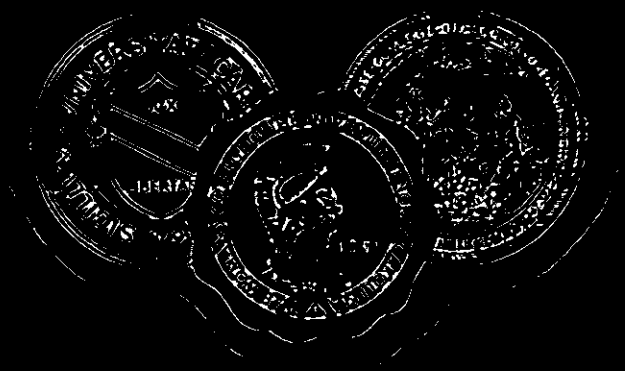
THE INSTITUTE OF STATISTICS

THE CONSOLIDATED UNIVERSITY OF NORTH CAROLINA



TRANSFORMATIONS TO REDUCE BOUNDARY BIAS IN KERNEL DENSITY ESTIMATION

by

J.S. Marron

and

D. Ruppert

January 1993

Mimeo Series #2300

DEPARTMENT OF STATISTICS
Chapel Hill, North Carolina

TRANSFORMATIONS TO REDUCE BOUNDARY BIAS IN KERNEL DENSITY ESTIMATION[1]

J. S. Marron
Department of Statistics
University of North Carolina
Chapel Hill, NC 27514

D. Ruppert
School of Operations Research
and Industrial Engineering
Cornell University
Ithaca, NY 14853

January 22, 1993

SUMMARY

We consider kernel estimation of a univariate density whose support is a compact interval. If the density is non-zero at either boundary, then the usual kernel estimator can be seriously biased. "Reflection" at a boundary removes some bias, but unless the first derivative of the density is 0 at the boundary, the estimator with reflection can still be much more severely biased at the boundary than in the interior. We propose to transform the data to a density that has its first derivative equal to 0 at both boundaries. The density of the transformed data is estimated, and an estimate of the density of the original data is obtained by change-of-variables. The transformation is selected from a parametric family, which is allowed to be quite general in our theoretical study. We propose algorithms where the transformation is either a quartic polynomial, a beta CDF, or a linear combination of a polynomial and a beta CDF. The last two types of transformations are designed to accommodate possible poles at the boundaries. The first two algorithms are tested on simulated data and compared with an adjusted kernel method of Rice. We find that our proposal performs similarly to Rice's for densities with one-sided derivatives at the boundaries. Unlike Rice's method, our proposal is guaranteed to produce nonnegative estimates. This can be a distinct advantage when the density is 0 at either boundary. Our algorithm for densities with poles outperforms the Rice adjustment when the density does have a pole at a boundary.

Key words and phrases: Empirical processes, kernel density estimation, local bandwidths, local least-squares estimation, parametric data transformations, pole at boundary, polynomial transformations, reflection about the boundary.

---

[1] Research of J.S. Marron supported by NSF Grant DMS-8902973. Research of D. Ruppert supported by NSF Grant DMS-9002791.

1. Introduction. Many methods for estimating a probability density f_x have been proposed. When f_x is compactly supported, most estimators, including the popular kernel method, are more biased near the endpoints than in the interior. In practical settings, the boundary region is often 20% to 50%, and sometimes more, of the support of f_x, so that boundary bias can be a serious problem. In this article, we propose a data transformation method for reducing the boundary bias of kernel estimators.

Let $X_1, \ldots, X_n$ be a random sample from f_x. Suppose for convenience that the support of f_x is $[0, 1]$. The conventional kernel estimate of $f_x(x)$ is

$$\tilde{f}_x(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left\{\frac{x - X_i}{h}\right\},$$

where K is the kernel function, and h is a window width depending on n. Here K is assumed to be a symmetric probability density, and to simplify the discussion, we assume its support is $[-1, 1]$.

The poor behavior of $\tilde{f}_x$ at the boundaries can be understood by noting that the support of $\tilde{f}_x$ is $[\min(X_i) - h, \max(X_i) + h]$, which typically is rather larger than $[0, 1]$, and, as shown in Section 2, the expected value of $\tilde{f}_x(0)$ is approximately $1/2 f_x(0)$, with a similar bias existing at the right boundary. This bias at the left boundary can be corrected by reflecting the mass of $\tilde{f}_x$ to the left of 0 into $[0, 1]$, i.e., letting the new estimate for $0 \leq x \leq h$ be $\tilde{f}_x(-x) + \tilde{f}_x(x)$ and setting the estimate equal to 0 for $x < 0$. A similar correction would be made at the right boundary. This reflection reduces the bias, but as seen in Section 2, the bias at 0 after reflection is approximately proportional to $h f_x^{(1)}(0)$ as $h \to 0$. In contrast, the bias in the interior is approximately proportional to $h^2 f_x^{(2)}/2$. Thus, reflection works well for densities whose derivative is equal to 0 at both boundaries, but otherwise when h is sufficiently small the bias at the boundaries will be larger than in the interior.

The methodology proposed in this paper estimates f_x by a three step process. First, a transformation g is selected from a parametric family so that the density of $Y_i = g(X_i)$ has a first derivative that is approximately equal to 0 at the boundaries of its support. Next, a kernel estimator with reflection is applied to the Y_i's. Finally, this estimator is converted, by the change of variables formula, to an estimate of f_x.

We are assuming in this article that the support of f_x is known. This is a common situation because the support is often defined by known constraints, e.g., a proportion must be between 0 and 1. If the support is unknown, it is natural to use the minimum and maximum of the sample as estimates of the boundaries of the support. In this case, the error in estimating the boundaries is of order $O_P(n^{-1})$ if the density is positive and continuous at the boundaries. If the density is 0 but has a nonzero, continuous derivative at the boundaries, then the error in estimating the

boundaries is worse but still only of order $O_P(n^{-1/2})$, smaller than the usual $O_p(n^{-2/5})$ error in density estimation. If the density and its first derivative are both 0 at a boundary, then the boundary will be difficult to estimate but the consequences of mis-estimating the boundary will be minor. For example, if $f_x(x) = (A + o(1))x^{k-1}/(k-1)$ as $x \to 0$, with $A > 0$ and $k \geq 1$, then for $u > 0$

$$Pr\{n^{1/k} \min(X_i) \leq u\} = 1 - \{1 - F_x(u/n^{1/k})\}^n$$

$$= 1 - \{1 - (A + o(1))u^k n^{-1}\}^n \to 1 - \exp(-Au^k) \text{ as } n \to \infty.$$

If the left boundary of the support is estimated by $\min(X_i)$ and the density estimate is set equal to 0 below $\min(X_i)$, then the contribution to the integrated squared error (ISE) from the region $[0, \min(X_i)]$ is of order $O_P(\int_0^{\min(X_i)}(x^{k-1})^2 dx) = O_P(n^{-2+1/k}) = O_P(n^{-1})$ since $k \geq 1$. It is well-known that the contribution to the ISE from the interior is of larger order of magnitude, $O_P(n^{-4/5})$. There are more sophisticated estimators of the endpoints of the support of a density than the sample extremes—see Smith (1987) for some results and earlier references. These might be used in the context of density estimation. Also, if $f_X^{(1)}$ exists everywhere and is absolutely continuous and if $f_X^{(2)}$ exists and is bounded and absolutely continuous on $(0,1)$, then reflection is not needed in the sense that the mean integrated squared error (MISE) of $\tilde{f}_X$ is $O(n^{-4/5})$ (Cline and Hart, 1991).

