Statistica Sinica **10**(2000), 1345-1367

BOUNDARY ADJUSTED DENSITY ESTIMATION AND BANDWIDTH SELECTION

Shean-Tsong Chiu

Colorado State University

Abstract: This paper studies boundary effects of the kernel density estimation and proposes some remedies to the problems. Since the kernel estimate is designed for estimating a smooth density, it introduces a large bias near the boundaries where the density is discontinuous. Bandwidth selectors developed for the kernel estimate that select a small bandwidth to reduce the bias can dramatically increase the variation and roughness of the density estimate. In this paper, several boundary adjusted procedures for estimating the density, as well as selecting the bandwidth, are introduced. The proposed procedures greatly reduce the boundary effects and is shown that these density estimates have the same optimal convergence rate as that of the kernel density estimate of a smooth density. Some asymptotic results about the boundary adjusted procedures are provided. Simulation studies were carried out to check the empiric performance of the proposed procedures compared to some existing boundary-corrected estimation procedures. In general, simulation results indicate that for moderate to large sample sizes, the proposed procedures reduce the boundary effects substantially, and are better than comparable existing methods. As an example, we estimate a relevant density connected with some coal-mining disaster data.

Key words and phrases: Bandwidth selection, boundary effects, characteristic function, cross-validation, kernel density estimation.

1. Introduction

Given a random sample $X_1, \ldots, X_n$ from a distribution with density f, the most commonly used nonparametric density estimation method is the kernel estimate

$$\hat{f}_\beta(x) = (n\beta)^{-1} \sum_{j=1}^{n} w\{(x - X_j)/\beta\} \tag{1.1}$$

(Rosenblatt (1956)), where w is a symmetric probability density function, and β is the bandwidth. In practice, a critical step in density estimation is the selection of β, which controls the smoothness of the density estimate.

The problem of bandwidth selection has been studied extensively, and several data-driven methods have been proposed. Most of the research assumes that f is a smooth function over the whole real line. Relatively little is known about the case when f is discontinuous at the boundary of its support. It is well known, see Gasser and Müller (1979), Rice (1984), Gasser, Müller and Mammitzsch (1985) and Schuster (1985), that boundary effects cause many difficulties when procedures designed for estimating smooth densities are applied to densities with discontinuities. In Section 2, we give a brief background.

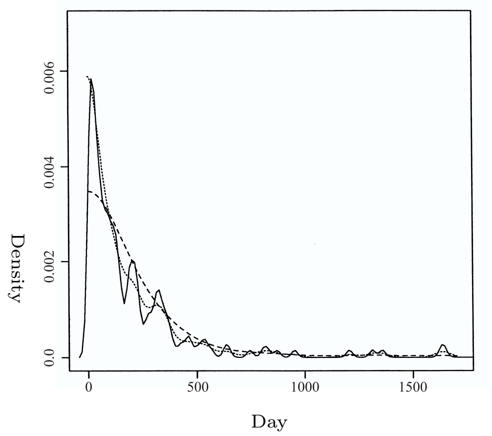

Figure 1. The density estimates of the coal mining disaster data set. The solid curve is the classical kernel density with the bandwidth 14.19; the dotted and dashed curves are the symmetrized estimates with the bandwidths, respectively, 35.64 and 130.79; the kernel is the normal density with bandwidth the standard deviation.

Figure 1 illustrates the problems of applying the kernel estimate (1.1) to a sample from a density discontinuous at the boundary. We use the Gaussian kernel with standard deviation as bandwidth. The data are the periods, in days, between consecutive major disasters of Britain coal mines from 1851 to 1962. More details about the data are given in Section 6. The solid curve is the density estimate (1.1) with bandwidth 14.19, selected by the stabilized selector of Chiu (1991, 1992). The estimate is very rough, and the bandwidth is apparently too small to smooth the density. Using a larger bandwidth introduces a large bias at or near the origin, yet a smaller bandwidth dramatically increases the variation.

Several approaches for handling the boundary effects, for either nonparametric regression or density estimation, have been proposed. Gasser and Müller (1979), Rice (1984) and Gasser *et al.* (1985) consider using boundary kernels near the boundaries. Hall and Wehrly (1991) and Cowling and Hall (1996) propose creating some pseudo data by reflecting the data around the boundaries; the procedures for estimating smooth functions are then applied to the enhanced data to select the bandwidth and estimate the density or the regression function. The main problem with these approaches is that the kernels used near the boundaries are different from the one used in the interior; thus the simple structure of the estimate (1.1) is lost. As a consequence, these approaches cannot easily apply the better bandwidth selection procedures, and have to rely on the less reliable cross-validation method.

Schuster (1985) suggests creating a mirror image of the data on the other side of the boundary, and then applying the estimate (1.1) to the data set and its reflection. As pointed out by Cowling and Hall (1996), this does not remove the asymptotic bias caused by the discontinuity in the first derivative. Another approach is based on semi-parametric models (cf. Eubank and Speckman (1990)), in which a parametric function is first fit to the data, and a nonparametric method is then applied to the residuals. The main problem here is that fitting a global parametric model may not reduce the boundary effects enough.

The difficulties caused by the boundary effects can be explained easily from the perspective of time series analysis. The estimate (1.1) is basically a low-pass filter, which passes the low frequency components and suppresses the noise at the high frequencies. For smooth densities, most information about f concentrates at the low frequencies and the estimate (1.1) can efficiently pass the information without passing too much noise. However, when f has discontinuities, either at the boundary or other places, the high frequency components still contain substantial information about f (particularly about the discontinuities). Thus a low-pass filter will either suppress much of the information or pass too much noise in order to pass most of the information. The basic idea of the proposed procedures is to use the higher frequency components to estimate the boundary effects and then adjust the boundary effects by subtracting or adding a function to the empiric cumulative distribution function before applying the procedures designed for estimating smooth densities.

On the surface, our proposed procedures may look like a hybrid of the reflection method of Schuster (1995) and the semi-parametric approach. However, there is an important difference. We do not attempt to fit, even roughly, a parametric model to the data; instead, the sole goal is to reduce the boundary effects on the high frequency components. Actually, our approaches is in the spirit of the well-known prewhitening in spectrum density estimation, as suggested by Blackman and Tukey (1959). We should remark that while it is not necessary to use a function resembling the true density to adjust the boundary effects, using one with a similar behavior at the boundary is usually beneficial.

In Section 3, we provide a detailed description of the boundary effects on the density estimate (1.1). The optimal bandwidth for the estimate is of the order $n^{-1/3}$, which is extremely small. The impact from first order discontinuities at the boundaries can be removed easily by using a symmetrized density estimate, which is equivalent to using only the real part of the sample characteristic function. This simple step reduces the bias dramatically, and the optimal bandwidth for the estimate is of the order $n^{-1/4}$ when f' is discontinuous at the boundary. To go further, we suggest some estimates of f' at the boundary in Section 4. For the boundary adjusted density estimate, the optimal bandwidth is of the order $n^{-1/5}$, which is the order of the optimal bandwidth for estimating smooth densities. We also briefly discuss a case when f is continuous at the boundary but f' is not. For this case, the imaginary part can be used, and it is not necessary to estimate f' to obtain the optimal convergence rate. Using the imaginary part only is equivalent to the negative-reflection method of Silverman (1986).

The proposed boundary adjusted procedures were applied to some simulated data. For comparison, we also applied the procedure of Cowling and Hall (1996), and the boundary adjusted kernel of Rice (1984) and Gasser *et. al* (1985). Simulation studies and results are summarized in Section 5. The simulation results clearly show the practical advantages of the proposed procedures. Section 6 provides a detailed description of the application to the coal-mining disasters data.

2. MISE and Bandwidth Selection for Smooth Densities

In this section, we provide a brief review of the density estimate, the mean integrated squared error, the optimal bandwidth and bandwidth selection for estimating smooth densities.

A commonly used measure of the performance of a density estimate $\hat{f}_\beta$ is the mean integrated squared error, $\mathrm{MISE}_n(\beta) = E\{\mathrm{ISE}_n(\beta)\}$, where $\mathrm{ISE}_n(\beta) = \int\{\hat{f}_\beta(x) - f(x)\}^2 dx$. Unless indicated otherwise, the integration is over the real line. Let $\phi(\lambda) = \int \exp(i\lambda x) f(x)dx$ be the characteristic function of f, and $\tilde{\phi}(\lambda) = \frac{1}{n}\sum \exp(i\lambda X_j) = \int \exp(i\lambda x)dF_n(x)$ be the sample characteristic function, where F_n is the empiric cumulative distribution function. In the following discussion, we borrow the terminology "frequency" for λ from time series analysis. The decay rate of $|\phi(\lambda)|$ depends on the smoothness of f. For a smooth f, $|\phi(\lambda)|$ decays quickly, and the information about f concentrates at the low frequencies. Also, $\mathrm{Var}\{\tilde{\phi}(\lambda)\} \approx 1/n$ at high frequencies.

