Improved Bandwidth Selection for Boundary Correction using the Generalized Reflection Method*

Joris Pinkse† and Karl Schurter‡

Department of Economics

Penn State

November 15, 2019

Abstract

We show that the main results in papers proposing commonly used and attractive boundary correction methods are incorrect. Indeed, we show that the theorems are false and that the problem can be addressed by assuming more smoothness and changing the recommendation of how to choose a secondary input parameter.

1 Introduction

In a series of papers, Zhang et al. (1999, ZKJ), Karunamuni and Alberts (2005, KA), and Karunamuni and Zhang (2008, KZ) propose some attractive boundary correction methods for nonparametric kernel density estimators. These results have been used by a substantial number of authors. For example, KZ has been adopted for the estimation of auction models in economics by Hickman and Hubbard (2015), which has gained some popularity.

*We thank Sung Jae Jun for valuable comments and Ana Enriquez for guidance on the fair use doctrine.

†joris@psu.edu

‡kes380@psu.edu

Unfortunately, the main results in KA and KZ are false for reasons that we point out in this paper. Indeed, no estimator can achieve the claimed results under the stated conditions. We focus in our discussion on KZ, noting that a similar critique applies to KA.

The source of the problem lies in the fact that the KZ methodology requires an auxiliary estimate of the derivative of the log density at the boundary point. This estimator does not converge at the rate stated in the paper for reasons we explain in detail in section 2. The problem is that the variance decreases at a slower rate than what is claimed.

We show that all is not lost, however: the procedure can be fixed. Doing so requires more smoothness (one extra derivative at the boundary) and a different recommendation for the choice of the bandwidth h_1 for the auxiliary estimate. Whereas KZ require the auxiliary bandwidth to converge faster than the main bandwidth used in the paper, one should in fact make the auxiliary bandwidth vanish at a slower rate.

The reason for this is simple: having an extra derivative makes the bias of the auxiliary estimate vanish at the rate h_1^2 instead of h_1. This allows h_1 to converge more slowly yet the bias to decrease faster than would be the case without the extra smoothness. Having h_1 converge more slowly also speeds up convergence of the variance.

The fix carries over to papers that use the KZ methodology, including Hickman and Hubbard (2015). There, also, one should assume an extra derivative and pick the auxiliary bandwidth to converge more slowly than $n^{-1/5}$ but faster than $n^{-1/10}$, ideally at a rate $n^{-1/7}$.

We provide a detailed description of the problem in section 2 and propose a fix in section 3.

2 Problem

Given independent and identically distributed random variables $X_1, \ldots, X_n$ with unknown density f, KZ's estimator is defined by

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} \left\{ K\left(\frac{x - X_i}{h}\right) + K\left(\frac{x - \hat{g}_n(X_i)}{h}\right) \right\} \tag{KZ-2.8}$$

$$\hat{g}_n(y) = y + \hat{d}_n y^2 + A\hat{d}_n^2 y^3$$

$$\hat{d}_n = \frac{\log f_n(h_1) - \log f_n(0)}{h_1}$$

$$f_n(h_1) = f_n^*(h_1) + \frac{1}{n^2}$$
$$f_n(0) = \max\left\{ f_n^*(0), \frac{1}{n^2} \right\}$$
$$f_n^*(h_1) = \frac{1}{nh_1} \sum_{i=1}^{n} K\left(\frac{h_1 - X_i}{h_1}\right)$$
$$f_n^*(0) = \frac{1}{nh_0} \sum_{i=1}^{n} K_{(0)}\left(\frac{-X_i}{h_0}\right)$$

where $A > 1/3$ and the bandwidths (h, h_1, h_0) are chosen by the researcher, K is a second–order kernel, and $K_{(0)}$ is a boundary kernel. We quote the statement of their main result regarding the asymptotic distribution of the above estimator.

