Statistica Sinica **6**(1996), 1005-1013

A SIMPLE NONNEGATIVE BOUNDARY CORRECTION METHOD FOR KERNEL DENSITY ESTIMATION

M. C. Jones and P. J. Foster

Open University and University of Manchester

Abstract: Without correction, kernel density estimates suffer from boundary effects. Many boundary corrections now exist, but almost all those with good theoretical performance allow the corrected estimator to become negative. An exception is provided by some recently proposed sophisticated transformation methodology. In this paper, we propose much simpler nonnegative boundary corrected estimators which are analogues of the wide class of simple, but possibly negative, boundary corrections based on generalized jacknifing.

Key words and phrases: Boundary kernels, generalized jackknifing, positivity, smoothing.

1. Introduction

Suppose we are doing kernel density estimation (e.g. Silverman (1986), Scott (1992), Wand and Jones (1995)) on data on the positive real line. Near the support boundary at the origin, the estimator is poor and, in fact, has considerable bias. This is because the kernel density estimator has no knowledge of the boundary and, in general, assigns probability mass outside the support.

A variety of boundary correction methods for kernel density estimation now exists, and most are referred to in Jones (1993). He sets up a unified approach to many of the more straightforward methods using "generalised jackknifing" (Schucany, Gray and Owen (1971)). To describe this, suppose that $\breve{f}(x) = n^{-1}\sum_i K_h(x-X_i)$ is a kernel density estimator based on data $X_1,\ldots,X_n$ employing the kernel function K. Here, $K_h(\cdot) = h^{-1}K(h^{-1}.)$ and will be taken to be a probability density function itself. Write $p = x/h$ and

$$a_l(p) = \int_{-S_K}^{\min\{p,S_K\}} u^l K(u)du,$$

where $[-S_K, S_K]$ is the support of K. Divide $\breve{f}(x)$ by $a_0(p)$ to give $\bar{f}$. The "local renormalisation" by division by $a_0(p)$ is, on its own, an inadequate form of boundary correction, as discussed in Section 2 of Jones (1993); also, it could be replaced by a reflection technique. Let $\tilde{f}$ be like $\bar{f}$ only with kernel function L, on $[-S_L, S_L]$, replacing K, let $c_l(p) = \int_{-S_L}^{\min\{p,S_L\}} u^l L(u)du$ and make the division by $c_0(p)$. Think of $\bar{f}$ and $\tilde{f}$ as being defined only on $[0, \infty)$. Then, in a minor reformulation of the presentation of Jones (1993), generalised jackknifing seeks a linear combination

$$\hat{f}(x) \equiv \alpha_x \bar{f}(x) + \beta_x \tilde{f}(x) \tag{1}$$

with good asymptotic bias properties. Away from the boundary, kernel density estimation typically affords a bias of order h^2 as $h = h(n) \to 0$. It turns out that the choices

$$\alpha_x = c_1(p)a_0(p)/\{c_1(p)a_0(p) - a_1(p)c_0(p)\}, \tag{1a}$$

$$\beta_x = -a_1(p)c_0(p)/\{c_1(p)a_0(p) - a_1(p)c_0(p)\}, \tag{1b}$$

allow $O(h^2)$ bias at and near the boundary also. (Note that $c_1(p)a_0(p)$ must not equal $a_1(p)c_0(p)$.) Observe that boundary corrected kernel density estimates typically do not integrate to unity, but could be renormalised to do so.

There are many possible choices for L. It is usually preferred to make L a function of K because then one has a boundary correction derived solely from the "interior kernel" K. Examples include taking $L(u)$ to be $K_c(u) = c^{-1}K(c^{-1}u)$ or $K'(u)$ or $K(2p - u)$ or $uK(u)$. The last of these is particularly popular. It results in the simple linear boundary kernel $(l_x + m_x u)K(u)$ where

$$l_x = a_2(p)/\{a_2(p)a_0(p) - a_1^2(p)\} \quad \text{and} \quad m_x = -a_1(p)/\{a_2(p)a_0(p) - a_1^2(p)\}.$$

This is used in the literature in many places, often as an ad hoc technique and sometimes as an automatic consequence of something else, e.g. local linear fitting (see Jones (1993)). It seems that the performance of many generalised jackknives is broadly equivalent, and hence linear boundary correction is as good as any.

