n^{-1}b^{-1/2}) & \text{interior } x \\ n^{-1}b^{-1}\dfrac{\sigma^2(x)\Gamma(2\kappa+1)}{2^{2\kappa+1}\Gamma^2(\kappa+1)f(x)} + o(n^{-1}b^{-1}) & \text{boundary } x. \end{cases} \tag{4.5}$$

So, the variance decreases when x increases as $x^{-1/2}$ appeared in the leading term of the asymptotic variance. This property is highly desirable when estimating curves with sparse regions in the upper tail of the design density f. This is a property not shared by using other fixed symmetric kernels including the Gaussian kernel. The reduced variance for large x is gained at the price of increasing bias. This is equivalent to the strategy of using larger bandwidth values in areas where the design is sparse. The gamma kernel carries this strategy automatically in a natural manner. The varying kernel shape is not the only device for changing the amount of smoothing. The smoothing bandwidth can be altered from one place to another, like one does in the adaptive kernel smoothing (Abramson (1982)). However, its effectiveness would be reduced for beta/gamma kernel smoothing due to the varyin kernels.

5. Finite sample variance

One problem with a compact kernel based local polynomial smoother, as revealed in Seifert and Gasser (1996a), is that its finite sample variance can be infinite. This is because, in the case of the local linear smoother $\sum w_j(x) = S_2(x)S_0(x) - S_1^2(x)$, the denominator of (2.2) has a positive probability of being zero or arbitrarily small. To avoid this, Fan (1993) added n^{-2} in the denominator. It is shown in the following that using the beta or gamma kernels can eliminate the problem. The results are valid for local polynomial smoothers rather than just local linear smoothers.

Let

$$X_p = \begin{pmatrix} 1(x-X_1)\cdots(x-X_1)^p \\ \cdots\cdots\cdots \\ 1(x-X_n)\cdots(x-X_n)^p \end{pmatrix} \quad \text{and} \quad W = \mathrm{diag}(K_{x,b}(X_j)).$$

LEMMA 1. *If there are at least $p+1 \le n$ different design points not being on the boundary of $\mathcal{S} = [0,1]$, then X^TWX is non-singular.*

PROOF. Without loss of generality, we assume that the first $p+1$ design points are different. Write

$$W^{1/2} = \begin{pmatrix} W_{11}^{1/2} & 0 \\ 0 & W_{22}^{1/2} \end{pmatrix} \quad \text{and} \quad X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}$$

where $W_{11}^{1/2}$ and $X^{(1)}$ are both non-singular square matrices of $p+1$ order. This means $rank(W^{1/2}X) = p+1$. So, $rank(X_p^TWX_p) = p+1$ which finishes the proof.

In the case of local linear ($p=1$), it may be shown by some matrix algebra that

$$\sum w_j(x) = |X_p^TWX_p| = \frac{1}{2}\sum_{i \neq j}(X_i - X_j)^2 K_{x,b}(X_i)K_{x,b}(X_j), \tag{5.1}$$

which is a weighted measure of the spread of the design points. So, if there are two different design points not being on the boundary of $\mathcal{S}$ $\sum w_j(x) \neq 0$ as maintained in the lemma. Note in passing that (5.1), which is valid for symmetric kernels and appears has not been noticed before, is useful in computation to develop recursive formula.

Based on the lemma, it is readily true that the probability of $X_p^TWX_p$ being singular is zero if the design variable is either continuous or fixed but there are at least $p+1$ different points not being on the boundary of $\mathcal{S}$. The later case includes any equally spaced design as long as $n \ge p+3$. The condition on the design variable is very weak and is satisfied in almost all the practical cases. Therefore, in finite samples the local polynomial beta or gamma smoother has finite variance with probability 1. This generalizes a corollary in Seifert and Gasser (1996*a*) which states that local linear fits have finite variance if and only if the kernel function has noncompact support.

6. Simulation results

A simulation study was carried out and designed to investigate the performance of the beta or gamma kernel based local linear smoother. For comparison purposes, the local linear ridge smoother proposed in Seifert and Gasser (1996*a*) was also considered, which added a matrix $H = \binom{0\ 0}{0\ c_1}$ to X^TWX before matrix inversion. This was designed to stabilize the variance of the original local linear smoothers in areas of sparse design.