In Section 2, we analyze the boundary bias of kernel estimation both without and with reflection. We also define the transformation/kernel estimator. Section 3 discusses the asymptotic behavior of the transformation/kernel estimator when a data-based transformation is selected from a parametric family. In Section 4, we use the results of Section 3 to propose a method where g is a quartic polynomial. This proposal is tested on simulated data and compared to Rice's (1984) modified boundary kernels. In Section 5, we consider the case where $f_X(x) = Ax^\alpha$, $A > 0$, and $\alpha \in (-1, 1)$. If $\alpha < 0$, then the bias of a kernel estimator with reflection approaches ∞ as $x \to 0$. Empirical evidence indicates that the bias for x near 0 can be dramatically reduced if the transformation is a beta CDF with parameters estimated appropriately (see Algorithm P). It is common for the support of f_X to have only one boundary, e.g., to be $[0, \infty)$. An appropriate algorithm for this situation is discussed briefly in Section 6. Section 7 discusses other papers on transformations applied to kernel density estimation and suggests some possible future work. Proofs are found in the Appendix.

2. Boundary bias and transformation/kernel estimation. Figure 1a shows the boundary problem for $\tilde{f}_X$. In particular at the left boundary, $x = 0$, the estimates (solid lines) are roughly half the height of the true density (dotted line), and a substantial part of the probability

mass of $\tilde{f}_x$ is placed outside $[0,1]$. This problem can be quantified in terms of bias. Useful insights are obtained through asymptotic analysis as $h \to 0$. For simple presentation, we consider only the left boundary area, $x \in [0,h]$. Since this region vanishes in the limit, one may parametrize a "typical point" by $x = Ch$, for $C \in [0,1]$. As shown in the appendix, assuming f_x has two continuous derivatives on a neighborhood of $x = 0$,

$$\mathbf{E}\tilde{f}_x(x) = \mu_0(C)f_x(x) - h\mu_1(C)f_x^{(1)}(x) + \frac{h^2}{2}\mu_2(C)f_x^{(2)}(x) + o(h^2), \text{ if } x = Ch,\ 0 \le C \le 1, \quad (2.1)$$

where $\mu_k(C) = \int_{-1}^{C} u^k K(u)du$. Since $\mu_0(1) = 1$, $\mu_1(1) = 0$, and $\mu_2(1) = \int u^2 K(u)du$, we have the usual interior bias expansion given by (2.3) below when $C = 1$. The "average half height" of the estimates at $x = 0$ in Figure 1a is explained by $\mu_0(0) = 1/2$. Since $\mu_0(C) < 1$ for $C \in [0,1)$, the kernel estimator is generally inconsistent in the boundary region, i.e., we have a "0-order bias."

Schuster (1985) proposes an adjustment by "reflection about the boundaries" or "boundary folding." This is implemented as

$$\begin{aligned}
\hat{f}_x(x) &= \tilde{f}_x(x) + \tilde{f}_x(-x) \quad \text{if} \quad x \in [0,h) \\
&= \tilde{f}_x(x) \quad \text{if} \quad x \in [h, 1-h] \\
&= \tilde{f}_x(x) + \tilde{f}_x(2-x) \quad \text{if} \quad x \in (1-h, 1] \\
&= 0 \quad \text{if} \quad x \notin [0,1].
\end{aligned}$$

Figure 1b shows the usual effect of this adjustment. Note there is still substantial bias present at both boundaries. This is quantified analytically, under the above assumptions, by

$$\begin{aligned}
\mathbf{E}\hat{f}_x(x) &= f_x(x) + \frac{h^2}{2}\mu_2(1)f_x^{(2)}(x) - 2h[C\mu_0(-C) + \mu_1(-C)]f_x^{(1)}(x) \\
&+ 2h^2[C^2\mu_0(-C) + C\mu_1(-C)]f_x^{(2)}(x) + o(h^2), \text{ if } x = Ch,\ 0 \le C \le 1. \quad (2.2)
\end{aligned}$$

(2.2) is proved in the Appendix. The first two terms are the same as for the familiar interior bias expansion

$$\mathbf{E}\hat{f}_x(x) = f_x(x) + \frac{h^2}{2}f_x^{(2)}(x) + o(h^2), \text{ if } h \le x \le 1-h. \quad (2.3)$$

The other terms in (2.2) represent additional bias due to the boundary reflection. Note they are 0 for $C = 1$, but give a bias of order h, i.e. a "first order bias," for $C < 1$, if $f_x^{(1)}(0) \neq 0$. This bias is still serious, as indicated visually in Figure 1b, and also in terms of integrated L^p, $p > 1$, norm for the error over $[0,1]$, which is dominated by the poor behavior on the vanishing intervals $[0,h]$ and $[1-h,1]$—see van Eeden (1985) and Cline and Hart (1991) for the case $p = 2$.

However, when $f_x^{(1)}(0) = 0$, the bias of $\hat{f}_x$ is of second order, i.e., $O(h^2)$. In this paper we develop a methodology which exploits this fact.

Our notation for the transformed data is $Y_i = g(X_i)$, where g is a smooth, monotonically increasing function. We shall assume that $g(0) = 0$ and $g(1) = 1$, but this is only for convenience and does not restrict the generality of our methodology. The density f_Y of Y_i is related to f_X by $f_X(x) = f_Y(g(x))g'(x)$. Although g will depend on $\{X_i\}$, as a heuristic device we will for the moment treat g as deterministic. If $\hat{f}_Y(y)$ is an estimate of f_Y, the corresponding transformed estimator of $f_X(x)$ is $\hat{f}_X(x;g) = \hat{f}_Y(g(x))g'(x)$. This estimator has bias

$$\mathbf{E}\hat{f}_X(x;g) - f_X(x) = g'(x)\{\mathbf{E}\hat{f}_Y(g(x)) - f_Y(g(x))\}.$$

Our goal is to choose g so that f_Y is flat at the boundaries, and $|g'(x)|$ is bounded. Then if $\hat{f}_Y$ is the boundary reflected kernel estimator, even at the boundaries the resulting transformed estimator will have second order bias.