Defining $W(\lambda) = \int w(x)\exp(i\lambda x)dx$, the characteristic function of the estimate $\hat{f}_\beta$ at (1.1) is

$$\hat{\phi}_\beta(\lambda) = \tilde{\phi}(\lambda)W(\beta\lambda). \tag{2.1}$$

The estimate $\hat{f}_\beta$ can also be obtained as the Fourier transform: $\hat{f}_\beta(x) = (2\pi)^{-1} \int \hat{\phi}_\beta(\lambda)\exp(-i\lambda x)d\lambda$. This and (2.1) show that the kernel estimate can be viewed as a weighted average of sinusoid waves with the coefficients determined by the sample characteristic function with weight concentrating at the low frequencies.

By applying Parseval's formula, $\mathrm{ISE}_n(\beta)$ can be written as

$$\mathrm{ISE}_n(\beta) = \int \{\hat{f}_\beta(x) - f(x)\}^2 dx = \frac{1}{2\pi} \int |\phi(\lambda) - \tilde{\phi}(\lambda) W(\beta\lambda)|^2 d\lambda. \tag{2.2}$$

Letting $\tilde{\phi}_d(\lambda) = \tilde{\phi}(\lambda) - \phi(\lambda)$ denote the noise part of $\tilde{\phi}$, we expand (2.2) to get

$$\begin{aligned}\mathrm{ISE}_n(\beta) = &\frac{1}{2\pi} \int |\phi(\lambda)|^2 \{1 - W(\beta\lambda)\}^2 d\lambda + \frac{1}{2\pi} \int |\tilde{\phi}_d(\lambda)|^2 W^2(\beta\lambda) d\lambda \\ &- \frac{2}{2\pi} \int \phi(\lambda) \tilde{\phi}_d(-\lambda) W(\beta\lambda) \{1 - W(\beta\lambda)\} d\lambda. \end{aligned} \tag{2.1}$$

Define $w_2 = w{*}w$ and note that $|\tilde{\phi}_d(\lambda)|^2$ is approximately an exponential random variable with mean $\{1 - |\phi(\lambda)|^2\}/n$. From (2.3),

$$\begin{aligned}\mathrm{MISE}_n(\beta) &= \frac{1}{2\pi} \int |\phi(\lambda)|^2 \{1 - W(\beta\lambda)\}^2 d\lambda + \frac{1}{2\pi n} \int \{1 - |\phi(\lambda)|^2\} W^2(\beta\lambda) d\lambda \\ &\approx \frac{1}{2\pi} \int |\phi(\lambda)|^2 \{1 - W(\beta\lambda)\}^2 d\lambda + \frac{w_2(0)}{n\beta}. \end{aligned} \tag{2.4}$$

Now, since $1 - W(\lambda) \approx 1 - \lambda^2 \int x^2 w(x) dx/2$ as $\lambda \to 0$, (2.4) is approximately equal to, when f'' is of bounded variation,

$$\begin{aligned} &\frac{1}{2\pi} (\beta^4/4) \int x^2 w(x) dx \int |\phi(\lambda)|^2 \lambda^4 d\lambda + \frac{w_2(0)}{n\beta} \\ &= (\beta^4/4) \int \{f''(x)\}^2 dx \int x^2 w(x) dx + \frac{w_2(0)}{n\beta}. \end{aligned} \tag{2.5}$$

Therefore, the optimal $\mathrm{MISE}(\beta)$ is of order $n^{-4/5}$ and the optimal bandwidth, which minimizes $\mathrm{MISE}_n(\beta)$, is of order $n^{-1/5}$. The asymptotic form of $\mathrm{MISE}(\beta)$ is obtained in Hall and Marron (1987) and Scott and Terrell (1987).

Since f or ϕ is unknown, the MISE and the optimal bandwidth have to be estimated from the data. Rudemo (1982) and Bowman (1984) proposed least squares cross-validation, $\mathrm{CV}_n(\beta) = \int \hat{f}_\beta^2(x) - n^{-1} \sum_{j=1}^n \hat{f}_{\beta,j}(X_j)$, where $\hat{f}_{\beta,j}$ is the kernel density estimate with the jth observation deleted from the sample. Silverman (1986, p.62) shows that the cross-validation can be approximately written as

$$\mathrm{CV}(\beta) \approx \frac{1}{2\pi} \int \{|\tilde{\phi}(\lambda)|^2 - 1/n\} \{W^2(\beta\lambda) - 2W(\beta\lambda)\} d\lambda + \frac{w_2(0)}{n\beta}. \tag{2.6}$$

Comparing (2.6) with (2.4), we see that cross-validation uses the first term in (2.6) to estimate the bias term in $\mathrm{MISE}(\beta)$.

Up to a constant shift, CV_n is an unbiased estimate of MISE_n, see Scott and Terrell (1987). The asymptotic properties of the bandwidth estimate are established in Scott and Terrell (1987) and Hall and Marron (1987): the bandwidth estimate is consistent and is asymptotically normal, with a slow convergence rate of $n^{-1/10}$. In the simulation studies of Scott and Terrell (1987) and Chiu (1992), it was found that cross-validation often selects smaller bandwidths. The difficulty with cross-validation, as pointed out in Chiu (1991), is caused by including too many higher frequency components in the estimation of the bias term in MISE.

Several bandwidth selection procedures have been proposed to remedy the difficulty of cross-validation. For example, see Scott and Terrell (1987), Park and Marron (1990), Chiu (1991b, 1992), Hall, Marron and Park (1992), Sheather and Jones (1991), Hall, Sheather, Jones and Marron (1991), and Jones, Marron and Park (1991). Except for the biased cross-validation of Scott and Terrell (1987), these bandwidth estimates have a fast convergence rate. In particular, the bandwidth estimates of Chiu (1991, 1992), Hall, Jones, Sheather and Marron (1991), and Jones, Marron and Park (1991) are root n consistent when f is sufficiently smooth. Chiu (1996) provides a review.

We modify the stabilized criterion of Chiu (1991),

$$\mathrm{S}(\beta) = \frac{1}{2\pi}\int_{-\Lambda}^{\Lambda}\{|\tilde{\phi}(\lambda)|^2 - 1/n\}\{1 - W(\beta\lambda)\}^2 d\lambda + \frac{w_2(0)}{n\beta}. \tag{2.7}$$

The cut-off frequency Λ is used to reduce the variation from the high frequency components which do not contain significant information about f. We select Λ as the minimizer of

$$C_n^{\infty}(\Lambda) = -\int_0^{\Lambda}|\tilde{\phi}(\lambda)|^2 d\lambda + 2.55\Lambda/n. \tag{2.8}$$

This is a modification of the CV_n^{∞} proposed in Chiu (1992). A heavier penalty coefficient 2.55 (instead of 2) is used to reduce the chance of selecting an unnecessarily large Λ. The value 2.55 is chosen because the random walk with increments $-Z_j + 2.55$, where Z_j are independent exp(1) random variables, has a probability 0.9 to have its minimum at zero. The probabilities and critical values are computed by using the discrete arc sin distribution, see Woodroofe (1982) and Feller (1966, 1968) for details.

Under some smoothness conditions, the stablized criterion provides a root n consistent bandwidth estimator. Simulation results confirm that the procedure gives a much better bandwidth estimate and density estimate than does cross-validation criterion.

3. Boundary Effects and Boundary Adjusted Procedures

In deriving the asymptotic MISE (2.5) and the optimal bandwidth, we need a critical assumption that $\int \lambda^4|\phi(\lambda)|d\lambda$ exists. This assumption fails when either f or f' are discontinuous. Compared with the problem of estimating a smoother

f, the optimal bandwidth for the density estimate (1.1) is much smaller and the optimal MISE is much larger. As shown below, since MISE is dominated by the discontinuities, applying the procedures for estimating a smooth density cannot provide a satisfactory result.

To simplify the discussion, we consider the case that f has support $[0, \infty)$. Except for a discontinuities at zero, f is assumed to be a smooth function. The case of finite support can be handled in a similar way and will be addressed in Section 7.

Suppose that on $[0, \infty)$, f and f' are continuous and f'' is of bounded variation. The characteristic function of f can be written, by integration by part, as $\phi(\lambda) = -f(0)/(i\lambda) - f'(0)/\lambda^2 - \phi_2(\lambda)/\lambda^2$, where $\phi_2(\lambda) = \int_0^\infty f''(x)\exp(i\lambda x)dx = O(1/\lambda)$. Therefore $\phi(\lambda)$ is dominated by $f(0)/(i\lambda)$ at high frequencies when $f(0) \neq 0$. Since $\phi(\lambda)$ decays slowly a low-pass filter, such as the kernel estimate (1.1), needs a much smaller bandwidth in order to pass most of the information about f.