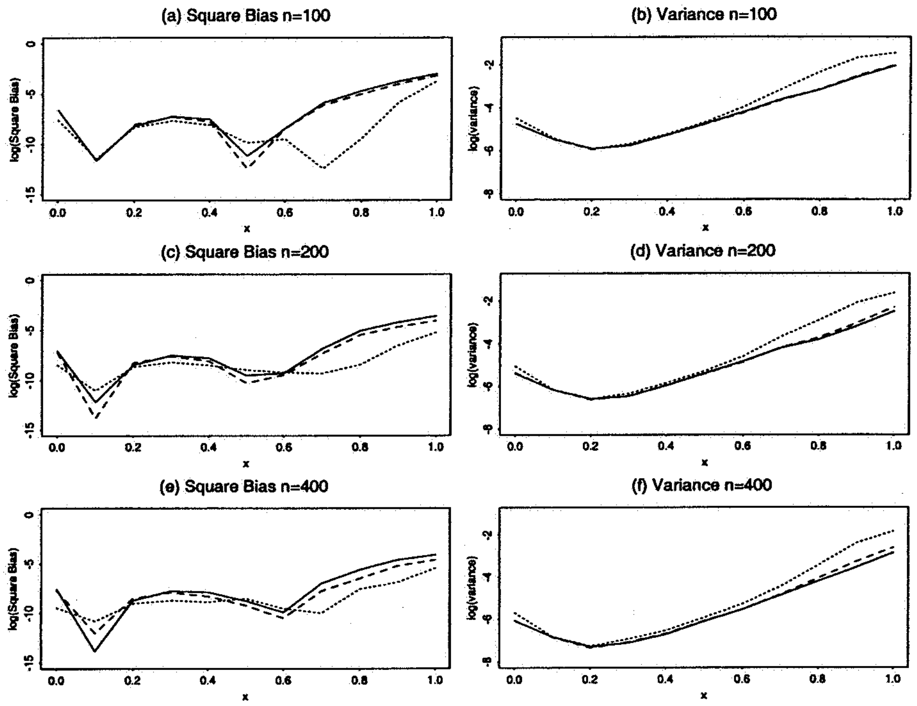

Fig. 1. Squared bias and variance of the local linear smoother using the beta kernels (solid lines), the Epanechnikov kernel (dotted lines) and the Gaussian kernel (dashed lines); $m(x) = (x - 0.5)^2 I(0 \le x \le 1)$ and the design points x_i are truncated $N(0, 0.3^2)$.

Readers should consult the paper for details. In the simulation, the Epanechnikov and the Gaussian kernels were employed with the ridged smoother.

The data were generated according to two regression models based on: $Y_i = m(X_i) + \epsilon_i$ where ϵ_i are independent $N(0, 0.05^2)$ random variables. In the first model, $m(x) = (x - 0.5)^2 I(0 \le x \le 1)$ and X_i were independent $|N(0, 0.3^2)|$ random variables truncated on $[-1, 1]$. In the second model, $m(x) = \exp(-x) + \exp\{-4(x - 1)^2\}$ for $x > 0$ and X_i were independent Gamma(2) random variables. So, in both models the design density was relatively sparse towards the right end of the support. As the regression curve is compactly supported in the first case, the gamma kernel based local linear smoother is not considered here. And similarly, the beta kernel based smoother was not considered in the second case. The simulation results were all based on 1000 simulations with the random variables generated using the algorithm given in Press *et al.* (1992). The sample sizes used in each model were $n = 100, 200$ and 400.

The performances of the smoothers were evaluated over a grid of equally spaced points within $[0, 1]$ in the first and $[0, 5.5]$ in the second model. The optimal global smoothing bandwidths which minimizes the MISE of each smoother over the above defined range were used in each simulation. After 1000 simulation, the average squared bias and variance were obtained as measures of performance.

The simulation results are summarized in Fig. 1 for the first model and Fig. 2 for the second model. Both figures show that local linear smoothers using the beta or gamma kernels had the best variance for almost all the sample sizes considered across the entire