One example of a transformation that would be effective in this regard is $g = F_X$, the CDF transformation, which would result in Y_i being uniformly distributed on $[0,1]$. Similarly, if $g(x) = DF_X(x)$, for some $D > 0$, in neighborhoods of 0 and 1, then $f_Y^{(1)}(y) = 0$ for $y = 0$ or 1. In this article, we mimic this behavior by estimating $F_X^{(j)}$, $j = 1,2$ at $x = 0$ and 1, and choosing a transformation with the same derivatives.

3. Parametric transformations. In this article, it is assumed that the transformation applied to the data is estimated from a parametric family, that this family contains a "target transformation" such that $f_Y^{(1)}$ is 0 at the boundaries if this transformation is used, and that the estimation transformation converges to the target. The results in this section relate the rate of convergence of the transformation to the asymptotic behavior of the transformation/kernel estimator.

Let $\Theta \subset R^k$ and let $G = \{g_\theta : \theta \in \Theta\}$ be a parametric family of transformations with domain $[0,1]$. We will assume throughout this paper that each element $g_\theta(x)$ of G is three times continuously differentiable with respect to x and has a strictly positive derivative. Derivatives at 0 and 1 are, of course, one-sided. The kernel K is assumed to be Lipschitz continuous. We continue the convenient assumption that $g_\theta(0) = 0$ and $g_\theta(1) = 1$. Since g is no longer considered fixed, we change notation slightly and write $f_Y(\cdot;g)$ in place of f_Y, and let $\hat{f}_X(x;g) = \hat{f}_Y(g(x);g)g'(x)$. We shall be concerned with the selection of $g_n = g_{\theta_n}$ from G based on $\{X_1,\ldots,X_n\}$. We will assume throughout this section that $h = C_0 n^{-1/5}$ for some $C_0 > 0$ and that f_X has two bounded derivatives on $[0,1]$.

Suppose that for some θ_0, $\theta_n \to \theta_0$ in probability at some rate to be specified later. How close are $\hat{f}_Y(g_{\theta_n}(x);g_{\theta_n})$ and $\hat{f}_Y(g_{\theta_0}(x);g_{\theta_0})$? We will first compare them in terms of "variation about

the mean" and then in terms of bias. We need the following assumption about the family G.

Condition 1. *There exists* $K_1 > 0$ *such that*

$$\max_{j=1,2,3} \sup_{x' \in [0,1]} |g_\theta^{(j)}(x') - g_{\theta_0}^{(j)}(x')| \leq K_1 \|\theta - \theta_0\|.$$

Here and in what follows, derivatives of g_θ are with respect to x, not θ, and $\|\cdot\|$ denotes the Euclidean norm on R^k. An example of when Condition 1 is satisfied is when g_θ is a polynomial in x whose coefficients are linear functions of θ.

Theorem 1. *Suppose* $\theta_0 \in \Theta$, $x \in [0,1]$, *and Condition 1 holds. Define* $G_\Delta = \{g_\theta : \|\theta - \theta_0\| \leq \Delta$ *and* $\theta \in \Theta\}$. *Let* $\Delta_n = (\log n)^{-1/2-\epsilon}$ *for some* $\epsilon > 0$. *Then for all sequences* $\{x_n\} \subset [0,1]$

$$\sup_{g \in G_{\Delta_n}} \left| [\hat{f}_Y(g(x_n); g) - \mathbf{E}\hat{f}_Y(g(x_n); g)] - [\hat{f}_Y(g_{\theta_0}(x_n); g_{\theta_0}) - \mathbf{E}\hat{f}_Y(g_{\theta_0}(x_n); g_{\theta_0})] \right| = o((nh)^{-1/2}) \tag{3.1}$$

almost surely.

All proofs are in the appendix.

Note that the RHS of (3.1) is of smaller order than the standard deviation of $\hat{f}_Y(g_{\theta_0}(x_n); g_{\theta_0})$, which is asymptotic to a constant multiple of $(nh)^{-1/2}$. If $\|\theta_n - \theta_0\| = o_P(\Delta_n)$, then $Pr\{g_{\theta_n} \in G_{\Delta_n}\} \to 1$, so by Theorem 1

$$\hat{f}_Y(g_{\theta_n}(x_n); g_{\theta_n}) = \hat{f}_Y(g_{\theta_0}(x_n); g_{\theta_0}) + \left\{ [\mathbf{E}\hat{f}_Y(g(x_n); g)]\Big|_{g=g_{\theta_n}} - \mathbf{E}\hat{f}_Y(g_{\theta_0}(x_n); g_{\theta_0}) \right\} + o_P((nh)^{-1/2}).$$

Therefore, in terms of variation about the mean $\hat{f}_Y(g_{\theta_n}(x_n); g_{\theta_n})$ and $\hat{f}_Y(g_{\theta_0}(x_n); g_{\theta_0})$ are asymptotically equivalent. The following results follow directly from (2.2).

Theorem 2. *Assume Condition 1 holds. Then*

$$[\mathbf{E}\hat{f}_Y(g(x_n); g)]\Big|_{g=g_{\theta_n}} = \mathbf{E}\hat{f}_Y(g_{\theta_0}(x_n); g_{\theta_0}) + o_P(h^2) \tag{3.2}$$

either if $\theta_n = \theta_0 + o_P(h)$ *and* $x_n = Ch$ *or* $x_n = (1 - Ch)$ *for* $C \in [0,1]$, *or if* $\theta_n \to \theta_0$ *and* $x_n \equiv x \in (0,1)$ *is fixed. Under (3.2)* $|\hat{f}_X(x_n; g_{\theta_n}) - \hat{f}_X(x_n; g_{\theta_0})| = o_p(h^2)$. *If* $\theta_n = \theta_0 + O_P(h)$, *then*

$$\begin{aligned} \hat{f}_X(x_n; g_{\theta_n}) &= \hat{f}_X(x_n; g_{\theta_0}) + O_P(n^{-2/5}) \quad \text{if } x_n = Ch \text{ or } x_n = (1 - Ch), C \in [0,1] \\ &= \hat{f}_X(x; g_{\theta_0}) + o_P(n^{-2/5}) \quad \text{if } x_n \equiv x \in (0,1). \end{aligned}$$

If $g_1 = F_X$ then $f_Y^{(1)}(g_1(x); g_1) \equiv 0$. Theorem 3 below gives a simple condition on the first two derivatives of g_2 so that $f_Y^{(1)}(g_2(x); g_2) = 0$ for $x = 0$ or 1: at the boundaries the first two derivatives of g_2 should equal a constant multiple of those of F_X.