The bias term in (2.4) is dominated by $f^2(0)\int\{1 - W(\beta\lambda)\}^2/\lambda^2 d\lambda = O(\beta)$. Thus, the optimal bandwidth is of the order $n^{-1/3}$, which is extremely small, cf. Hall (1981). From this, it is clear that the kernel estimate (1.1) and the bandwidth selectors designed for smooth densities cannot be applied directly.

It is interesting to note that, at the high frequencies, a discontinuity of f at zero affects only the imaginary part of ϕ, while, regardless the value of $f(0)$, the real part of $\phi(\lambda)$ is dominated by $f'(0)/\lambda^2$, which decays much faster than $1/\lambda$. This observation leads us to consider using only the real part of $\tilde{\phi}(\lambda)$ to estimate the density. A concern is whether the real part of ϕ contains all the information about f.

Note that the real part $\phi_r = \text{Re}\{\phi\}$ is the characteristic function of the symmetric density $f(|x|)/2$. Based on the original sample, we can create a sample from the density $f(|x|)/2$ by defining $\tilde{X}_i = Z_i X_i,\ i = 1, \ldots, n$, where Z_i are iid random variables with $\Pr(Z_i = 1) = \Pr(Z_i = -1) = 1/2$. The real part of the sample characteristic function of $\{\tilde{X}_i\}$ is identical to $\tilde{\phi}_r = \text{Re}\{\tilde{\phi}\}$. Since for a symmetric density, the imaginary part of the sample characteristic function contains no information about the density, it is sufficient to use only the real part to estimate the symmetric density.

Accordingly, we modify (2.1) and propose the density estimate $\hat{f}_\beta(|x|)/2 = (2\pi)^{-1}\int \hat{\phi}_\beta(\lambda)\cos(x\lambda)d\lambda$. whose characteristic function has the real part

$$\hat{\phi}_\beta(\lambda) = W(\beta\lambda)\tilde{\phi}_r(\lambda). \tag{3.1}$$

The derivative of the density estimate based on (3.1) is zero at $x = 0$, since $\cos(\lambda x)$ is symmetric about zero. The MISE is obtained by modifying (2.4),

$$\text{MISE}_n(\beta) = \frac{1}{2\pi}\int \phi_r^2(\lambda)\{1 - W(\beta\lambda)\}^2 d\lambda + \frac{w_2(0)}{2n\beta}. \tag{3.2}$$

Note that the denominator in the second term above is $2n$ instead of n, as in (2.4), since $\mathrm{Var}\{\tilde{\phi}_r(\lambda)\} = \{1 - \phi_r^2(\lambda)\}/(2n)$. Also, the MISE of $\hat{f}_\beta$ is twice the MISE in (3.2).

If $f'(0) = 0$, (3.2) has an asymptotic form similar to (2.5) and the optimal bandwidth is of order $n^{-1/5}$. Otherwise, the bias term in (3.2) is dominated by $\{f'(0)\}^2 \int \{1 - W(\beta\lambda)\}^2/\lambda^4 d\lambda = O(\beta^3)$. Thus, the optimal bandwidth for the estimate (3.1) is of order $n^{-1/4}$, larger than $n^{-1/3}$ but smaller than $n^{-1/5}$. Also, (3.2) at the optimal bandwidth is of order $n^{-3/4}$, which is bigger than the rate $n^{-4/5}$ when $f'(0) = 0$.

When $f'(0) \neq 0$, we consider the density estimate whose characteristic function has the real part

$$\hat{\phi}_\beta(\lambda) = W(\beta\lambda)\{\tilde{\phi}_r(\lambda)\} - \psi(\theta, \lambda)\} + \psi(\theta, \lambda), \tag{3.3}$$

where $\psi(\theta, \lambda) = f'(0)/\lambda^2 + O(1/\lambda^3)$ is used to remove the boundary effects from $f'(0)$. The parameter θ is a function of $f'(0)$; the precise relationship depends on the particular form of ψ being used.

Roughly speaking, the estimate (3.3) is similar to applying a kernel smoother to the residuals obtained by subtracting from $dF_n(x)$ a function with the same derivative $f'(0)$ at zero. The function subtracted is added back after the smoothing. In the simulation study and application to the coal-mine data set, we use $\psi(\theta, \lambda) = f'(0)/\{|f'(0)| + \lambda^2\}$. this is, ignoring the sign, the real part of the characteristic function of the exponential distribution with mean $\theta = 1/|f'(0)|^{1/2}$.

The MISE of the estimate $\hat{f}_{\beta,\theta}(|x|)/2$ based on (3.3) is

$$\mathrm{MISE}_n(\beta) = \frac{1}{2\pi} \int \{\phi_r(\lambda) - \psi(\theta, \lambda)\}^2 \{1 - W(\beta\lambda)\}^2 d\lambda + \frac{w_2(0)}{2n\beta}. \tag{3.4}$$

For (3.4), since $\phi_r(\lambda) - \psi(\theta, \lambda) = O(1/\lambda^3)$, $\int \lambda^4 \{\phi_r(\lambda) - \psi(\theta, \lambda)\}^2 d\lambda$ exists, and the optimal bandwidth minimizing (3.4) is of order $n^{-1/5}$; the optimal MISE (3.4) is of order $n^{-4/5}$.

In practice, $f'(0)$ is unknown and $\psi(\theta, \lambda)$ in (3.3) is replaced by $\psi(\hat{\theta}, \lambda) \approx \hat{f}'(0)/\lambda^2$, where $\hat{f}'(0)$ is an estimate of $f'(0)$. We then modify the stabilized criterion (2.7) and use

$$\mathrm{S}(\beta) = \frac{1}{2\pi} \int_{-\Lambda}^{\Lambda} [\{\tilde{\phi}_r(\lambda) - \psi(\hat{\theta}, \lambda)\}^2 - 1/(2n)]\{1 - W(\beta\lambda)\}^2 d\lambda + \frac{w_2(0)}{2n\beta} \tag{3.5}$$

to estimate (3.4). The cut-off frequency used in (3.5) is the minimizer of

$$C(\Lambda) = \int_0^{\Lambda} \min\{\tilde{\phi}_r(\lambda) - \psi(\hat{\theta}, \lambda), 1\}^2 d\lambda + \frac{3.23\Lambda}{2n}. \tag{3.6}$$

Note that the penalty coefficient is 3.23 instead of 2.55 as in (2.8). The reason is that $\tilde{\phi}_r^2(\lambda)$ is approximately a χ_1^2 instead of an exponential random variable.

Theorem 1 describes some asymptotic properties of the procedure. The assumptions of the theorem are stated below, and the proofs are given in Section 8.

Assumption 1. Assume $f^{(k)}$ exists on $[0,\infty)$ for $k = 1,\dots,2l+1$, and that $f^{(2l+1)}$ is of bounded variation with $f^{(2l+1)}(0) \neq 0$.

Assumption 2. Let $g(\theta,x)$, $x \geq 0$, be the function defined by $\psi(\theta,\lambda) = \mathrm{Re}\{\int g(\theta,x)\exp(i\lambda x)dx\}$. Assume $g(x)$ satisfies the conditions in Assumption 1 and that $f^{(2k+1)}(0) = g^{(2k+1)}(\theta_0,0)$ for $k = 0,\dots,l-1$.

Assumption 3. For any θ_1 and θ_2, $|\psi(\theta_1,\lambda) - \psi(\theta_2,\lambda)| \leq M\{(\theta_1-\theta_2)/\lambda^2$ for some constant $M > 0$.

Theorem 1. *Suppose that $f(x)$ and $\psi(\theta,\lambda)$ satisfy Assumptions* 1 *to* 3 *with $l \geq 1$, and that $\hat{f}'(0) - f'(0) = O(n^{-2/7})$. Then $n^{1/5}(\hat{\beta} - \beta_{0n}) = O\{\hat{f}'(0) - f'(0)\}$, where β_{0n} is the bandwidth minimizing* (3.4) *and $\hat{\beta}$ is the bandwidth estimate minimizing* (3.5)*.*

In the next section we propose an estimate of $f'(0)$ which satisfies the requirement of Theorem 1. We note that, in general, the procedure does not give a root n consistent bandwidth estimate.