Theorem (Theorem 2.1 in Karunamuni and Zhang (2008))**.** *Let $\hat{f}_n$ be defined by (2.8) with $h = O(n^{-1/5})$. Let $h_1 = o(h)$. Assume that $f(x) > 0$ for $x = 0$, h, and that $f^{(2)}$ is continuous in a neighborhood of 0. Then for $x = ch$, $0 \le c \le 1$, we have*[1]

$$E\hat{f}_n(x) - f(x) = \frac{h^2}{2}\left\{ f^{(2)}(0)\int_{-1}^{1} t^2 K(t)\,\mathrm{d}t - 6(A-1)\frac{(f^{(1)}(0))^2}{f(0)}\int_{c}^{1}(t-c)^2 K(t)\,\mathrm{d}t \right\} + o\left(h^2\right) \tag{KZ-2.10}$$

and

$$Var\hat{f}_n(x) = \frac{f(0)}{nh}\left\{ \int_{-1}^{1} K^2(t)\,\mathrm{d}t + 2\int_{-1}^{c} K(t)K(2c-t)\,\mathrm{d}t \right\} + o\left((nh)^{-1}\right) \tag{KZ-2.11}$$

The asserted contribution of KZ's theorem 2.1 is that the remainder terms in (2.10) and (2.11) are little o instead of big O as they were in earlier papers. This is purportedly achieved by choosing a bandwidth h_1 for estimation of the derivative of the log density at the boundary for which $h_1 = o(h)$, where $h \sim n^{-1/5}$ is the main bandwidth used in the paper with n the sample size.

The density function f is assumed to be twice differentiable, so its derivative f' is once differentiable. The optimal nonparametric convergence rate for estimates of $(\log f)'$ is the same as that of estimates of f', namely $\sqrt[5]{n}$ (see e.g. Stone, 1982). This is true since the bias is $O(h_1)$ and the variance $O(1/nh_1^3)$, such that the root mean square error is $O(n^{-1/5})$ if $h_1 \sim n^{-1/5}$. Undersmoothing,

[1]There is likely a typo in the second integral in 2.11 in KZ. Based on ZKJ, we state what we believe to be the intended formula. This discrepancy is immaterial for the point that we make here.

i.e. choosing $h_1 = o(n^{-1/5})$, removes the asymptotic bias, but makes the variance vanish more slowly: the convergence rate is worse.

Lemma A.2. *Let $f_n^*(h_1)$ and $f_n^*(0)$ be defined by* (2.4) and (2.5), *respectively, with $h = h_1$. Suppose that $f^{(2)}$ exists and is continuous near $x = 0$. Then*

$$E[|f_n^*(x) - f(x)|^3 | X_k = x_k, X_\ell = x_\ell] = \mathrm{O}(h_1^6),$$

for any integers $1 \leqslant k, \ell \leqslant n$, and $x = 0, h$.

Proof. Follows from Lemma A.1 of Zhang et al. (1999). □

Lemma A.3. *Let $\widehat{d}_n$ be defined by* (2.3) *with h replaced by h_1, where $h_1 = \mathrm{o}(h)$. Assume that $f(x) > 0$ for $x = 0, h$ and that $f^{(2)}$ exists and is continuous near $x = 0$. Then*

$$E[|\widehat{d}_n - d|^3 | X_k = x_k, X_\ell = x_\ell] = \mathrm{O}(h_1^3),$$

for any integers $1 \leqslant k, \ell \leqslant n$, where $d = f^{(1)}(0)/f(0)$.

Proof. Similar to the proof of Lemma A.2 of Zhang et al. (1999). □

Figure 1: Lemma statements in Karunamuni and Zhang (2008)

We thus agree with (2.10), but (2.11) is false. Indeed, the term that is claimed to be $o\{(nh)^{-1}\}$ in (2.11) would dominate the first right hand side term in (2.11) if $h_1 = o(n^{-1/5})$. To see that this is true, consider lemmas A.2 and A.3 in figure 1, both of which claim a convergence rate faster than the optimal one (if $h_1 = o(n^{-1/5})$) and both of which are false.

Indeed, consider lemma A.2. Its proof is claimed to follow from lemma A.1 of Zhang et al. (1999, ZKJ), which is depicted in figure 2. Lemma A.3 in KZ refers to lemma A.2 of ZKJ, which also depends on lemma A.1 in ZKJ.