Theorem 3. *Suppose that for some* $D > 0$ $g_1'(x) = Dg_2'(x) > 0$ *and* $g_1''(x) = Dg_2''(x)$ *for some* $x \in [0,1]$. *Suppose that* $f_X^{(1)}(x)$ *exists. Then*

$$f_Y^{(j)}(g_1(x); g_1) = f_Y^{(j)}(g_2(x); g_2)/D^{j+1} \quad j = 0, 1.$$

To apply Theorem 3, we need the following assumption.

Condition 2. *There exist* $\theta_0 \in \Theta$ *and* $c_1, c_2 > 0$ *such that*

$$(g_{\theta_0}'(0), g_{\theta_0}''(0), g_{\theta_0}'(1), g_{\theta_0}''(1)) = (c_1 f_X(0), c_1 f_X^{(1)}(0), c_2 f_X(1), c_2 f_X^{(1)}(1)).$$

In the following theorem we look at $f_X(\cdot; g)$ at a sequence $\{x_n\}$ contained in $[0,1]$. This includes as special cases x_n fixed (independent of n) and x_n equal to Ch or $(1 - Ch)$ for some $C \in [0,1]$.

Theorem 4. *Assume Conditions 1 and 2 hold. Then* $\hat{f}_X(x_n; g_{\theta_0}) = f_X(x_n) + O_P(n^{-2/5})$ *for all sequences* $\{x_n\} \subset [0,1]$. *If*

$$\theta_n = \theta_0 + O_P(h), \tag{3.3}$$

then

$$\hat{f}_X(x_n; g_{\theta_n}) = f_X(x_n) + O_P(n^{-2/5}) \quad \text{for all } \{x_n\} \subset [0,1]. \tag{3.4}$$

4. A proposal and simulations. In this section we propose a specific transformation with asymptotic properties considered in Section 2, and investigate its performance on simulated data. In the first step of the algorithm the data are binned to form a histogram estimate of f_X. In the next step, f_X and $f_X^{(1)}$ are estimated at 0 and 1 by a local linear least-squares fit to this histogram—see Cleveland (1979) and Müller (1987) for an introduction to local least-squares regression. The local least-squares estimated coefficients are truncated to satisfy the constraints of a density and its derivatives at the boundaries. In Steps 3 and 4 a quartic polynomial transformation g_n is found that is monotonically increasing and that approximates F_X near 0 and 1.

Algorithm D—for bias reduction when f_X has finite, continuous Derivatives at the boundaries

(1) Let m be a positive integer. Divide $[0,1]$ into m equal length subintervals with centers $z_j = (j - 1/2)/m$, $j = 1, \ldots, m$. Let d_j be the number of X_i in the jth subinterval. (In the simulations we used m.)

(2) Let $h_1 \in [0,1]$. Let b_0^* and b_1^* minimize

$$\sum_{j=1}^{m} \left[\frac{md_j}{n} - (b_0^* + b_1^* z_j)\right]^2 I\{z_j \le h_1\}. \tag{4.1}$$

Let b_2^* and b_3^* minimize

$$\sum_{j=1}^{m} \left[\frac{md_j}{n} - (b_2^* + b_3^* z_j)\right]^2 I\{z_j \ge 1 - h_1\}.$$

Let $b_0 = \max(0, b_0^*)$ and $b_2 = \max(0, b_2^*)$. Let $b_1 = 0$ if $b_0 = 0$ and $b_1^* < 0$ and let $b_1 = b_1^*$ otherwise. Let $b_3 = 0$ if $b_2 = 0$ and $b_3^* > 0$ and let $b_3 = b_3^*$ otherwise. (b_0, b_1, b_2, b_3) will be used as an estimate of $(f_x(0), f_x^{(1)}(0), f_x(1), f_x^{(1)}(1))$.

(3) Let $p_n^*(x) = \sum_{k=0}^{3} a_k x^k$ be the unique cubic polynomial such that $(p_n^*(0), (p_n^*)^{(1)}(0), p_n^*(1), (p_n^*)^{(1)}(1)) = (b_0, b_1, b_2, b_3)$. Thus, (a_0, a_1, a_2, a_3) solves

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{pmatrix},$$

so that $a_0 = b_0$, $a_1 = b_1$, $a_2 = -3b_0 - 2b_1 + 3b_3 - b_4$, and $a_3 = 2b_0 + b_1 - 2b_2 + b_3$. Fix $\gamma \ge 0$. If $\min_{x\in[0,1]} p_n^*(x) \ge \gamma$, then let $p_n = p_n^*$. Otherwise, let $p_n = p_n^* + \gamma - \min_{x\in[0,1]} p_n(x)$. (In the simulations, we used $\gamma = 0$.)

(4) A transformation with correct endpoint behavior is $g_n(x) = \int_0^x p_n(u)\,du \,/\, \int_0^1 p_n(u)\,du$.

(5) Estimate $f_x(x)$ by $\hat{f}_x(x; g_n)$ using bandwidth $h \in [0,1]$.

Condition 3. *Letting p_0 be the unique cubic polynomial such that $(p_0(0), p_0^{(1)}(0), p_0(1), p_0^{(1)}(1)) = (f_x(0), f_x^{(1)}(0), f_x(1), f_x^{(1)}(1))$, p_0 satisfies $\min_{x\in[0,1]} p_0(x) > \gamma$.*

Let $G_k = \{g : g \text{ is a } k\text{th degree polynomial, } g(0) = 0, g(1) = 1, \text{ and } g'(x) > \gamma\ \forall x \in [0,1]\}$, where γ is the same as in Step 3 of Algorithm D. Our algorithm uses $G = G_4$. Condition 3 implies that G_4 satisfies Condition 2 with $c_1 = c_2$.

Theorem 5. *Let $N = \lfloor mh_1 + 1/2 \rfloor$ where $\lfloor\cdot\rfloor$ is the greatest integer function. Suppose $h_i = C_i n^{-1/5}$, $C_i > 0$ for $i = 1, 2$ and $N \ge 2$. Assume that f_x has two bounded derivatives on $[0,1]$ and that Condition 3 holds. Then $\hat{f}_x(x_n; g_n) = f_x(x_n) + O_P(n^{-2/5})$ for all sequences $\{x_n\} \subset [0,1]$.*

If Condition 3 does not hold, especially if $\min_{x\in[0,1]} p_0(x) < 0$, then quartic polynomials may not be a satisfactory family of transformations for the purpose of bias reduction at the boundaries. If $\min_{x\in[0,1]} p_n^*(x)$ is considerably smaller than 0 then this is evidence that Condition 3 is violated

for any $\gamma > 0$ and Step 3 should be modified, say by using a cubic spline instead of a cubic polynomial. One might first consider placing a single knot at 1/2. A local linear squares fit to the histogram at 1/2 could provide the information necessary to determine the spline coefficients. We will not pursue this possibility here.