A disadvantage of all generalised jackknife boundary corrections, however, is their propensity for taking negative values near the boundary. See the dashed curves in Fig. 1 where $n = 50$ data points are simulated from the Gamma $(3, 1)$ distribution (but only the boundary region $0 < x < h$ is shown). The purpose of this paper is to describe how to attain much the same boundary performance whilst retaining nonnegativity via a simple "nonnegativisation" device. This device can actually be applied to any boundary corrected density estimate and in particular yields a nonnegative version of each and every generalized jackknife boundary corrected method. Non–unit integral remains a feature of the result, but again renormalisation is possible (simulations suggest that renormalisation would make little difference). The successful results of applying the nonnegative analogue of $(l_x + m_x u)K(u)$ are included, as solid lines, in Fig. 1. Here, K is the biweight kernel $K(u) = (15/16)\{(1 - u^2)_+\}^2$, where $(\cdot)_+$ is the "nonnegative part" function, and $h = 1.3$, chosen as explained in Section 4.

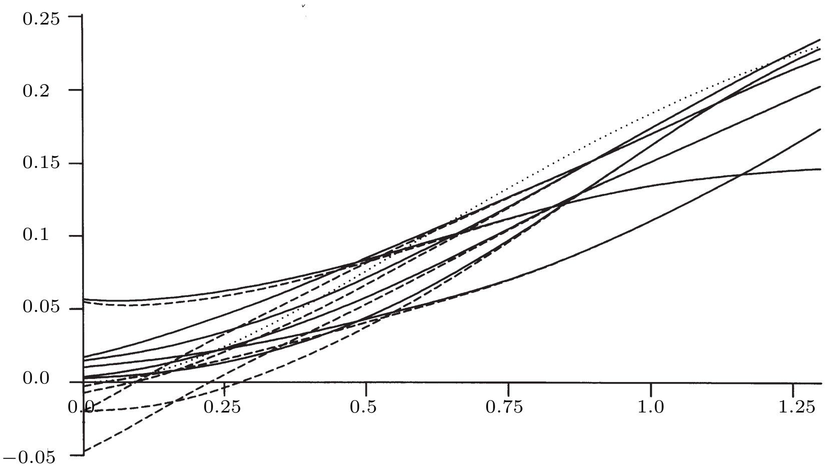

Figure 1. Six pairs of density estimates based on random samples of size 50 from the Gamma (3,1) distribution, $L(u) = uK(u)$, K the biweight kernel, and $h = 1.3$. Solid lines: $\hat{f}_P$, dashed lines: $\hat{f}$, dotted curve: f. The boundary region $0 < x < h$ only is shown.

Marron and Ruppert's (1994) excellent work on transformation-based boundary correction includes methods that remain nonnegative everywhere, and one that also has unit integral. But their methodology is complicated. Our competitors are very much simpler (see Section 2 for details). Properties of the new proposals will be derived in Section 3. Some further simulation evidence will be provided in Section 4. In cases where negative-allowing boundary corrections work well, so too do ours, and these are the cases dealt with in almost all of the boundary correction literature. But one intriguing arm of Marron and Ruppert's (1994) methodology which affords good estimation even when the underlying density has a pole at the boundary remains superior to what we can achieve. In the closing Section 5, we indicate why our approach is not quite as obvious as one might expect from, for example, related work of Jones and Foster (1993), and we mention other possible approaches.

2. The Methodology

Here is the basic idea. Recall that $\bar{f}(x)$ denotes the basic kernel density estimator $(0 \le x < \infty)$ divided by $a_0(p)$ and that $\hat{f}$ is the boundary corrected kernel density estimator given by (1), (1a) and (1b). Then the combination of $\bar{f}$ and $\hat{f}$ given by

$$\hat{f}_P(x) \equiv \bar{f}(x) \exp\left\{\frac{\hat{f}(x)}{\bar{f}(x)} - 1\right\} \tag{2}$$

is the proposed modified boundary corrected estimator. It is clearly nonnegative because, since K and $a_0(p)$ are nonnegative, $\bar{f}$ is nonnegative, and the rest of the formula is exponentiated. That it is a modification of $\bar{f}$ "in the direction of" $\hat{f}$ is clear, and thus to each $\hat{f}$ there corresponds a nonnegative $\hat{f}_P$. Indeed, there is no requirement here of generalised jackknifing to obtain $\hat{f}$, so the proposal is a completely general nonnegativisation; it might also be used, for instance, with the boundary kernels of Müller (1991). That $\hat{f}_P$ has the properties required of a boundary corrected estimator is verified in Section 3. The $L(u) = uK(u)$ special case of the generalised jackknifing version of this will be utilised in Section 4.