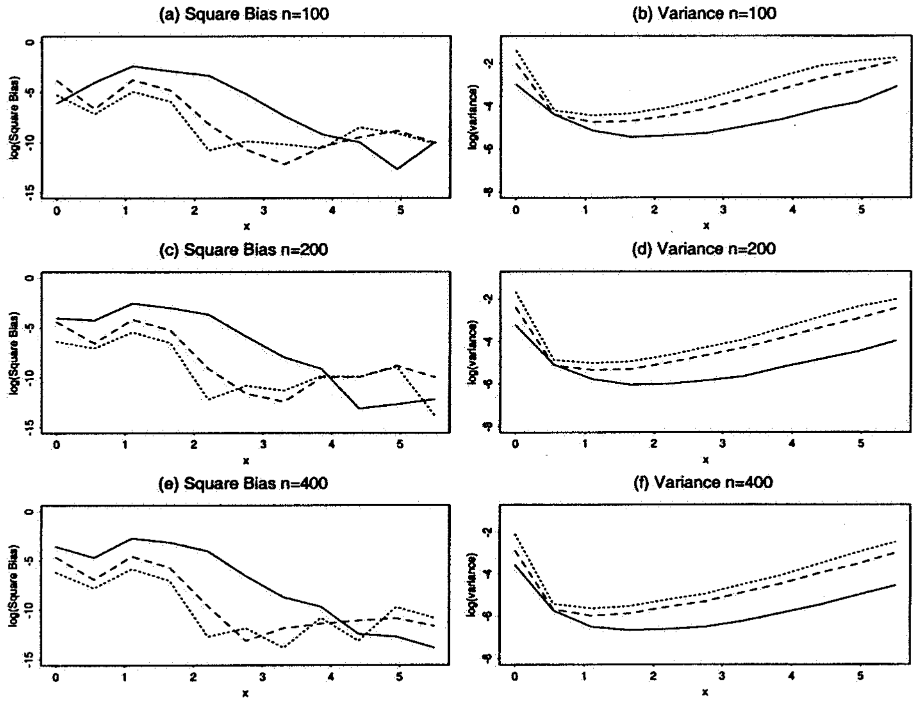

Fig. 2. Squared bias and variance of the local linear smoother using the gamma kernels (solid lines), the Epanechnikov kernel (dotted lines) and the Gaussian kernel (dashed lines); $m(x) = \exp(-x) + \exp\{-4(x-1)^2\}I(x > 0)$ and the design points x_i are Gamma(2).

range. They had significantly smaller variance when $x > 1$ in the second model where the design density becomes sparse as revealed in Fig. 2. In the first model, the performance of the beta and Gaussian smoothers were close to each other, and both had smaller variance and larger bias than those of the Epanechnikov smoother when $x > 0.8$. As the variance dominated the bias, as showed by the relative scale of the left versus the right panels, the overall MSE of the beta and Gaussian smoothers were smaller than that of the Epanechnikov kernel.

The increase in both bias and variance as x increases in Fig. 1 was due to (i) $m''(x)$ being constant and (ii) the design density is monotonic decreasing within $[0, 1]$. The shape of the variance in Fig. 2 was due to the Gamma(2) design distribution used which has a mode at $x = 1$. The increase in bias in Fig. 2 when $x \in [1, 2]$ was due to the fact that $\{m''(x)\}^2$ peaks at $x = 1.5$. The rise in bias at $x = 0$ was due to the combination of the boundary effect and the relatively lower level design. As expected, the bias of the gamma kernel based smoother was the largest in a large range of area, but was a good price paid in return for a much larger reduction in the variance.

Acknowledgements

The author would like to thank two referees for constructive comments which improve the presentation of the paper.

Appendix

PROOF OF THEOREM 1. We only prove the theorem for the case of beta kernels as that for the gamma kernels can be derived in a similar way.

Let $\mu_l(x) = E\{S_l(x)\}$, $T_l(x) = n^{-1}\sum_{j=1}^{n}(x-X_j)^l m(X_j)K_{x,b}(X_j)$ and $\nu_l(x) = E\{T_l(x)\}$ for $l = 0, 1, 2$, and $r(x) = m(x)f(x)$. Notice that

$$\begin{aligned}\mu_l(x) &= E\{K_{x,b}(X)(x-X)^l\} = \int_0^1 K_{x,b}(y)(x-y)^l f(y)dy \\ &= E\{f(\xi_x)(x-\xi_x)^l\}\end{aligned}$$

where ξ_x is a Beta$\{x/b+1, (1-x)/b+1\}$ random variable. Taylor expanding $f(\xi_x)$ at x,

$$\begin{aligned}f(\xi_x)(x-\xi_x)^l &= f(x)(x-\xi_x)^l - f^{(1)}(x)(x-\xi_x)^{l+1} + \frac{1}{2}f^{(2)}(x)(x-\xi_x)^{l+2} \\ &\quad + O_p\{(x-\xi_x)^{l+3}\}.\end{aligned} \tag{A.1}$$

It may be shown using the properties of beta random variables that $E(\xi_x - x) = (1-2x)b + O(b^2)$, $E\{(x-\xi_x)^2\} = bx(1-x) + O(b^2)$ and $E\{(x-\xi_x)^k\} = O(b^2)$ for $k \geq 3$. Apply the expectation on either side of (A.1), we have

$$\mu_l(x) = (-1)^l \sum_{j=0}^{2-l} f^{(j)}(x)p_{l+j}(x)/j! + o\{p_2(x)\}. \tag{A.2}$$