We performed a modest simulation study with 3 densities on $[0,1]$:

(1) Parabolic density where $f_x(x) = (\frac{4}{3} - 3(x - \frac{1}{3})^2)$. This is an example where Condition 3 is just violated because $\min_{x\in[0,1]} p_0(x) = \gamma = 0$.

(2) Uniform squared (or U^2) density where $f_x(x) = \frac{1}{2}x^{-1/2}$.

(3) Mixture density where $f_x(x) = \frac{10}{3}\phi\left(10(x - \frac{1}{3})\right) + 2x^2$. Here ϕ is the standard normal density. This density is the mixture of a $N(1/3, 1/100)$ density and a Beta$(3,1)$ density with mixing proportions of 1/3 and 2/3, respectively. Strictly speaking, this density has infinite support, but the amount of probability outside of $[0,1]$ is negligible.

For each density, we simulated 8 independent random samples, each of size $n = 500$. Figures 1–3 show the true underlying densities (dotted line) and the eight estimates (solid lines) based on various estimators, including a modification of a proposal of Rice (1984). Rice uses a kernel depending on x such that $\mu_{1,x} = 0$ for all x. By (1.2), Rice's estimator has bias of order $O(h^2)$ for all x. Rice's estimator has the disadvantage that it can take negative values. The bandwidth $h = h_2$ of the kernel estimate is shown in each plot and h_1 used in Step 1 is given in the captions. The bandwidths were chosen subjectively with the intention that the different estimators would do roughly the same amount of smoothing.

Figure 1—Parabolic density. As discussed in the Introduction, Figure 1a shows the estimator without adjustment for boundary effects and Figure 1b shows the estimator with reflection but without transformation. In Figure 1a we see a negative 0-order bias at 0 because $f_x(0) > 0$, while in Figure 1b we see a positive 1st order bias at 0 because $f_x^{(1)}(0) > 0$. Since $f_x(1) = 0$, there is no 0-order bias at 1 in Figure 1a, but the positive 1st order bias at 1 is evident in Figures 1a and 1b.

Figure 1c shows the estimate from Algorithm D with bandwidths $h = 0.18$. Rice's estimator with $h = 0.12$ is shown in Figure 1d. Notice that Figures 1c and 1d are quite similar. It appears that Algorithm D and Rice's estimator are competitive in this sampling situation, but the bandwidths for the two estimators are not comparable—but see a modification of Algorithm D called Algorithm D-LB below which is comparable to Rice's algorithm with the *same* bandwidth.

The Algorithm D estimator is still somewhat biased at the boundaries. Since $f_x(1) = 0$, the bias at the right boundary is unavoidable when one uses a nonnegative estimator. The bias at the

left boundary can be reduced by lowering the value of h_1 used there. Although Algorithm D as presented here uses the same value of h_1 at the left and right boundary, in practice one might use different h_1's. We tried using $h_1 = .1$ at the left boundary and $h_1 = .4$ at the right. The result was that $f_X(0)$ was underestimated by four of eight estimates (rather than one of eight as in Figure 1c), but the curves were more variable.

Figure 2—U^2 density. This density has a pole at 0, so it is not clear that either Algorithm D or Rice's estimator will be satisfactory here. Figures 2a and 2b show Algorithm D and Rice's estimator, both with $h = 0.06$. They are quite similar, both perhaps somewhat undersmoothed. Figure 2c illustrates Rice's estimator with a larger bandwidth, $h = 0.12$. The wiggles seen in Figures 2a and 2b are mostly gone, but the bias near 0 is severe. Figure 2d illustrates an algorithm designed for densities with poles. This algorithm and that figure will be discussed in the next section.

Figure 3—Mixture density. Figures 3a and 3b show Algorithm D and Rice's estimator, respectively. Here Rice's estimator seems better. A side effect of transformation is that Algorithm D oversmooths near 0 and undersmooths near 1. Wand, Marron, and Ruppert (1991) discuss the effects of transformation on the amount of local smoothing of the estimator. They show that in terms of variance, a transformation/kernel estimator has an "effective" local bandwidth roughly equal to $h/g'(x)$. This can be shown by a Taylor expansion:

$$\begin{aligned}\hat{f}_X(x;g) &= \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{g(x)-g(X_i)}{h}\right) g'(x) \\ &\simeq \frac{1}{n(h/g'(x))}\sum_{i=1}^{n} K\left(\frac{x-X_i}{(h/g'(x))}\right). \qquad (4.2)\end{aligned}$$

Approximation (4.2) is not sufficiently precise to show the effect of a transformation of the higher derivatives of f_Y (and hence on the bias of $\hat{f}_X(x;g)$), but it does show the effect of local variation in effective bandwidth upon the variance $\hat{f}_X(x;g)$. Wand et al. recommended transformations for estimation of densities where an effective bandwidth that is constant with respect to location is undesirable. They developed a methodology for choosing a transformation to minimize an estimate of mean integrated squared error (MISE). Since MISE depends on both bias and variance, the Wand et al. methodology seeks an optimal compromise between the transformations's effects on the higher derivatives of f_Y and on the local variance of the estimator. Algorithm D is different—it reduces bias by reducing higher derivatives of f_Y without regard to the local effective bandwidth. For the mixture density, the effective local bandwidth of Algorithm D is large near 0 and small near 1. This variation in local effective bandwidth is not desirable here, and can be removed by

using the following estimate:

$$\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{g(x)-g(X_i)}{g'(x)h}\right). \tag{4.3}$$

(4.3) was suggested by M. C. Jones's comments (personal communication) on Ruppert and Cline (1991). We were not certain how best to apply reflection to (3.3), so instead of using reflection we divided (4.3) by

$$\frac{1}{h}\int_0^1 K\left(\frac{g(x)-g(y)}{g'(x)h}\right) dy. \tag{4.4}$$

(4.4) is equal to 1 when x is sufficiently far in the interior of $[0,1]$. This adjustment is similar to the reflection method described in Section 1 in that it removes the 0-order bias. See Diggle and Marron (1987) for further discussion of this rescaling type of boundary modification, including comparison to the reflection method. Rescaling is a natural adjustment and an associate editor suggested using it for the other algorithms in this article. However, reflection is slightly more efficient than rescaling (Diggle and Marron, 1987). We call this estimate a "local bandwidth estimator", and Algorithm D with estimate (4.3) divided by (4.4) will be called Algorithm D-LB. This algorithm is illustrated in Figure 3c. It is similar to Rice's estimator except near 0 where the latter is seriously negative. Algorithm D-LB was also tested on the Parabolic and U^2 densities, where its performance was quite similar to Rice's algorithm with the same bandwidth.