3. Theoretical Performance

The asymptotic means and variances of both $\hat{f}(x)$ and $\hat{f}_P(x)$ are given in the following theorem. By way of notation, write

$$B(p) = \frac{c_1(p)a_2(p) - a_1(p)c_2(p)}{c_1(p)a_0(p) - a_1(p)c_0(p)}$$

and

$$V(p) = \frac{c_1^2(p)b(p) - 2c_1(p)a_1(p)e(p) + a_1^2(p)g(p)}{\{c_1(p)a_0(p) - a_1(p)c_0(p)\}^2},$$

where $b(p)$, $e(p)$ and $g(p)$ are $\int_{-S_K}^{\min\{p,S_K\}} K^2(u)du$, $\int_{-\min(S_K,S_L)}^{\min\{p,S_K,S_L\}} K(u)L(u)du$ and $\int_{-S_L}^{\min\{p,S_L\}} L^2(u)du$, respectively.

Theorem. *Suppose that f has at least two continuous derivatives. Then, as $n \to \infty$, $h = h(n) \to 0$ and $nh \to \infty$,*

$$E\{\hat{f}(x)\} \simeq f(x) + \frac{1}{2}h^2 B(p) f''(x),$$

$$E\{\hat{f}_P(x)\} \simeq f(x) + \frac{1}{2}h^2 \Big\{B(p)f''(x) + \frac{a_1^2(p)}{a_0^2(p)} \frac{f'(x)^2}{f(x)}\Big\}$$

and

$$V\{\hat{f}_E(x)\} \simeq (nh)^{-1} V(p) f(x),$$

where $\hat{f}_E$ denotes either $\hat{f}$, given by (1), *or $\hat{f}_P$, given by* (2).

The results for $\hat{f}$ are taken from Jones (1993), and are presented for comparison with the results for $\hat{f}_P$. The asymptotic variance terms are the same, although in finite samples one might expect the new estimator to have less variance because of nonnegativity. If the terms containing f'' and f' in the bias are, respectively, $B_1(x)$ and $B_2(x)$, then while the additive jackknife has asymptotic bias $B_1(x)$, the multiplicative one has bias $B_1(x) + B_2(x)$. The results for $\hat{f}_P(x)$

in the theorem can be obtained by making the following approximation to (2) using a Taylor expansion:

$$\begin{aligned} f(x)\Big[1+\frac{\{\bar f(x)-f(x)\}}{f(x)}\Big]\exp\Big[1+\frac{\{\hat f(x)-f(x)\}}{f(x)}-\frac{\{\bar f(x)-f(x)\}}{f(x)} \\ -\frac{\{\bar f(x)-f(x)\}}{f(x)}\frac{\{\hat f(x)-f(x)\}}{f(x)}+\frac{\{\bar f(x)-f(x)\}^2}{f^2(x)}\Big] \\ \simeq f(x)\Big[1+\frac{\{\hat f(x)-f(x)\}}{f(x)}+\frac{1}{2}\frac{\{\bar f(x)-f(x)\}^2}{f^2(x)}\Big]. \end{aligned}$$

Complete the manipulations by using the mean and variance of $\hat f$ as in the theorem together with

$$E\{\bar f(x)\}\simeq f(x)-h(a_1(p)/a_0(p))f'(x)+(1/2)h^2(a_2(p)/a_0(p))f''(x).$$

Because of the different dependencies of these terms on the underlying density f, it is difficult to compare biases in general. We plotted these bias terms (not shown) for the three representative members of the gamma family used in Section 4. A similar result pertained to each, namely, identical biases from around $p = 0.7$ upwards, the bias of $\hat f_P$ remaining positive while that of $\hat f$ crosses the zero line and that the larger in absolute value at the boundary itself changed from one situation to the other. For the Gamma (2,1) density, the two biases had a generally similar shape throughout, but because $(f'^2/f)(x) \to \infty$ as $x \to 0$, the bias of $\hat f_P$ exploded there. The higher order contact of many densities that have $f(0) = 0$ means that this difficulty is not too widespread. Overall, we believe there is no serious deterioration of the asymptotic bias of $\hat f_P$ when compared to that of $\hat f$.