In a similar derivation,

$$\nu_l(x) = (-1)^l \sum_{j=0}^{2-l} r^{(j)}(x)p_{l+j}(x)/j! + o\{p_2(x)\}. \tag{A.3}$$

Notice that

$$\begin{aligned}\mathrm{Cov}\{S_{l_1}(x), S_{l_2}(x)\} &= n^{-1}\int K_{x,b}^2(t)(x-y)^{l_1+l_2}f(y)dy - n^{-1}\mu_{l_1}(x)\mu_{l_2}(x) \\ &= n^{-1}A_b(x)E\{(x-\gamma_x)^{l_1+l_2}f(\gamma_x)\} - n^{-1}\mu_{l_1}(x)\mu_{l_2}(x)\end{aligned} \tag{A.4}$$

where γ_x is the Beta$\{2x/b+1, 2(1-x)/b+1\}$ random variable. From Chen (2000a), $A_b(x) = O\{b^{-1/2}x^{-1/2}(1-x)^{-1/2}\}$. Also $E\{(x-\gamma_x)^{l_1+l_2}f(\gamma_x)\} = O\{\mu_{l_1+l_2}(x)\}$. Thus, it may be shown that the second term on the right hand side of (A.4) is of a smaller order than the first term. Therefore,

$$\mathrm{Cov}\{S_{l_1}(x), S_{l_2}(x)\} = \begin{cases} O\{n^{-1}A_b(x)\} & \text{if } l_1 + l_2 = 0; \\ o\{n^{-1}A_b(x)\} & \text{if } l_1 + l_2 \geq 1. \end{cases}$$

In general, we have

$$\mathrm{Cov}\{H_{l_1}(x), H_{l_2}(x)\} = \begin{cases} O\{n^{-1}A_b(x)\} & \text{if } l_1 + l_2 = 0; \\ o\{n^{-1}A_b(x)\} & \text{if } l_1 + l_2 \geq 1 \end{cases} \tag{A.5}$$

where H_l can be either S_l or T_l.

Let $\hat{r}(x) = S_2(x)T_0(x) - S_1(x)T_l(x)$ and $\hat{q}(x) = S_2(x)S_0(x) - S_1^2(x)$. Then $\hat{m}(x) = \hat{r}(x)/\hat{q}(x)$ and based on (A.5) we have

$$E\{\hat{m}(x)\} = \frac{\mu_2(x)\nu_0(x) - \mu_1(x)\nu_1(x)}{\mu_2(x)\mu_0(x) - \mu_1^2(x)} + O\{n^{-1}A_b(x)\}.$$

Substituting (A.2) and (A.3), and using the standard derivation for the bias of local linear smoothers, we may derive (2.4) of the theorem.

Note that

$$\begin{aligned}\text{(A.6)} \qquad \mathrm{Var}\{\hat{m}(x)\} &= \mathrm{Var}\left[E\left\{\frac{\sum w_j(x)Y_j}{\sum w_j(x)}\,\middle|\, X_1,\ldots,X_n\right\}\right] \\ &\quad + E\left[\mathrm{Var}\left\{\frac{\sum w_j(x)Y_j}{\sum w_j(x)}\,\middle|\, X_1,\ldots,X_n\right\}\right] \\ &= \mathrm{Var}\{\hat{r}(x)/\hat{q}(x)\} + E\left[\frac{\sum w_j^2(x)\sigma^2(X_j)}{\{\sum w_j(x)\}^2}\right]\end{aligned}$$

as $\sum w_j(x)m(X_j) = \hat{r}(x)$ and $\sum w_j(x) = \hat{q}(x)$.

Let $\eta(x) = \mu_2(x)\mu_0(x) - \mu_1^2(x)$ and $\beta(x) = \mu_2(x)\nu_0(x) - \mu_1(x)\nu_1(x)$. Then,

$$\begin{aligned}\mathrm{Var}\{\hat{r}(x)/\hat{q}(x)\} &= \eta^{-2}(x)\,\mathrm{Var}\{\hat{r}(x)\} - 2\eta^{-3}(x)\beta(x)\,\mathrm{Cov}\{\hat{r}(x),\hat{q}(x)\} \\ &\quad + \eta^{-4}(x)\beta^2(x)\,\mathrm{Var}\{\hat{q}(x)\} + o\{n^{-1}A_b(x)\} \\ &= \eta^{-2}(x)\mu_2^2(x)[\mathrm{Var}\{T_0(x)\} - 2\eta^{-1}(x)\beta(x)\,\mathrm{Cov}\{T_0(x),S_0(x)\} \\ &\qquad + \eta^{-2}(x)\beta^2(x)\,\mathrm{Var}\{S_0(x)\}] \\ &\quad + o\{n^{-1}A_b(x)\}.\end{aligned}$$