Both asymptotics and this limited simulation study indicate that for densities with finite, continuous derivatives at the boundaries, Algorithm D-LB and Rice's algorithm will be nearly equally effective in practice and which one uses is mostly a matter of convenience. Algorithms D and D-LB are not comparable. The Mixture density is an example where Algorithm D-LB is preferable to Algorithm D, but there will be densities where the opposite is true. Of course for many densities, especially those for which f_X is bounded well above 0 on its support, Algorithms D and D-LB will be similar. One advantage of Algorithm D is that it always produces an estimate that is a bona fide density, whereas Rice's estimate may be negative and Algorithm D-LB may produce an estimate not integrating exactly to 1. However, the integration error should be small and can be removed by rescaling.

5. Poles at the boundaries. The methodology of the previous sections can accommodate densities with (one-sided) first derivatives at the boundaries. It is not designed to handle poles at the boundaries, i.e. densities with either $\lim_{x\downarrow 0} f_X(x)$ or $\lim_{x\uparrow 1} f_X(x)$ equal to ∞. We saw in the Monte Carlo study that Algorithm D is only partly successful at reducing boundary bias at the U^2 density that has a pole at 0.

Suppose now that

$$\lim_{x \downarrow 0} \frac{f_x(x)}{x^{\alpha_1}} = A_1 \text{ and } \lim_{x \uparrow 1} \frac{f_x(x)}{(1-x)^{\alpha_2}} = A_2, \qquad A_1, A_2 > 0. \tag{5.1}$$

Since f_x is a density, $\alpha_1, \alpha_2 > -1$. Boundary bias will be most severe when $\alpha_i < 0$ so that f_x has a pole, but bias will also be a more than usually serious concern when $\alpha_i \in (0,1)$ so that $f_x^{(1)}$ has a pole, i.e., f_x is vertical at the boundary.

When (5.1) holds, algorithm D should be modified. Our proposal is to fit power functions to the binned data separately at 0 and 1 to estimate α_1 and α_2. Then $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are used to select a beta CDF as the transformation. Since $\text{Var}(md_j/n) \simeq (m/n)\, A z_j^{\alpha}$ by (5.1), weighted nonlinear least-squares should be used to fit power functions with weights proportional to $z_j^{-\alpha}$.

Algorithm P—for bias reduction when f_x has Poles at the boundaries

Replace Steps (2) and (3) of Algorithm D by the following:

(2*) Let $h_1 \in [0,1]$. Let $\hat{A}_1$ and $\hat{\alpha}_1$ minimize

$$\sum_{j=1}^{m} z_j^{-\hat{\alpha}_{1,p}} \Big[\frac{md_j}{n} - (\hat{A}_1 z_j^{\hat{\alpha}_1}) \Big]^2 I\{z_j \le h_1\}, \tag{5.2}$$

where $\hat{\alpha}_{1,p}$ is a preliminary estimate of α. Our recommendation is to solve (5.2) with $\hat{\alpha}_{1,p} = 0$, call the solution $(\hat{A}_1^{(1)}, \hat{\alpha}_1^{(1)})$, let $\hat{\alpha}_{1,p} = \hat{\alpha}_1^{(1)}$, solve (5.2) again to obtain $(\hat{A}_1^{(2)}, \hat{\alpha}_1^{(2)})$, and then to iterate this scheme till convergence.

Let $\hat{A}_2$ and $\hat{\alpha}_2$ minimize

$$\sum_{j=1}^{m} (1-z_j)^{-\hat{\alpha}_{2,p}} \Big[\frac{md_j}{n} - (\hat{A}_2 (1-z_j)^{\hat{\alpha}_2}) \Big]^2 I\{z_j \ge 1 - h_1\}.$$

where $\hat{\alpha}_{2,p}$ is a preliminary estimate of α_2. Again iteration is recommended.

(3*) Let $p_n(x) = x^{\hat{\alpha}_1}(1-x)^{\hat{\alpha}_2}$.

Except for these changes, Algorithm P is the same as Algorithm D. Algorithm P for density U^2 is shown in Figure 2d. The minimizations in step (2*) were performed by a Gauss-Newton algorithm. The estimates are remarkably good, which is not too surprising since this is a density that is ideally suited for Algorithm P, since the U^2 density is a beta density. Boundary bias is almostly completely eliminated. Since there are no "features" in the interior, a large bandwidth is possible and this gives very smooth estimates.

In the following we will discuss Algorithm P only near 0, since the situation near 1 is analogous. Since (5.1) implies that $f_X^{(2)}(x)$ is unbounded near 0 unless $\alpha_1 \geq 2$ or $\alpha_1 = 1$, the second order bias near 0 is, like the boundary bias, severe but both are reduced by transformation. If $0 < \alpha_1 < 1$, then $f_X(0) = 0$ so the effective bandwidth near 0 is very large. This is clearly inappropriate, so a local bandwidth modification of Algorithm P is needed. On the other hand, if $\alpha_1 < 0$ then the effective bandwidth decreases near 0, which is appropriate and is another reason why Algorithm P is effective at poles.

Our theoretical study of Algorithm P has not been extensive, but we have some preliminary results that we will now present. For simplicity we will consider only the right boundary, since the left boundary is analogous. Suppose that $-1 < \alpha_1 < 0$. Let the transformation be $g(x) = Bx^{\xi+1}$ for x near 0, where ξ is "close to" α_1. In these rough calculations we will treat ξ as deterministic and tending to α_1, but our interest is in $\xi = \hat{\alpha}_1$. The boundary region is $0 \leq x \leq (h/B)^{1/(\xi+1)}$. Suppose $x = (ch/B)^{1/(\xi+1)}$ for some $c < 1$. Then we have the following result on the relative bias of $\hat{f}_X(x; g)$.