It may not be clear why an $O(n^{-2/7})$ estimate of $f'(0)$ is sufficient for removing the boundary effects from MISE. This is due to the fact that reflecting already removes the effects caused by the discontinuity of f; thus we are making a secondary adjustment. The effects of the discontinuity of f' on MISE is approximately proportional to $\beta^3 f'(0)$. The residual effect after the adjustment is therefore of order $n^{-3/5}n^{-2/7}$, which is smaller than the order $n^{-4/5}$ of the optimal MISE

The discussion above concentrates on the case $f(0) \neq 0$. If $f(0) = 0$, the problem caused by the discontinuity of f' at zero can be handled more easily. In this case $\mathrm{Im}\{\phi(\lambda)\} = O(1/\lambda^3)$ and, similar to the consideration leading to (3.1), we can use only the imaginary part of $\tilde{\phi}$ to estimate the density. More specifically, the imaginary part of the characteristic function of the estimate is $W(\beta\lambda)\mathrm{Im}\{\tilde{\phi}(\lambda)\}$. This estimate is equivalent to the kernel estimate applied to $\mathrm{sign}(x)dF_n(|x|)/2$, an estimate of the function $g(x) = \mathrm{sign}(x)f(|x|)/2$ (which is not a density anymore). Since g' is continuous at zero, it is not necessary to remove the boundary effects to achieve the rate $n^{-1/5}$ and $n^{-4/5}$ for, respectively, the optimal bandwidth and the optimal MISE. The optimal bandwidth can be estimated by modifying (2.7) and (2.8) accordingly. Asymptotic results for the bandwidth estimate can be obtained by following the arguments in Chiu (1991).

One practical issue regarding the procedures based on $\mathrm{Im}\tilde{\phi}(\lambda)$ is the assumption $f(0) = 0$. This condition can be checked by comparing the graphs of $\tilde{\phi}_r^2(\lambda)$ with $\tilde{\phi}_i^2(\lambda)$. If $f(0)$ is substantially different from 0, $\tilde{\phi}_i^2$ should be much larger than $\tilde{\phi}_r^2$ at high frequencies. Otherwise, the assumption that $f(0) = 0$ is a reasonable one. Using the imaginary part only is equivalent to the negative reflection estimate discussed in Silverman (1986, p.31).

4. Estimation of the Derivative at the Boundary

In this section, we propose procedures for estimating $f'(0)$, which can be used in the density estimate (3.3) and the bandwidth selector (3.5). As noted before, $\phi_r(\lambda) \approx f'(0)/\lambda^2$ at high frequencies. Thus $\lambda^2\phi_r(\lambda) \approx f'(0)$ and $\lambda^2\tilde{\phi}(\lambda) \approx f'(0)+\lambda^2\mathrm{Re}\{\tilde{\phi}_d(\lambda)\}$. This observation suggests the following estimate, a weighted average of $\lambda^2\tilde{\phi}_r(\lambda)$.

$$\hat{f}'(0) = 3\Lambda^3 \int_\Lambda^\infty \tilde{\phi}_r(\lambda)/\lambda^2 d\lambda. \tag{4.1}$$

It is necessary to exclude lower frequencies to avoid having a large bias. The convergence rate of $\hat{f}'(0)$ is given in Theorem 2.

Theorem 2. *Suppose that $f^{(2l+1)}$ is of bounded variation on $[0, \infty)$, and $f^{(2k+1)}(0) = 0$ for $0 < k < l$, with $f^{(2l+1)}(0) \neq 0$. Assume that $\Lambda \to \infty$ as $n \to \infty$. Then the variance of $\hat{f}'(0)$ is of the order Λ^3/n and the bias is of the order Λ^{-2l}.*

Theorem 2 provides some hints to the practical issue on selecting the cut-off frequency. The best convergence rate of $\hat{f}'(0)$ depends on the smoothness of f. Under the assumptions of the theorem, the optimal convergence rate is $n^{-2l/(4l+3)}$, obtained by setting $\Lambda = n^{1/(4l+3)}$. Ideally, Λ could be selected to minimize the mean square error of $\hat{f}'(0)$. But, since $\phi_2(\lambda)$ is negligible at high frequencies when $f'(0) \neq 0$, it would be a difficult task to estimate the bias. Practically, we might be able to select Λ visually from the log-log plot of $\tilde{\phi}_r^2(\lambda)$. In any case, it is useful to have some automatic selection procedures.

We now discuss some methods of selecting Λ for $\hat{f}'(0)$ at (4.1). These methods are subsequently applied to simulated data and the coal-mining disaster data in Sections 5 and 6. It is natural to consider the criterion $C_1(\Lambda) = -\int_0^\Lambda \tilde{\phi}_r^2(\lambda) + 3.23\Lambda/(2n)$. When $f'(0) \neq 0$, $O(n^{1/4}) = \Lambda = o(n^{1/4+\delta})$ for some constant $\delta > 0$. Thus the convergence rate of $\hat{f}'(0)$ is approximately of order $n^{-1/8}$. On the other hand if $f'(0) = 0$, $\Lambda \approx O(n^{1/(4l)})$ and the convergence rate of $\hat{f}'(0)$ is approximately of order $n^{-1+3/(4l)}$. These results indicate that while $\hat{f}'(0)$, based on Λ selected by C_1, is a good estimate when $f'(0)$ is small, it has a large variation when $f'(0)$ is big. We see from this that when $f'(0)$ has a substantial effect, a much smaller Λ should be used than the one provided by C_1.

If $f^{(3)}(0) \neq 0$ the optimal convergence rate of $\hat{f}'(0)$ is $n^{-2/7}$, obtained by setting $\Lambda = O(n^{1/7})$. This leads to the consideration of the criterion

$$C_2(\Lambda) = -\int_0^{\Lambda} \tilde{\phi}_r^2(\lambda) + \Lambda(n/2)^{-4/7}. \tag{4.2}$$

For a Λ selected by C_2, the rate is approximately $n^{1/7}$ when $f'(0) \neq 0$ and $n^{1/(7l)}$ when $f'(0) = 0$. In both cases, the convergence rate of $\hat{f}'(0)$ is roughly $n^{-2/7}$, and thus satisfies the condition of Theorem 1 in the previous section.

Although this procedure provides a useful estimate of $f'(0)$ asymptotically, our empiric experience indicates that the Λ selected by C_2 could be too small when ϕ_r^2 has a side lobe. To reduce the problem, we suggest using the criterion

$$C_3(\Lambda) = -\int_0^{\Lambda} \min[1, \{\tilde{\phi}_r(\lambda) - \hat{f}_0'(0)/\lambda^2\}^2] + \Lambda/(n/2)^{-4/5} \tag{4.3}$$

to select Λ, where $\hat{f}_0'(0)$ is the initial estimate based on the Λ selected by C_2. We use this as the estimate of $f'(0)$ in the simulation studies.

5. Implementation and Simulation

To check the performance of the proposed procedures, we applied them to some simulated data sets. The results from several typical cases are summarized in this section. A detailed implementation of the procedures is also provided. For comparison, we also applied the procedures based on boundary kernels and on Cowling and Hall (1996).

For the procedure of Cowling and Hall (1996), the pseudodata $X_{(-j)}$ were created as

$$X_{(-j)} = -5X_{(j/3)} - 4X_{(2j/3)} + (10/3)X_{(j)},\ j = 1, \ldots, n,$$

where $X_{(0)} = 0$ and $X_{(j)}$ are the order statistics. For non-integer $j/3$ or $2j/3$, the values are obtained by linear interpolation. As the authors note, the pseudodata are not necessarily negative. They suggested truncating the pseudodata at the "turning point", which is not defined explicitly. In the following simulation studies, we considered the turning point to be the location of the global minimum. The density estimate was obtained by applying the kernel $w(x) = (3/4)(-x^2 + 1),\ |x| \leq 1$ to the pseudodata set created. The bandwidth is selected by cross-validation, $\int_0^\infty \{\hat{f}(x)\}^2 dx - 2/n \sum_{j=1}^n \hat{f}_{-j}(X_j)$, where $\hat{f}_{-j}(X_j)$ is the leave-one-out estimate at X_j.

For the procedure based on boundary kernels, we applied the kernel $w(x)$ mentioned above to the interior, and the kernel $u_q(x) = (ax + b)w(x), q > -1$, around the boundary, where a, b are constants satisfying $\int_{-q}^1 u_q(x)dx = 1$ and

$\int_{-q}^{1} xu_q(x)dx = 0$. Following Hall and Wherly (1991), the bandwidth at the boundary $x = 0$ is set at 2β, and $2\beta - x$ at $0 < x < \beta$. The bandwidth is selected by cross-validation. For both procedures, the minimum was searched over the interval $(0, 3]$.

Several densities were considered in the simulation studies and we report the results from four of them: (1) the standard half normal density; (2) $N(1, 1)$ truncated at zero; (3) $N(-1, 1)$ truncated at zero; and (4) a mixture of an exponential and a normal density, $0.75\exp(1) + 0.25N(3, 0.8^2)$. These densities are fairly smooth on $(0, \infty)$.