We use (A.5) in the derivation of the last equation. It may be shown in a similar fashion to that for deriving (A.5) that, with error terms of $o\{n^{-1}A_b(x)\}$,

$$\mathrm{Var}\{T_0(x)\} = n^{-1}A_b(x)f(x)m^2(x), \qquad \mathrm{Cov}\{T_0(x),S_0(x)\} = n^{-1}A_b(x)f(x)m(x) \quad \text{and}$$
$$\mathrm{Var}\{S_0(x)\} = n^{-1}A_b(x)f(x).$$

These and the fact that $\eta^{-1}(x)\beta(x) = m(x) + o(1)$ lead to

$$\text{(A.7)} \qquad \mathrm{Var}\{\hat{r}(x)/\hat{q}(x)\} = o\{n^{-1}A_b(x)\}.$$

To work out the second term on the right of (A.6), we define

$$W_l(x) = n^{-1}\sum_{j=1}^{n}(x-X_j)^l K^2_{x/b+1,(1-x)/b+1}(X_j)\sigma^2(X_j)$$

for non-negative integer l. It may be shown that

$$\text{(A.8)} \qquad \omega_l(x) = E\{W_l(x)\} = \begin{cases} A_b(x)\sigma^2(x)f(x) + o\{A_b(x)\} & \text{if } \ l=0; \\ o\{A_b(x)\} & \text{if } \ l \geq 1. \end{cases}$$

Then

$$
\begin{aligned}
\text{(A.9)} \quad & E\left[\frac{\sum w_j^2(x)\sigma^2(X_j)}{\{\sum w_j(x)\}^2}\right] \\
&= n^{-1}E\left[\frac{S_2^2(x)W_0(x) - 2S_1(x)S_2(x)W_1(x) + S_1^2(x)W_2(x)}{\{S_2(x)S_0(x) - S_1^2(x)\}^2}\right] \\
&= n^{-1}\left(\frac{\mu_2^2(x)\omega_0(x) - 2\mu_1(x)\mu_2(x)\omega_1(x) + \mu_1^2(x)\omega_2(x)}{\{\mu_2(x)\mu_0(x) - \mu_1^2(x)\}^2}\right)\{1 + O(1)\} \\
&= n^{-1}\omega_0(x)/\mu_0^2(x) + o\{n^{-1}A_b(x)\} \\
&= n^{-1}A_b(x)\sigma^2(x)/f(x) + o\{n^{-1}A_b(x)\}
\end{aligned}
$$

Combining (A.7) and (A.9) we prove (2.5) in the lemma, and thus finish the proof of the theorem.

References

Abramson, I. S. (1982). On bandwidth variation in kernel estimates—a square root law, *Ann. Statist.*, **10**, 1217–1223.

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

Chen, S. X. (2000*a*). Beta kernel smoother for regression curves, *Statist. Sinica*, **10**, 73–91.

Chen, S. X. (2000*b*). Probability density function estimation using gamma kernels, *Ann. Inst. Statist. Math.*, **52**, 471–480.

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

Cleveland, W. S. and Devlin, S. (1988). Locally weighted regression: An approach to regression analysis by local fitting, *J. Amer. Statist. Assoc.*, **83**, 596–610.

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

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

Fan, J. Q. and Gijbels, I. (1996) *Local Polynomial Modeling and Its Applications*, Chapman and Hall, London.

Hall, P. and Turlach, B. A. (1997). Interpolation methods for adapting to sparse design in nonparametric regression (with discussion), *J. Amer. Statist. Assoc.*, **92**, 466–476.

Hastie, T. and Loader, C. (1993). Local regression: Automatic kernel carpentry, *Statist. Sci.*, **8**, 120–143.

Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression, *Ann. Statist.*, **22**, 1346–1370.

Seifert, B. and Gasser, Th. (1996*a*). Finite sample variance of local polynomials: Analysis and solutions, *J. Amer. Statist. Assoc.*, **91**, 267–275.

Seifert, B. and Gasser, Th. (1996*b*). Variance properties of local polynomials and ensuring modifications (with discussion), *Statistical Theory and Computational Aspects of Smoothing* (eds. W. Härdle and M. G. Schimek), 5–79, Physica, Heidelberg.

Stone, C. J. (1977). Consistent nonparametric regression, *Ann. Statist.*, **5**, 595–645.

Press, W. H., Flannery, B. F., Teukolsky, S. A. and Vetterling, W. T. (1992). *Numerical Recipes in C. The Art of Scientific Computing*, 2nd ed., Cambridge University Press, Cambridge.