Theorem 6. *Suppose that* $(\xi - \alpha_1)/\log(h) \to 0$ *as* $n \to \infty$. *Then*

$$\frac{f_X(x) - \mathrm{E}\hat{f}_X(x; g)}{f_X(x)} = (\xi - \alpha_1)\left\{\log(x) - \frac{\log(h)}{\xi + 1}\right\}(1 + o_P(1))$$
$$= \frac{\xi - \alpha_1}{\xi + 1}\log(c/B)(1 + o_P(1)).$$

As $c \to \infty$, the relative bias goes to $+\infty$ or $-\infty$—depending on $\mathrm{sign}(\xi - \alpha_1)$—at the slow rate $|\log(x)|$ but, of course, the absolute bias grows much faster. The difference $(\xi - \alpha_1)$ is clearly the major factor influencing the magnitude of the bias. If we substitute $\hat{\alpha}_1$ for ξ then the left side of the display in Theorem 6 is no longer bias but instead is another component of the variance of $\hat{f}_X(x; g)$. To give a rough idea about the size of this component we need to know the variance of $\hat{\alpha}_1$, which can be found by standard weighted nonlinear least-squares calculations:

Theorem 7. *Let* $N = \lfloor h_1 m + 1/2 \rfloor$ *as in Theorem 5. Suppose that* $N \to \infty$. *Then, as* $n \to \infty$

$$\mathrm{Var}(\hat{\alpha}_1) = \frac{(\alpha_1 + 1)^3}{A_1 n h_1^{\alpha_1 + 1}}(1 + o(1)).$$

A referee mentioned that estimation of α_1 is a special case of the problem of estimating a "tail index" of a Pareto-like distribution. A number of other estimators of the tail index have been proposed. See Smith (1987) for several of these estimators and other references. Many of these estimators are based on a fixed number, k_n, of the smallest order statistics, rather than

all observations within a fixed distance of 0. The various estimators will differ in both bias and variance, and it is unlikely that any will be uniformly "best".

The following argument leads us to conjecture that the variance in Theorem 7 is the best possible for an estimator using only the X_i's that are in $[0, h_1]$. Suppose that $f_x(x) = A_1 x^{\alpha_1}$ for $0 \le x \le h_1$ but that the form of f_x is unknown elsewhere. The conditional density of X_i, given that $0 \le X_i \le h_1$, is $f_x^*(x) = x^{\alpha_1}/(\int_0^{h_1} u^{\alpha_1} du)$, $0 \le x \le h_1$. Define $\nu_j = \nu_j(\alpha_1) = \int_0^{h_1} (\log(u))^j u^{\alpha_1} du$. Then $d/d\alpha_1 \log f_x^*(x) = \log x - \nu_1/\nu_0$. Therefore, the Fisher information for f_x^* is $(\nu_2 - \nu_1^2/\nu_0)/\nu_0$ $= (\alpha_1 + 1)^{-2}$. The expected number of X_i's less than h_1 is $A_1 n h_1^{\alpha_1+1}/(\alpha_1 + 1)$. Let the conditional MLE be the maximizer of $\sum \log f_x^*(X_i) I(X_i \le h_1)$. Then, by standard results the asymptotic variance of the MLE is $(\alpha_1 + 1)^3/[A_1 n h_1^{\alpha_1+1}]$.

Hall (1982) considers a different type of maximum likelihood estimator, one that maximizes the likelihood of a fixed number, k_n, of the smallest order statistics. The asymptotic variance of Hall's estimator is different than the MLE based on all X_i's in $[0, h_1]$.

Algorithm P may not be suitable for some densities with a pole at only one boundary. In particular, consider the case of a pole at 0 but not at 1. The cases $\alpha_2 = 0$ and $\alpha_2 = 1$ do provide models where there is no pole at 1, the U^2 density being an example of the former. Nonetheless, Algorithm D does not allow one to adjust $g'(1)$ and $g''(1)$ independently of the behavior of g at 0. To do this we introduce a third algorithm. This algorithm replaces Steps (2) and (3) of Algorithm D by the following.

Algorithm PD—for reduction of boundary bias when there is a Pole at 0 but f_x is Differentiable at 1

(2**) Estimate b_2 and b_3 as in Step (2) of Algorithm D. Estimate $\hat{A}_1$ and $\hat{\alpha}_1$ as in Step (2*) of Algorithm P.

(3**) Let $p_n(x) = \hat{A}_1 x^{\hat{\alpha}_1} + \beta_1 + \beta_2(1 - x)$ where $p_n(1) = \hat{A}_1 + \beta_1 = b_2$ and $p_n'(1) = \hat{A}_1 \hat{\alpha}_1 - \beta_2 = b_3$.

6. Densities whose support has one boundary. If the support of f_x has one boundary, that is, supp (f_x) is $[\gamma, \infty)$ or $(\infty, \gamma]$ for some γ, then by a linear transformation of the X_i's we can make the support $[0, \infty)$. Then it is natural to choose g so that g' and g'' agree with $F_x^{(1)}$ and $F_x^{(2)}$ at 0 and such that g is approximately linear for large x. A suitable transformation would then be $g(x) = \int_0^x [A + Be^{-Cu}] du$ for appropriate A, B, and C. Let b_0 and b_1 be estimates of $f_x(0)$ and

$f_X^{(1)}(0)$ as in Step 2 of Algorithm D. We want to choose A, B, and C so that

$$g'(0) = A + B = b_0 \text{ and } g''(0) = -BC = b_1, \tag{6.1}$$

and such that e^{-Cx} is, roughly speaking, "scaled to the data." Our suggestion is to set $C = 1/\text{median}(X_i)$ so the slope of the transformation is one at the median, and then to define A and B as solutions to (6.1). After transformation to $Y_i = g(X_i)$, f_Y is estimated using reflection only at 0.

7. Related work and concluding remarks. Transformations and kernel estimation were mentioned by Devroye and Györfi (1985) and Silverman (1986), but the first methodology for data-based selection of a transformation prior to kernel estimation appears to be Wand, Marron, and Ruppert (1991). As mentioned before, the proposal in Wand et al. chooses the transformation by minimizing an estimate of MISE. Special attention was given to shifted power transformations for skewed distributions. Ruppert and Wand (1991) extended this methodology to transformations suitable for heavy tailed, nearly symmetric distributions. Neither of these papers discuss the possibility of reducing the order of the bias.

Ruppert and Cline (1991) study transformation by a (very) smooth estimate of the CDF—the CDF estimate is the indefinite integral of a kernel density estimate. This proposal transforms to a near uniform distribution. By (1.2) the reflected kernel estimate is unbiased at the uniform distribution. Of course, the Ruppert/Cline estimator is not exactly unbiased because f_Y is not exactly uniform, but they prove that their methodology reduces the order of the bias from $O(h^2)$ to $O(h^{2k})$ where k is the number of times that their method is iterated, the first iteration being the ordinary kernel estimate. These rates are only obtained for densities with sufficiently many derivatives.

Ruppert and Cline (1991) restricted their discussion to densities with infinite support. The difficulty with compact support is that the boundary bias of the initial kernel estimate persists during iteration. The algorithms presented here would provide excellent initial estimates for the Ruppert/Cline method. Such an hybrid would have bias of order $O(h^2)$ at the boundaries and of order $O(h^{2k})$ in the interior. Further reduction in boundary bias would be possible by using more sophisticated boundary transformations, using estimates of higher derivatives for the initial estimate.