The half normal density is very smooth and $f^{(2k+1)}(0) = 0$ for all k; the curve decays faster than any linear (on the log-log scale) trend. For the truncated normal (2), the plot of $\log|\phi|$ suggests that we would not get an accurate estimate of $f'(0)$ when the sample size is small. In fact, we expect $|\hat{f}'(0)|$ to overestimate $|f'(0)|$. But for the truncated normal (3) and the mixture (4), an accurate estimate of $f'(0)$ can be obtained from moderate samples. Also, since $f'(0)$ has dominant effects for the truncated normal (3) and the mixture (4), we expect the boundary adjusted procedures to make substantial improvement, in terms of ISE, over the non-adjusted estimates.

For each density, 1000 replicates were simulated for each of the sample sizes $n = 100$, 400 and 1600. Each simulated data set is discretized by defining $Y_t = F_n(z_t) - F_n(z_{t-1})$, $t = 1, \ldots, N$, where $z_t = tU/N$, $t = 0, \ldots, N$. In the simulation studies, we set $U = 32$ and $N = 16384 = 2^{14}$. We then applied the fast Fourier transform to Y_t to obtain an approximation of $\tilde{\phi}_r(\lambda_j)$, where $\lambda_j = 2\pi j/U$. Integrations needed for estimates or criteria were obtained from the corresponding summations based on the approximate $\tilde{\phi}_r(\lambda_j)$.

After obtaining an estimate of $f'(0)$ by the procedure described at the end of the previous section, we used $\psi(\hat{\theta}, \lambda) = \hat{f}'(0)/\{|\hat{f}'(0)| + \lambda^2\}$ to reduce the boundary effects on $\tilde{\phi}_r$. The function $\psi(\theta, \lambda)$ for a given $\theta \neq 0$ is, ignoring the sign, the real part of the characteristic function of an exponential density. We chose the exponential density for its simplicity and smoothness (except at zero).

The cut-off frequency Λ was obtained according to (3.6), and then the bandwidth estimate by minimizing (3.5). For an easier comparison, we used the same kernel $w(x) = (3/4)(-x^2 + 1)$, $|x| \leq 1$ to estimate $f(x)$. The minimizer was found by a numeric method over the region $(0, 4]$.

Table 1 summarizes the results concerning the estimates $\hat{f}'(0)$. The numbers in the parenthesis below each of the densities are the true values of $f'(0)$. For Cowling and Hall (1996), the estimate of $f'(0)$ is given by $\hat{f}'(0) = (n\beta^2)^{-1} \sum_j u'(X_j/\beta)$, where the summation is over all available data, including the pseudodata. For the boundary kernel based procedure, since the bandwidth is also a function of x, we estimated $f'(0)$ numerically by $\{\hat{f}(0.01) - \hat{f}(0)\}/0.01$.

From Table 1, we see that our procedure provides a reasonable estimate of $f'(0)$. In general, the proposed estimate is much better than the two it is compared with. Cowling and Hall (1996) does not work well when $f'(0) < 0$; the boundary kernel based estimate has a large variation for all cases.

Table 1. The sample means and sample standard deviations of various estimates of $f'(0)$.

Desity	Estimate	Sample size $n = 100$		$n = 400$		$n = 1600$	
		$f'(0)$	$SD\{f'(0)\}$	$f'(0)$	$SD\{f'(0)\}$	$f'(0)$	$SD\{f'(0)\}$
(1) Half Normal	Proposed	−0.43	0.37	−0.32	0.25	−0.21	0.15
(0.00)	Boundary kernel	−0.22	3.54	−0.11	1.31	−0.14	1.21
	Cowling and Hall	0.41	0.65	0.20	0.61	0.05	0.70
(2) Trunc. Normal	Proposed	0.21	0.26	0.34	0.18	0.42	0.14
(0.29)	Boundary kernel	0.25	1.22	0.27	0.75	0.28	0.46
	Cowling and Hall	0.17	0.22	0.19	0.26	0.20	0.21
(3) Trunc. Normal	Proposed	−1.74	0.80	−1.74	0.49	−1.74	0.27
(−1.53)	Boundary kernel	−1.21	4.50	−1.35	2.24	−1.43	1.36
	Cowling and Hall	1.41	2.41	1.04	1.36	0.75	0.97
(4) Mixture	Proposed	−0.45	0.47	−0.68	0.24	−0.71	0.13
(−0.75)	Boundary kernel	−0.50	1.65	−0.58	1.00	−0.60	0.84
	Cowling and Hall	0.45	0.61	0.36	0.43	0.21	0.90

In Tables 2 and 3, we summarize the sample means of $\hat{\beta}$, the ratios of the standard deviation to the corresponding sample mean of $\hat{\beta}$, and the sample means and standard deviations of ISE($\hat{\beta}$). In addition, we include the results when the true $f'(0)$ was used to adjust the boundary effects.

In general, the boundary adjusted procedures give a much larger bandwidth estimate when the boundary effects are substantial, as for the truncated normal (3) and the mixture (4). In these cases, the average ISE is dramatically reduced for moderate to large sample sizes. Adjusting the boundary effects provides a mild improvement when $f'(0)$ is small, such as for the truncated normal (2). As expected, for the half normal with $f'(0) = 0$, making the adjustment gives a larger bandwidth and increases the average ISE. For the small sample size $n = 100$, the average ISE of the proposed estimate is larger than the non-adjusted one for the cases of the half normal (1) and the mixture (4); the average ISE of the proposed estimate is smaller for the other two cases.

Comparing with the other two procedures we see that, except for one case, the average ISE($\hat{\beta}$) of the proposed estimate is much smaller. The case of the mixture (4) with $n = 100$ is the exception. When $f'(0) < 0$, the procedure of Cowling and Hall (1996) is substantially worse than the others. We note that the standard deviations of the ISE's are quite large (compared with the average

ISE). This is caused by the fact that the distributions of ISE have a heavy right-hand tail. We also note that the standard deviation of the ISE of the proposed estimate is substantially smaller than that of the compared estimates.

Table 2. Summary, for the half normal and the truncated normal density (2), of the sample means of $\hat{\beta}$, the ratios of the standard devation of $\hat{\beta}$ to the corresponding sample mean, and the sample means and standard deviations of $ISE(\hat{\beta})$ for various estimates. Results are based on 1000 replications for each case.

Density	n	Estimate/Adjust	$\hat{\beta}$	$SD(\hat{\beta})/Mean(\hat{\beta})$	$ISE(\hat{\beta})$	$SD(ISE)$
1: Half Normal	100	None, 0	0.711	0.167	8.4e-3	1.2e-2
		Proposed	1.334	0.351	9.3e-3	9.9e-3
		Boundary kernel	1.013	0.424	1.6e-2	2.2e-2
		Cowling and Hall	0.894	0.410	1.3e-2	1.5e-2
	400	None, 0	0.528	0.112	2.6e-3	3.0e-3
		Proposed	0.739	0.261	3.0e-3	2.7e-3
		Boundary kernel	0.707	0.404	4.4e-3	4.4e-3
		Cowling and Hall	0.748	0.392	4.2e-3	4.2e-3
	1600	None, 0	0.397	0.086	9.0e-4	9.4e-4
		Proposed	0.474	0.124	9.7e-4	9.0e-4
		Boundary kernel	0.454	0.383	1.5e-3	1.4e-3
		Cowling and Hall	0.557	0.456	1.6e-3	1.4e-3
2: Trunc. Normal	100	None, 0	0.939	0.304	8.9e-3	6.4e-3
		True $f'(0)$	0.911	0.139	6.2e-3	6.0e-3
		Proposed	1.021	0.238	8.4e-3	7.5e-3
		Boundary kernel	0.902	0.451	1.2e-2	1.2e-2
		Cowling and Hall	1.053	0.436	1.0e-2	1.0e-2
	400	None, 0	0.573	0.186	3.1e-3	2.3e-3
		True $f'(0)$	0.652	0.114	2.3e-3	2.0e-3
		Proposed	0.669	0.105	2.6e-3	2.2e-3
		Boundary kernel	0.616	0.316	3.8e-3	3.6e-3
		Cowling and Hall	0.677	0.332	3.6e-3	3.3e-3
	1600	None, 0	0.406	0.115	1.1e-3	6.3e-4
		True $f'(0)$	0.478	0.070	7.7e-4	5.2e-4
		Proposed	0.485	0.051	9.0e-4	5.7e-4
		Boundary kernel	0.465	0.283	1.2e-3	9.4e-4
		Cowling and Hall	0.494	0.294	1.2e-3	9.5e-4

For most cases, as expected, the relative standard deviation of the proposed bandwidth estimate is much smaller than the other two estimates. The proposed estimate has a large variation when it is difficult to distinguish the density from an exponential density. The results also indicate that the relative convergence rate of the proposed estimate better than that of the other two estimates.