4. Practical Performance

The purpose of this section is to show, via simulated examples, that nothing seems to be lost, in terms of finite sample performance relative to alternative methods, while nonnegativity is to be gained. We consider samples of size $n = 50$ from each of the Gamma (1,1), Gamma (2,1) and Gamma (3,1) densities. The biweight kernel was employed. The bandwidths used in each of these cases are simply obtained by entering knowledge of the true f into the standard asymptotic mean squared error optimal formula $h^5 = [\{\int_{-S_K}^{S_K} u^2K(u)du\}^2 \int_0^\infty (f'')^2(x)\, dx\ n]^{-1} \int_{-S_K}^{S_K} K^2(u)du$ (e.g. Silverman (1986), Wand and Jones (1995)); we find that $h = 1.07$, 0.89 and 1.30 respectively. Note that this formula takes no account of boundary effects nor, indeed, do any published practical automatic bandwidth selectors, but progress is now being made on remedying this (Cheng (1996)).

We made 10000 replications of the simulation setups above and in Table 1 present integrated squared biases, variances and mean squared errors of $\hat{f}$ and $\hat{f}_P$. These each employ $L(u) = uK(u)$. Overall, squared biases are increased a little by use of $\hat{f}_P$, which is not surprising. There is sometimes, however, a related decrease in variance. The latter serves to actually make $\hat{f}_P$ better in integrated mean squared error terms than $\hat{f}$ for Gamma (2,1) and only a little worse for Gamma (3,1). For the exponential density, Gamma (1,1), large f near zero makes for increased variance of $\hat{f}_P$ and hence a noticeable overall deterioration. By the way, $\hat{f}$ takes negative values in 0.0%, 34.5% and 76.9% of cases in these simulations for the densities in the order of Table 1.

Table 1. Results of simulations comparing negativity-allowing, $\hat{f}$, and nonnegative, $\hat{f}_P$, boundary corrected estimators for samples of size $n = 50$, averaging over 10000 simulations. Here, $L(u) = uK(u)$ and K is the biweight kernel.

Density	Estimator	h	Integrated Squared Bias ($\times 10^{-4}$)	Integrated Variance ($\times 10^{-4}$)	Integrated Mean Squared Error ($\times 10^{-4}$)
Gamma (1,1)	$\hat{f}$	1.07	7.284	104.9	112.2
	$\hat{f}_P$	1.07	9.803	133.9	143.7
Gamma (2,1)	$\hat{f}$	0.89	8.882	42.87	51.75
	$\hat{f}_P$	0.89	12.62	36.78	49.40
Gamma (3,1)	$\hat{f}$	1.30	1.283	11.06	12.34
	$\hat{f}_P$	1.30	2.565	9.999	12.56

We should add that we also did such simulations for the alternative Ls, K' and K''; the latter is not so immediately attractive in respect of its “joining” with interior kernels. Comparisons between additive and multiplicative proposals within each L were much the same. The choice $L(u) = uK(u)$ proved to be best in two of the three situations.

We also made analogues of Figures 1 to 3 of Marron and Ruppert (1994) for estimator $\hat{f}_P$, again using $L(u) = uK(u)$, but these are also not shown to save space. These concern, respectively, eight random samples for each of $n = 500$ observations from parabolic, “uniform squared” and mixture densities.

Using the biweight kernel and bandwidths matched with those used by Marron and Ruppert, we found (i) great similarities with Marron and Ruppert's (1994) transformation and an adjustment due to Rice (1984) for the parabolic density, (ii) great similarities also for the mixture density, but note that we have avoided the problem (disguised in Marron and Ruppert's Fig. 3(b) by truncation) that the Rice adjustment affords negativity near the origin, and (iii) made a much better indication of the pole at the origin than did Rice's method in the U^2 case. However, Marron and Ruppert's Algorithm P, designed specifically to deal with poles of this sort, has the edge if the precise form of the density near zero when there is a pole is of major concern.

5. Motivation for (2)

Generalised jackknifing was earlier employed to increase the "order" of kernels in the interior from the probability density's order 2 to order 4 (and more): a kernel of order k has the properties $\int_{-S_K}^{S_K} u^l K(u)du = 0$ if $1 \leq l < k$, $\int_{-S_K}^{S_K} K(u)du = 1$ and $\int_{-S_K}^{S_K} u^k K(u)du \neq 0, \infty$. Schucany and Sommers (1977) initiated this; Jones and Foster (1993) greatly developed it. If $\bar{f}$ and $\tilde{f}$ use K and L as kernels, this, again, means utilise an appropriate linear combination, $A_x\bar{f}(x) + B_x\tilde{f}(x)$. Higher order kernels also suffer from negativity problems. It was Terrell and Scott (1980) who suggested using generalised jackknifing on log estimates to alleviate this. This is equivalent to using a multiplicative combination of the form $\bar{f}^{A_x}\tilde{f}^{B_x}$. Nonnegativity is restored at the expense of a minor deviation from unity integral, a property which does hold for higher order kernels.