The source of the problems in KZ and KA is equation (A.8) in ZKJ, which says that

$$\bar{I}_1 = O\left(\frac{1}{n^2h^2}\right) = O(h^8), \quad \text{for} \quad h = O(n^{-1/5}). \tag{ZKJ-A.8}$$

The first equality is unobjectionable. The second equality, however, holds only if h vanishes *no faster than* $n^{-1/5}$.[2] If h converges faster than $n^{-1/5}$ then the second equality does not hold. If h in ZKJ vanishes at a rate faster than $n^{-1/5}$ then the rate in lemma A.1 of ZKJ is $O(h^8 + 1/n^2h^2)$, but *not* $O(h^8)$.

The consequence for KZ is that the convergence rates in lemmas A.2 and A.3 are $O(h_1^6 + 1/n^{3/2}h_1^{3/2})$ and $O(h_1^3 + 1/n^{3/2}h_1^{9/2})$ respectively instead of $O(h_1^6)$ and $O(h_1^3)$, which are the rates

[2] In ZKJ it is assumed that $h \sim n^{-1/5}$ so there the equality holds.

Lemma A.1. Let $f_n^*(h)$ and $f_n^*(0)$ be as defined by (13). suppose that $f^{(2)}(\cdot)$ is continuous near 0. Then

$$E[(f_n^*(x) - f(x))^4 | X_k = x_k, X_l = x_l] = O(h^8) \quad \text{(A.6)}$$

for any integer $1 \leq k, l \leq n; x = 0, h$.

Proof. Without loss of generality, we prove (A.6) only for the case $k = 1, l = 2$, and $x = h$. By the C_r inequality (Loève 1963, p. 157),

$$\begin{aligned}
&E[(f_n^*(h) - f(h))^4 | X_1 = x_1, X_2 = x_2] \\
&= E\{(f_n^*(h) - E[f_n^*(h)|X_1 = x_1, X_2 = x_2]) \\
&\quad + (E[f_n^*(h)|X_1 = x_1, X_2 = x_2] - f(h))\}^4 \\
&\quad |X_1 = x_1, X_2 = x_2\} \\
&\leq C\{E\{(f_n^*(h) - E[f_n^*(h)|X_1 = x_1, X_2 = x_2])^4 \\
&\quad |X_1 = x_1, X_2 = x_2\} \\
&\quad + E\{(E[f_n^*(h)|X_1 = x_1, X_2 = x_2] - f(h))^4 \\
&\quad |X_1 = x_1, X_2 = x_2\}\} \\
&= C(\bar{I}_1 + \bar{I}_2), \qquad \text{(A.7)}
\end{aligned}$$

where C is a constant (which, in different positions, may take different values):

$$\begin{aligned}
\bar{I}_1 &= \frac{1}{n^4h^4} E\left\{\left(\sum_{i=1}^{n} K\left(\frac{h - X_i}{h}\right)\right.\right. \\
&\qquad \left. - \sum_{i=1}^{n} E\left[K\left(\frac{h - X_i}{h}\right) | X_1 = x_1, X_2 = x_2\right]\right)^4 \\
&\qquad \left. |X_1 = x_1, X_2 = x_2\right\} \\
&= \frac{1}{n^4h^4} E\left[\sum_{i=3}^{n}\left(K\left(\frac{h - X_i}{h}\right) - EK\left(\frac{h - X_i}{h}\right)\right)\right]^4 \\
&= \frac{C}{n^4h^4} \sum_{i=3}^{n} E\left[K\left(\frac{h - X_i}{h}\right) - EK\left(\frac{h - X_i}{h}\right)\right]^4 \\
&\quad + \frac{2C}{n^4h^4} \sum_{3 \leq i < j \leq n} E\left[K\left(\frac{h - X_i}{h}\right) - EK\left(\frac{h - X_i}{h}\right)\right]^2 \\
&\quad \times \left[K\left(\frac{h - X_j}{h}\right) - EK\left(\frac{h - X_j}{h}\right)\right]^2.
\end{aligned}$$