Table 3. Same as Table 2, but for the truncated normal (3) and the mixture (4).

Density	n	Estimate/Adjust	$\hat{\beta}$	$SD(\hat{\beta})/Mean(\hat{\beta})$	$ISE(\hat{\beta})$	$SD(ISE)$
		None, 0	0.411	0.217	1.9e-	2.1e-2
		True $f'(0)$	1.103	0.255	1.0e-	1.7e-2
	100	Proposed	1.377	0.613	1.2e-	1.7e-2
		Boundary kernel	0.650	0.379	2.3e-	3.0e-2
		Cowling and Hall	0.517	0.445	2.8e-	2.4e-2
		None, 0	0.285	0.182	6.7e-	5.3e-3
		True $f'(0)$	0.680	0.262	3.4e-	3.6e-3
3: Trunc. Normal	400	Proposed	0.781	0.440	3.4e-	4.2e-3
		Boundary kernel	0.508	0.334	6.9e-	7.6e-3
		Cowling and Hall	0.456	0.387	1.0e-	7.9e-3
		None, 0	0.198	0.158	2.3e-	1.5e-3
		True $f'(0)$	0.400	0.174	9.7e-	1.1e-3
	1600	Proposed	0.512	0.172	9.7e-	1.1e-3
		Boundary kernel	0.404	0.304	1.9e-	1.6e-3
		Cowling and Hall	0.420	0.316	3.9e-	2.3e-3
		None, 0	0.747	0.335	1.2e-	9.9e-3
		True $f'(0)$	1.594	0.381	5.5e-	7.1e-3
	100	Proposed	2.278	0.438	1.5e-	9.6e-3
		Boundary kernel	1.047	0.449	1.3e-	1.5e-2
		Cowling and Hall	0.870	0.579	1.9e-	1.5e-2
		None, 0	0.4711	0.200	4.5e-	2.5e-3
		True $f'(0)$	0.978	0.139	1.8e-	1.6e-3
4: Mixture	400	Proposed	1.045	0.347	3.0e-	2.5e-3
		Boundary kernel	0.717	0.307	4.3e-	3.8e-3
		Cowling and Hall	0.642	0.404	6.3e-	4.0e-3
		None, 0	0.316	0.165	1.6e-	8.4e-4
		True $f'(0)$	0.679	0.111	6.0e-	4.7e-4
	1600	Proposed	0.670	0.102	7.7e-	5.8e-4
		Boundary kernel	0.542	0.267	1.3e-	1.2e-3
		Cowling and Hall	0.535	0.351	2.3e-	1.6e-3

Next, we make some pointwise comparison of the density estimates. In Table 4, we compare estimates of $f(0)$. For most cases, the MSE of the proposed estimate is substantially smaller than the other two estimates. The variation of the boundary-kernel estimate is often much larger than the others. The estimate of Cowling and Hall (1996) often has the largest bias. To check the MSE in the interior part, we plot pointwise MSE of the density estimates for $n = 400$ in Figure 2.

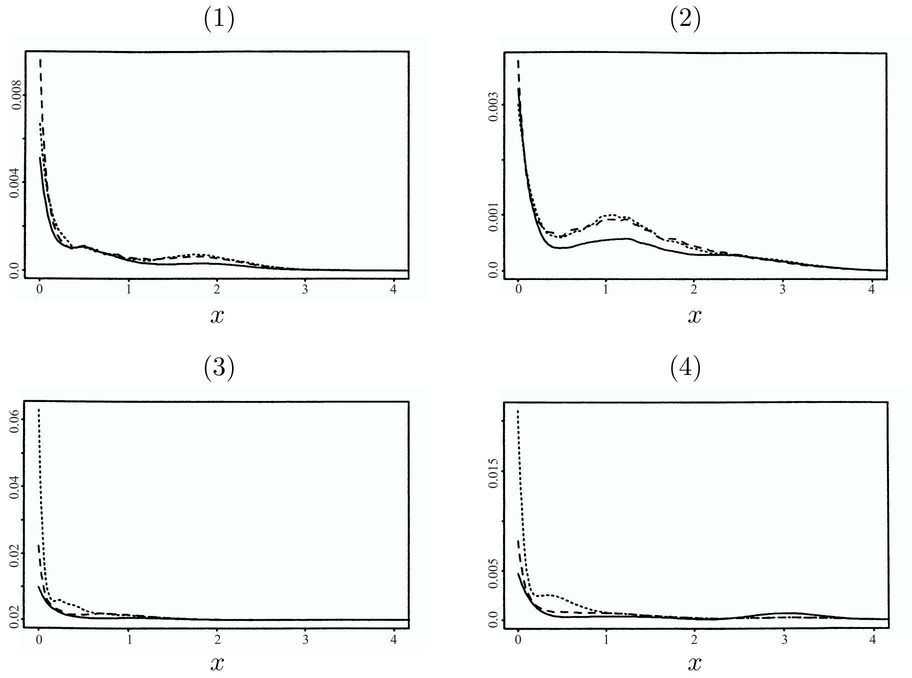

Figure 2. The pointwise mean squared errors of the estimates. The solid curves are of the proposed estimate, the dotted ones the estimate of Cowling and Hall (1996) and the dashed ones the boundary-kernel estimate.

In general, the MSE of the proposed estimate is substantially smaller than the compared estimate at the boundary, as well as at the interior points. The improvement in the interior is due to using the more stable bandwidth selector.

Besides the density estimate with $f'(0)$ based on Λ selected by C_3, we also check the performance of the estimate with Λ selected by C_2. There is not much difference between them for most cases, which indicates that the density estimate is not too sensitive to the selection of Λ.

Just as the procedures for estimating smooth densities should not be applied blindly, the boundary adjusted procedures should be applied with caution. It is vital to examine the plot of $\tilde{\phi}_r^2$. As the simulation results indicated, substantial improvement can be obtained when $f'(0)$ has a dominant effect. In this case, the boundary effects would be visible from the plot of $\tilde{\phi}_r^2$; otherwise, as in some small sample cases, adjustment made by using an inaccurate estimate can introduce more errors to the density estimate.

Table 4. Summary of the biases, standard deviations and the mean squared errors of the various estimates of $f(0)$.

1: Half Normal	100	Proposed	0.037	0.111	0.014
		Boundary kernel	0.017	0.188	0.036
		Cowling and Hall	-0.091	0.096	0.018
	400	Proposed	0.023	0.068	0.005
		Boundary kernel	0.010	0.097	0.010
		Cowling and Hall	-0.040	0.071	0.007
	1600	Proposed	0.013	0.038	0.002
		Boundary kernel	0.008	0.059	0.004
		Cowling and Hall	-0.013	0.052	0.003
2: Trunc. Normal	100	Proposed	0.027	0.094	0.010
		Boundary kernel	0.009	0.106	0.011
		Cowling and Hall	0.040	0.068	0.006
	400	Proposed	-0.003	0.057	0.003
		Boundary kernel	0.000	0.062	0.004
		Cowling and Hall	0.024	0.049	0.003
	1600	Proposed	-0.007	0.035	0.001
		Boundary kernel	0.001	0.035	0.001
		Cowling and Hall	0.016	0.030	0.001
3: Trunc. Normal	100	Proposed	-0.064	0.167	0.032
		Boundary kernel	-0.044	0.290	0.086
		Cowling and Hall	-0.304	0.179	0.124
	400	Proposed	-0.024	0.096	0.010
		Boundary kernel	-0.028	0.147	0.022
		Cowling and Hall	-0.228	0.104	0.063
	1600	Proposed	-0.003	0.052	0.003
		Boundary kernel	-0.015	0.073	0.006
		Cowling and Hall	-0.173	0.064	0.034
4: Mixture	100	Proposed	-0.085	0.126	0.023
		Boundary kernel	-0.034	0.156	0.026
		Cowling and Hall	-0.189	0.098	0.045
	400	Proposed	-0.011	0.066	0.004
		Boundary kernel	-0.017	0.088	0.008
		Cowling and Hall	-0.130	0.063	0.021
	1600	Proposed	-0.007	0.034	0.001
		Boundary kernel	-0.011	0.051	0.003
		Cowling and Hall	-0.090	0.046	0.010

6. Application to the Coal Mining Disaster Data

We applied the boundary adjusted procedures to the Coal Mining Disaster Data from Jarrett (1979). The data set consists of 190 intervals, in days, between consecutive major coal mining disasters in Britain from 1851 to 1962. This data set is a popular example in the change point literature. For example, Raftery and Akman (1986) and Worsley (1986) model the data set as independent exponential random variables with a step mean function.