Rice (1984) extended Schucany and Sommers (1977) to the boundary problem; Jones (1993) noticed that the same methodology applies quite generally to boundary kernel derivation. The natural nonnegative extension of this work via Terrell and Scott (1980) does not always work, however! The problem, which does not rear its head for L functions suitable for obtaining higher order kernels, is that for the kinds of L we prefer to work with here, e.g. $uK(u)$, $K'(u)$, $\tilde{f}$ is negative in places. That said, nonnegative Ls could be used in such a formula. However, in the Schucany and Sommers (1977) and Rice (1984) special case that $L(u) = K_c(u)$, Jones and Foster (1993) observed that, when logs were taken, the limiting case as $c \to 1$ was not of the multiplicative form above but rather of the exponential form (2). The observation here is that this formulation continues to work to good effect for boundary kernels too.

Referees have pointed out that there are other potential alternatives for nonnegative boundary–corrected estimates. The projection technique of Gajek (1986) proposed to nonnegativise higher order kernels could be adapted to boundaries. This consists essentially of iterating between truncation of the estimate where it goes negative and adding/subtracting a suitable constant so that the integral becomes one. Our method wins on explicitness, and hence simplicity, grounds. Certain semiparametric density estimates in which parametric models are fitted locally, such as Loader (1996), can also be nonnegative and have good boundary properties (Hjort and Jones (1996)). Another version of this is to apply semiparametric regression methodology as in Fan, Heckman and Wand (1995) to histogram counts. These ideas are interesting, although linked in with switching, perhaps appealingly, to semiparametric density estimation throughout. It is also reasonable to consider spline–based density estimation approaches on a bounded support in which the boundaries can be quite naturally accommodated as constraints. Complicated computational algorithms are a drawback, as is the lack of straightforward theory for such estimators.

Finally, write

$$\hat{f}_g(x) \equiv \bar{f}(x) g\Big\{\frac{\hat{f}(x)}{\bar{f}(x)} - 1\Big\} \tag{3}$$

as a generalisation of (2). The case $g(z) = \exp(z)$ that we have used seems to be the most natural. But in fact other g's will work too: what is required is that $g(\epsilon) \sim 1 + \epsilon$ for small ϵ and that $g(z) \geq 0, \forall z$. A family of examples is $g(z) = (1 + k^{-1}z)^k$ for even k. A referee has also noted how (2) — and hence also (3) — can be used with appropriate pairs $\{\bar{f}, \hat{f}\}$ to reduce bias in more general situations: the other obvious application of this, to nonnegativising higher order kernels, is already mentioned in Jones and Foster (1993).

References

Cheng, M. Y. (1996). A bandwidth selector for local linear density estimators. To appear.

Fan, J., Heckman, N. E. and Wand, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *J. Amer. Statist. Assoc.* **90**, 141-150.

Gajek, L. (1986). On improving density estimators which are not bona fide functions. *Ann. Statist.* **14**, 1612-1618.

Hjort, N. L. and Jones, M. C. (1996). Locally parametric nonparametric density estimation. *Ann. Statist.* **24**, to appear.

Jones, M. C. (1993). Simple boundary correction for kernel density estimation. *Statist. Comput.* **3**, 135-146.

Jones, M. C. and Foster, P. J. (1993). Generalized jackknifing and higher order kernels. *J. Nonparametric Statist.* **3**, 81-94.

Loader, C. R. (1996). Local likelihood density estimation. *Ann. Statist.* **24**, to appear.

Marron, J. S. and Ruppert, D. (1994). Transformations to reduce boundary bias in kernel density estimation. *J. Roy. Statist. Soc. Ser.B* **56**, 653-671.

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

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

Schucany, W. R., Gray, H. L. and Owen, D. B. (1971). On bias reduction in estimation. *J. Amer. Statist. Assoc.* **66**, 524-533.

Schucany, W. R. and Sommers, J. P. (1977). Improvement of kernel type density estimators. *J. Amer. Statist. Assoc.* **72**, 420-423.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization.* John Wiley, New York.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London, New York.

Terrell, G. R. and Scott, D. W. (1980). On improving convergence rates for nonnegative kernel density estimators. *Ann. Statist.* **8**, 1160-1163.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing.* Chapman and Hall, London, New York.

Department of Statistics, The Open University, Walton Hall, Milton Keynes MK7 6AA, United Kingdom.

Statistical Laboratory, The University of Manchester, Manchester M13 9PL, United Kingdom.

(Received February 1995; accepted December 1995)