Because

$$E\left[K\left(\frac{h - X_i}{h}\right)\right]^k = h \int_{-1}^{1} K(t)^k f((1 - t)h)\, dt = O(h)$$

for $3 \leq i \leq n$ and $1 \leq k \leq 4$, we have

$$\bar{I}_1 = O\left(\frac{1}{n^2h^2}\right) = O(h^8), \quad \text{for} \quad h = O(n^{-1/5}). \quad \text{(A.8)}$$

Similarly, we can prove

$$\bar{I}_2 = O(h^8). \quad \text{(A.9)}$$

(A.6) is now proved by combining (A.7), (A.8), and (A.9).

Figure 2: Lemma in Zhang et al. (1999)

asserted in KZ. If $h_1 \sim n^{-1/5}$ then lemmas A.2 and A.3 in KZ would be correct, but that rate would not be sufficient to bound I_5 in the proof of KZ's theorem, see figure 3. Indeed, the bound obtained in (A.12) in KZ would be $O(h^6/nh_1^3)$ such that $I_5 = O(h^2/nh_1^3) \neq O(1/nh)$.

3 Fix

A simple way of fixing the problem is to assume f possesses one more derivative at the boundary *and* to let h_1 vanish at a rate *slower* than $n^{-1/5}$, *not* faster (which is the recommendation in KZ). All numerical examples considered in KZ are thrice differentiable, hence they possess the additional derivative.

If f is indeed thrice differentiable at the boundary then f' is twice differentiable, which implies that the bias is $O(h_1^2)$ and the variance $O(1/nh_1^3)$, producing a root mean square error that is $o(n^{-1/5})$ if $nh_1^5 \to \infty$ and $nh_1^{10} \to 0$ as $n \to \infty$. The optimal rate for h_1 is then $n^{-1/7}$, producing a root mean square error of $O(n^{-2/7})$.

References

Hickman, B. R. and Hubbard, T. P. (2015). Replacing sample trimming with boundary correction in nonparametric estimation of first-price auctions. *Journal of Applied Econometrics*, 30(5):739–762.

Karunamuni, R. J. and Alberts, T. (2005). On boundary correction in kernel density estimation. *Statistical Methodology*, 2(3):191–212.

Karunamuni, R. J. and Zhang, S. (2008). Some improvements on a boundary corrected kernel density estimator. *Statistics & Probability Letters*, 78(5):499–507.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of statistics*, pages 1040–1053.

Zhang, S., Karunamuni, R. J., and Jones, M. C. (1999). An improved estimator of the density function at the boundary. *Journal of the American Statistical Association*, 94(448):1231–1240.

Observe that from (A.8),

$$I_2 \leqslant \frac{1}{(nh)^2} E\left\{ \sum_{i=1}^{n} \left[K\left(\frac{x+\widehat{g}_n(X_i)}{h}\right) - K\left(\frac{x+g(X_i)}{h}\right) \right] \right\}^2$$
$$= I_4 + I_5,$$

where,

$$I_4 = \frac{1}{(nh)^2} \sum_{i=1}^{n} E\left[K\left(\frac{x+\widehat{g}_n(X_i)}{h}\right) - K\left(\frac{x+g(X_i)}{h}\right) \right]^2$$

and

$$I_5 = \frac{2}{(nh)^2} \sum_{1\leqslant i<j\leqslant n} E\left[K\left(\frac{x+\widehat{g}_n(X_i)}{h}\right) - K\left(\frac{x+g(X_i)}{h}\right) \right] \left[K\left(\frac{x+\widehat{g}_n(X_j)}{h}\right) - K\left(\frac{x+g(X_j)}{h}\right) \right]. \tag{A.9}$$