The range of the data is from 0 to 2366. We replaced the only 0 in the data set by 0.5 to make sure that the data point was counted when we discretized the data. We next divided the data by the sample mean 213.4 and applied the procedures, with the setting $U = 32$, $N = 2^{14}$ and the normal kernel. Since all procedures used are scale equivariant, the bandwidth and density estimate can be easily transformed back to the original scale.

First, we blindly applied the kernel density estimate (1.1) to the data. The bandwidth $0.0665 \cdot 213.4 = 14.19$ is selected by the stabilized procedure (2.7) described in Section 2. The density estimate (1.1) with the bandwidth is shown in Figure 1 as the solid curve. The curve is very rough due to the use of an extremely small bandwidth. Next, we considered the estimates and the bandwidth selection procedures based on $\tilde{\phi}_r$. For the estimate (3.1) based on $\tilde{\phi}_r$, the bandwidth obtained from (3.5) is $0.1670 \cdot 213.4 = 35.64$. The density estimate based on (3.1) with the bandwidth 35.64 is also shown in Figure 1 as the dotted curve. The density is obtained by applying the Fourier transform on (3.1). Although the bandwidth is more than double the bandwidth 14.19 of the estimate (1.1), the density estimate is still rough.

To estimate $f'(0)$ we first obtain $\Lambda = 2.553/(213.4)^2$, selected by (4.2), and the estimate of $\hat{f}'(0)$ is $-2.302/(213.4)^2$. Using this as the initial estimate, we then obtain $\Lambda = 2.749/(213.4)^2$ from (4.3) and the final estimate $\hat{f}'(0) = 2.531/(213.4)^2$ =5.52e-5. Based on this, criterion (3.5) is used to select the bandwidth for the density estimate based on (3.3). The bandwidth selected is $0.6129 \cdot 213.4 = 130.79$, almost 4 times wider than the bandwidth 35.64 selected when no boundary adjustment is made. Figure 3 shows the proposed density estimate (solid line) with the bandwidth 130.79. For comparison, we show the density estimate (3.1) with bandwidth 130.79 as the dashed curve in Figure 1. It is clear that the estimate severely underestimates $f(0)$ and overestimates the density near zero.

The data set may contain some change points. Raftery and Akman (1986) use a step function with one discontinuity at 124 to fit the data. The sample mean for the data points from 1 to 124 is 114.8, and the sample mean for the rest of the data points is 398.6. Under this model, the density is a mixture of two exponential densities. We also plot the parametric density estimate

$$(124/190)\exp(-x/114.8)/114.8 + (66/190)\exp(-x/398.6)/398.6.$$

in Figure 3 to compare (dotted line) with the nonparametric estimate obtained earlier. The two densities are in good agreement.

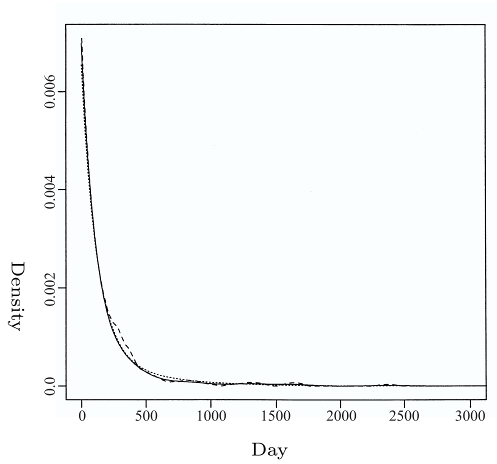

Figure 3. The proposed density estimate (solid curve), the parametric density estimate (dashed curve) and the boundary-kernel estimate of the coal mining disaster data set.

For comparison, we also applied the boundary kernel based procedure and Cowling and Hall (1996) to the same data set. The bandwidths are, respectively, 115 and 55 and the estimated $f'(0)$ are -6.62e-5 and 3.56e-5. Adjusting the bandwidths 115 and 55 by the factor 1.6, the asymptotically equivalent bandwidths for the normal kernel is 71.88 and 34.38, much smaller than 130.79 obtained by the proposed procedure. Figure 3 also shows the boundary-kernel density estimate (dashed line). It is clear that the estimates are still too rough.

7. Bounded Support

In this section, we extend the proposed procedures to the case that the density f has a finite support. Without loss of generality, we assume that the support is $[0,1]$.

The characteristic function of f is $\int_0^1 f(x)\exp(i\lambda x)dx$. Integration by parts yields

$$\phi(\lambda) = \{\exp(i\lambda)f(1) - f(0)\}/(i\lambda) + \{\exp(i\lambda)f'(1) - f'(0)\}/\lambda^2 - \phi_2(\lambda)/\lambda^2, \quad (7.1)$$

where $\phi_2(\lambda) = \int_0^1 f''(x)\exp(i\lambda x)dx$. A function with a bounded support can be represented by a Fourier series, instead of an integration, see Edwards (1979, Chap. 1). More precisely, f can be represented as a Fourier series with coefficients $\phi(\lambda)$ at the frequencies $\lambda_{2j} = 2\pi j$. At the frequencies $\lambda_{2j} = 2\pi j$, (7.1) becomes $\phi(\lambda_{2j}) = \{f(1) - f(0)\}/(i\lambda_{2j}) + \{f'(1) - f'(0)\}/\lambda_{2j}^2 - \phi_2(\lambda_{2j})$. As in the case with the support $[0,\infty)$, the discontinuities in f dominate the imaginary part of $\phi(\lambda_{2j})$. Following the arguments in Section 3, we try to use only the real part of

$\tilde{\phi}$ to estimate f. Using the real part is essentially the same as using the expanded data set including the original ones and their images on $[-1,0]$.

Since the length of the support of the expanded density is doubled, we need to consider the characteristic functions at the frequencies $\lambda_j = \pi j$ at odd j's as well as at the even j's. For an odd j, $\phi(\lambda_j) = -\{f(1)+f(0)\}/(i\lambda_j) - \{f'(1) + f'(0)\}/\lambda_j^2 - \phi_2(\lambda_j)$. Letting $c_1 = -\{f'(1) - f'(0)\}$ and $c_2 = -\{f'(1)+f'(0)\}$, we have $\phi_r(\lambda_j) \approx c_1/\lambda_j^2$ for odd j and $\phi_r(\lambda_j) \approx c_2/\lambda_j^2$ for even j

Similar to (4.1), c_1 and c_2 are estimated by, respectively, $\hat{c}_1 = \sum_{j>J_1} \tilde{\phi}_r(\lambda_{2j+1}) /\lambda_{2j+1}^2 / \sum_{j>J_1}(1/\lambda_{2j+1}^4)$ and $\hat{c}_2 = \sum_{j>J_2} \tilde{\phi}_r(\lambda_{2j})/\lambda_{2j}^2/\sum_{j>J_1}(1/\lambda_{2j}^4)$, where J_1 and J_2 are selected as described in Section 4, with some modification for switching from integration to summation. For the current case, we suggest using a quadratic function ax^2+bx to reduce the boundary effects. The coefficients are $a = -c_2/2$ and $b = (c_2 - c_1)/2$.

Finally, we outline the implementation of the procedures. We first discretize the data in the interval $[0,1]$ by defining $Y_t = F_n(t/N) - F_n\{(t-1)/N\}$, $t = 1,\ldots,N$. In order to obtain $\tilde{\phi}(\lambda)$ at $\lambda = \pi j$, N zeros are appended to the series Y_t. Applying the Fourier transform to the expanded series gives us $\tilde{\phi}(\lambda_j)$. Except for estimating c_1 and c_2, the remaining part of the implementation is essentially the same as in the case of infinite support. We estimate c_1 by using $\tilde{\phi}_r(\lambda_j)$ at odd j's and c_2 at even j's.

8. Proofs

We first prove Theorem 1 when $f^{(3)}$ is of bounded variation and $f^{(3)} \neq 0$ $(l=1)$. We redefine $\tilde{\phi}_d = \tilde{\phi}_r - \phi_r$ as the noise part of $\tilde{\phi}_r$. Since $\tilde{\phi}_r(\lambda) - \psi(\hat{\theta},\lambda) = \phi_r(\lambda) - \psi(\theta,\lambda) + \tilde{\phi}_d(\lambda) - \{\psi(\hat{\theta},\lambda) - \psi(\theta,\lambda)\}$, the difference between the first term in (3.4) and (3.5) can be decomposed to a few terms, $D_1+D_2+D_3+D_4+D_5+D_6$, where

$$2\pi D_1 = \int_{|\lambda|>|\Lambda} \{\phi_r(\lambda) - \psi(\theta,\lambda)\}^2\{1-W(\beta\lambda)\}^2 d\lambda,$$

$$2\pi D_2 = \int_{-\Lambda}^{\Lambda} \{\phi_d^2(\lambda) - 1/(2n)\}\{1-W(\beta\lambda)\}^2 d\lambda,$$

$$2\pi D_3 = \int_{-\Lambda}^{\Lambda} \phi_d(\lambda)\{\phi_r(\lambda) - \psi(\theta,\lambda)\}\{1-W(\beta\lambda)\}^2 d\lambda,$$

$$2\pi D_4 = \int_{-\Lambda}^{\Lambda} \{\psi(\theta\lambda) - \psi(\hat{\theta},\lambda)\}^2\{1-W(\beta\lambda)\}^2 d\lambda,$$

$$2\pi D_5 = \int_{-\Lambda}^{\Lambda} \phi_d(\lambda)\{\psi(\theta\lambda) - \psi(\hat{\theta},\lambda)\}\{1-W(\beta\lambda)\}^2 d\lambda,$$

$$2\pi D_6 = \int_{-\Lambda}^{\Lambda} \{\phi_r(\lambda) - \psi(\theta,\lambda)\}\{\psi(\theta,\lambda) - \psi(\hat{\theta},\lambda)\}\{1-W(\beta\lambda)\}^2 d\lambda.$$

Under the assumption of Theorem 1, the Λ selected by (3.6) satisfies $O(n^{1/7}) = \Lambda = o(n^{1/7+\delta})$ for any constant $\delta > 0$. In the following, δ is used as a positive generic constant whose meaning depends on the context in which it is used. The convergence of the terms D_j/β^4, $j = 1, 2, 3$ can be established by using the arguments in Chiu (1992). For the remaining terms, we have $\beta^{-4}D_4 = O\{(\hat{\theta} - \theta)^2\Lambda\} = o(n^{-1/2})$, since $1 - W(\beta\lambda) = O(\beta^2\lambda^2)$ as $\beta\lambda \to 0$. Since $\tilde{\phi}_d(\lambda) = O(n^{-1/2})$, $\beta^{-4}D_5 = O\{n^{-1/2}(\hat{\theta} - \theta)\Lambda^3\} = o(\hat{\theta} - \theta)$. For the last term , $\beta^{-4}D_6 = O(\hat{\theta} - \theta)$, since $\phi_r(\lambda) - \psi(\theta, \lambda) = O(1/\lambda^4)$. The consistency of the bandwidth estimate follows from these results.

To obtain the rate of convergence, it is essential to find the order of the terms $dD_j/d\beta$, $j = 1, \ldots, 6$, when $\beta = O(n^{-1/5})$. Similar to the argument above, or in Chiu (1992), $\beta^{-3}D'_j = O(n^{-1/2})$, for $j = 3, 4, 5$. For the second term, since the variance of $\int_{-\Lambda}^{\Lambda} \tilde{\phi}_r^2(\lambda)\lambda^4 d\lambda$ is of the order $n^{-2}\Lambda^9$, $\beta^{-3}D'_2 = O(n^{-1}\Lambda^{9/2}) = o(n^{-2/7})$. The bias is dominated by the first term. We have $\beta^{-3}D'_1 = O(\Lambda^{-3}) = O(n^{-3/7})$. The last term dominates the variance. Under Assumption 1, $\psi(\hat{\theta}, \lambda) - \psi(\theta, \lambda) = O\{(\hat{\theta} - \theta)/\lambda^2\}$. Therefore $\beta^{-3}D'_6 = O(\hat{\theta} - \theta)$.

It is easy to extend the above proof under the stronger condition that $f(x)$ and $\psi(\theta, \lambda)$ satisfy Assumptions 1 and 2 with $l \geq 2$. In this case, $\phi_r(\lambda) - \psi(\theta, \lambda) = O(1/\lambda^{2l+2})$ and $O(n^{1/(4l+3)}) = \Lambda = o(n^{1/(4l+3)+\delta})$. Thus, $\beta^3 D'_1(\beta) = O(n^{-(4l-1)/(4l+3)})$, which is of the order $o(n^{-1/2})$ when $l \geq 2$. Also, $\beta^{-3}D'_2(\beta) = o(n^{-1/2})$. Therefore the last term is the only one with an order greater than $n^{-1/2}$, and the proof of Theorem 1 is finished.

To prove Theorem 2, we write $\hat{f}'(0) - f'(0) = B_1 + B_2$, where

$$(3\Lambda^3)^{-1}B_1 = \int_{\Lambda}^{\infty} \{\phi_r(\lambda) - f'(0)/\lambda^2\}/\lambda^2 d\lambda \tag{8.1}$$

and $(3\Lambda^3)^{-1}B_2 = \int_{\Lambda}^{\infty} \phi_d(\lambda)/\lambda^2 d\lambda$. Since (8.1) is of order Λ^{-2l-3}, the bias term $B_1 = O(\Lambda^{-2l})$. For the variance, we note that

$$\mathrm{Var}\{\Lambda^{-3}B_2\} = n^{-1} \int_{\Lambda}^{\infty} \int_{\Lambda}^{\infty} \lambda^{-2}\mu^{-2}\{\phi_r(\lambda - \mu) - \phi_r(\lambda)\phi_r(\mu)\} d\lambda d\mu.$$

Since $\int_{\Lambda}^{\infty} \mu^{-2} \int_{\Lambda-\mu}^{\mu-\Lambda} (\lambda + \mu)^{-2} \phi_r(\lambda) d\lambda d\mu$ is of order Λ^{-3}, the variance of $\hat{f}'(0)$ is of order Λ^3/n, as stated in the theorem.

Acknowledgement

We thank the Associate Editor and the referees for comments that greatly enhanced the paper.

References

Bowman, A. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* **71**, 353-360.

Blackman, R. B. and Tukey, J. W. (1959). *The Measurement of Power Spectrum, from the Point of View of Communications Engineering*. Dover, New York.

Chiu, S. T. (1991). Bandwidth selection for kernel density estimation. *Ann. Statist.* **19**, 1883-1905.

Chiu, S. T. (1992). An automatic bandwidth selector for kernel density estimation. *Biometrika* **79**, 771-782.

Chiu, S. T. (1996). A comparative review of bandwidth selection for kernel density estimation. *Statist. Sinica* **6**, 129-145.

Cowling, A. and Hall, P. (1996). On pseudodata methods for removing boundary effects in kernel density estimation. *J. Roy. Statist. Soc. Ser. B* **58**, 551-563.

Edwards, R. E. (1979). *Fourier Series: A Modern Introduction*. Vol. 1. 2nd edition. Springer-Verlag, New York.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications.* Vol. 1. 3rd edition. Wiley, New York.

Feller, W. (1966). *An Introduction to Probability Theory and Its Applications*. Vol. 2. Wiley, New York.

Gasser, T., Müller, H. G. and Mammitzsch, V. (1985). Kernels for nonparametric curve estimation. *J. Roy. Statist. Soc. Ser. B* **47**, 238-252.

Hall, P. (1980). On trigonometic series estimates of densities. *Ann. Statist.* **9**, 683-685.

Hall, P. and Marron, J. S. (1987). Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation. *Probab. Theory Related Fields* **74**, 567-581.

Hall, P., Marron, J. S. and Park, B. (1992). Smoothed cross-validation. *Probability Theory and Related Fields* **92**, 1-20.

Hall, P., Sheather, S. J., Jones, M. C. and Marron, J. S. (1991). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika* **78**, 263-269.

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

Jarrett, R. G. (1979). A note on the intervals between coal-mining disasters. *Biometrika* **66**, 191-193.

Jones, M. C., Marron, J. S. and Park, B. U. (1991). A simple root n bandwidth selector. *Ann. Statist.* **19**, 1919-1932.

Park, B. and Marron, J. S. (1990). Comparison of data-driven bandwidth selectors. *J. Amer. Statist. Assoc.* **85**, 66-72.

Raftery, A. E. and Akman, V. E. (1986). Bayesian analysis of a Poisson process with a change-point. *Biometrika* **73**, 85-89.

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

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27**, 832-837.

Rosenblatt, M. (1971). Curve estimates. *Ann. Math. Statist.* **42**, 1815-1842.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.* **9**, 65-78.

Schuster, E. F. (1985). Incorportating support constraints into nonparametric estimation of densities. *Commun. Statist. Theory Meth.* **14**, 1123-1136.

Scott, D. W. and Terrell, G. R. (1987). Biased and unbiased cross-validation in density estimation. *J. Amer. Statist. Assoc.* **82**, 1131-1146.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. Ser. B.* **53**, 683-690.

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

Woodroofe, M. (1982). On Model selection and the arc sine laws. *Ann. Statist.* **10**, 1182-1194.

Worsley, K. J. (1986). Confidence regions and tests for a change-point in a sequence of exponential family random variables. *Biometrika* **73**, 91-104.

Department of Statistics, Colorado State University, Fort Collins, Colorado 80523, U.S.A.

E-mail: chiu@stat.colostate.edu

(Received July 1998; accepted March 2000)