By an application of Taylor expansion of order 1 on K we obtain using Lemma A.3 that

$$\begin{aligned}
I_4 &= \frac{1}{(nh)^2} \sum_{i=1}^{n} E\left[\left(\frac{\widehat{g}_n(X_i) - g(X_i)}{h}\right) K^{(1)}\left(\frac{x+(1-\delta)g(X_i)+\delta\widehat{g}_n(X_i)}{h}\right) \right]^2 \\
&\leqslant \frac{C_1}{n^2h^4} \sum_{i=1}^{n} E(\widehat{g}_n(X_i) - g(X_i))^2 [0\leqslant X_i \leqslant ph] \\
&= \frac{C_1}{n^2h^4} \sum_{i=1}^{n} E\left\{ (\widehat{d}_n - d)X_i^2 + (\widehat{d}_n^2 - d^2)X_i^3 \right\}^2 [0\leqslant X_i \leqslant ph] \\
&\leqslant \frac{2C_1}{n^2h^4} \sum_{i=1}^{n} \left\{ E(\widehat{d}_n - d)^2 X_i^4 [0\leqslant X_i \leqslant ph] + E(\widehat{d}_n^2 - d^2)^2 X_i^6 [0\leqslant X_i \leqslant ph] \right\} \\
&\leqslant \frac{2C_1}{n^2} \sum_{i=1}^{n} \left\{ E(\widehat{d}_n - d)^2 [0\leqslant X_i \leqslant ph] + E(\widehat{d}_n^2 - d^2)^2 [0\leqslant X_i \leqslant ph] \right\} \\
&\leqslant \frac{C_2}{n} \{h_1^2 \cdot h + h_1^2 \cdot h\} \\
&= \mathrm{o}((nh)^{-1}),
\end{aligned} \tag{A.10}$$

using an argument similar to obtain (A.5) together with (A.6) and (A.7), where $C_i > 0$ $(i = 1, 2)$ are constants independent of n. A similar argument yields that

$$|I_5| \leqslant \frac{C_3}{n^2h^4} \sum_{1\leqslant i<j\leqslant n} E|\widehat{g}_n(X_i) - g(X_i)||\widehat{g}_n(X_j) - g(X_j)| [0\leqslant X_i \leqslant ph,\ 0\leqslant X_j \leqslant ph], \tag{A.11}$$

where $C_3 > 0$ is a constant independent of n. Again using Lemma A.3, (A.6) and (A.7), we obtain

$$\begin{aligned}
&E|\widehat{g}_n(X_i) - g(X_i)||\widehat{g}_n(X_j) - g(X_j)|[0\leqslant X_i \leqslant ph, 0\leqslant X_j \leqslant ph] \\
&= E|(\widehat{d}_n - d)X_i^2 + A(\widehat{d}_n^2 - d^2)X_i^3||(\widehat{d} - d)X_j^2 + A(\widehat{d}_n^2 - d^2)X_j^3| [0\leqslant X_i \leqslant ph, 0\leqslant X_j \leqslant ph] \\
&\leqslant C_4 E\{h^2 E|\widehat{d}_n - d| + h^3 |\widehat{d}_n^2 - d^2|\}^2 [0\leqslant X_i \leqslant ph,\ 0\leqslant X_j \leqslant ph] \\
&\leqslant 2C_4 h^4 \{E(\widehat{d}_n - d)^2 [0\leqslant X_i \leqslant ph, 0\leqslant X_j \leqslant ph] + E(\widehat{d}_n^2 - d^2)^2 [0\leqslant X_i \leqslant ph,\ 0\leqslant X_j \leqslant ph]\} \\
&\leqslant C_5 h^4 \{h_1^2 E[0\leqslant X_i \leqslant ph, 0\leqslant X_j \leqslant ph]\} \\
&= \mathrm{O}(h^4 \cdot h_1^2 \cdot h^2) \\
&= \mathrm{o}(h^8),
\end{aligned} \tag{A.12}$$

where C_i $(i = 4, 5)$ are positive constants independent of n. Now by combining (A.9)–(A.12), we have $I_2 = \mathrm{o}((nh)^{-1})$. Similarly, it is easy to show that $I_3 = \mathrm{o}((nh)^{-1})$ using the covariance inequality. This completes the proof of (2.11) and the proof of Theorem 2.1. □

Figure 3: Theorem proof in Karunamuni and Zhang (2008)