There are, of course, other methods that could have been used to estimate the parameters $b_0, \ldots, b_3$ in Algorithm D or A_1, α_1, A_2, and α_2 of Algorithm P. We choose least-squares fitting

to bin counts because we use binned data to compute kernel estimates and there exist fast binning algorithms. The linear least square estimates of Algorithm D are very easy and quick to compute, and the nonlinear least squares estimates of Algorithm P, though slower to compute, are not burdensome.

The modeling approach of this article is that algorithms are data-driven, but the choice of algorithm definitely is not. For example, to decide between Algorithms D, P, and DP, one must decide at each boundary whether f_x or $f_x^{(1)}$ has a pole at that boundary or instead f_x and $f_x^{(1)}$ are continuous there. In practice, parametric models are often chosen subjectively, and this seems also appropriate for nonparametric models. Nonetheless, the extent to which the choice between the various algorithms we have proposed can be automated would be an interesting area for further research. We recommend that in practice the regression estimates in Step 2 of Algorithm D or Step 2* of Algorithm P be accompanied by standard model checking. Model inadequacy would indicate that the assumptions behind the algorithm are incorrect or that h_1 is too large. Although Theorem 5 only requires that N, the number of "observations" in the least-squares fitting, be 2, we suggest larger N to facilitate model checking.

Acknowledgements. We thank a referee and an associate editor for their helpful suggestions, and Sid Resnick for providing references to the literature on estimation of tail indices.

APPENDIX—Proofs (2.1), (2.2), and Theorems 1 and 3–7

Proof of (2.1) and (2.2). First

$$\begin{aligned}
\mathbf{E}\tilde{f}_x(x) &= \frac{1}{h}\int_0^1 K\left(\frac{x-y}{h}\right) f_x(y)dy = \frac{1}{h}\int_0^{h(1+C)} K\left(C-\frac{y}{h}\right) f_x(y)dy \\
&= \int_{-1}^{C} K(u) f_x(x-uh)du = \int_{-1}^{C} K(u)\{f_x(x) - f_x^{(1)}(x)uh + \frac{f_x^{(2)}(x)}{2}(uh)^2\}du + o(h^2) \\
&= f_x(x)\mu_0(C) - f_x^{(1)}(x)h\mu_1(C) + \frac{f_x^{(2)}(x)}{2}h^2\mu_2(C) + o(h^2), \qquad (A.1)
\end{aligned}$$

which proves (2.1). Next,

$$\begin{aligned}
\mathbf{E}\tilde{f}_x(-x) &= \frac{1}{h}\int_0^1 K\left(\frac{-x-y}{h}\right) f_x(y)dy = \frac{1}{h}\int_0^{h(1-C)} K\left(\frac{-Ch-y}{h}\right) f_x(y)dy \\
&= -\int_{-C}^{-1} K(u) f_x(-x-hu)du = \int_{-1}^{-C} K(u) f_x(x-(2C+u)h)du
\end{aligned}$$

$$= \int_{-1}^{-C} K(u)\{f_X(x) - f_X^{(1)}(x)h(2C+u) + \frac{f_X^{(2)}(x)}{2}h^2(2C+u)^2\}du + o(h^2)$$
$$= f_X(x)\mu_0(-C) - f_X^{(1)}(x)h[2C\mu_0(-C) + \mu_1(-C)]$$
$$+ f_X^{(2)}(x)h^2\left[2C^2\mu_0(-C) + 2C\mu_1(-C) + \frac{\mu_2(-C)}{2}\right] + o(h^2). \qquad (A.2)$$

(A.1) and (A.2) prove (2.2) since $\mu_0(C) + \mu_0(-C) = 1$, $\mu_1(C) = \mu_1(-C)$, and $\mu_2(C) + \mu_2(-C) = \mu_2(1)$.

Proof of Theorem 1. This proof makes use of empirical process results of Pollard (1984). Since it is similar to the proof of Lemma 5.2 in Ruppert and Cline (1991) we only sketch the argument. Let $g_0 = g_{\theta_0}$. Define

$$k_n(x'; g) = K\left\{\frac{g(x') - g(x_n)}{h}\right\} - K\left\{\frac{g_0(x') - g_0(x_n)}{h}\right\}$$

and $\mathcal{H}_n = \{k_n(\cdot; g) : g \in G_{\Delta_n}\}$. Because $g_0' > 0$ and $\text{supp}K \subset [-1, 1]$, there exist $M_1 > 0$ such that for all large n and all $g \in G_{\Delta_n}$

$$[\hat{f}_Y(g(x_n); g) - \mathbf{E}\hat{f}_Y(g(x_n); g)] - [\hat{f}_Y(g_0(x_n); g_0) - \mathbf{E}\hat{f}_Y(g_0(x_n); g_0)] =$$
$$\frac{1}{nh}\sum_{i=1}^{n}\Big\{[k_n(X_i; g)I\{|X_i - x_n| \le M_1 h\}] - [\mathbf{E}k_n(X_i; g)I\{|X_i - x_n| \le M_1 h\}]\Big\}.$$

By Condition 1, on the event $|X_1 - x_n| \le M_1 h$,

$$|g(X_1) - g(x_n) - g_0(X_1) + g_0(x_n)| = |g'(x_n)(X_1 - x_n) - g_0'(x_n)(X_1 - x_n)| + O(h^2)$$
$$= O(|g'(x_n) - g_0'(x_n)|\,|X_1 - x_n| + (h^2)) = O(\Delta_n h).$$

Therefore, since K is Lipschitz continuous there exists M_3, M_4 such that for all $g \in G_{\Delta_n}$

$$\mathbf{E}k_n^2(X_1; g) \le \frac{M_3}{h^2}\mathbf{E}\Big\{I\{|X_1 - x_n| \le M_1 h\}\Big(g(X_1) - g(x_n) - g_0(X_1) + g_0(x_n)\Big)^2\Big\} \le M_4 h\Delta_n^2.$$

It can be shown that the $L_1(F_X)$ covering numbers (Pollard 1984, p. 25) of $\mathcal{H}_n$ satisfy $N_1(\epsilon, \mathcal{H}_n) \le M_2\epsilon^{-k}$ for some M_2 that does not depend on n or ϵ—recall that $\dim\Theta = k$. We can apply Theorem 37 of Pollard (1984, p. 34) with $\delta_n = \sqrt{M_4 h}\Delta_n$ and, for some $\eta > 0$, $\alpha_n = \left((\log n)^{1+\eta}/(n\delta_n^2)\right)^{1/2}$ to obtain the following

$$\sup_{g \in G_{\Delta_n}}\Big|[\hat{f}_Y(g(x_n); g) - \mathbf{E}\hat{f}_Y(g(x_n); g)] - [\hat{f}_Y(g_0(x_n); g_0) - \mathbf{E}\hat{f}_Y(g_0(x_n); g_0)]\Big|$$
$$o(h^{-1}\delta_n^2\alpha_n) = o((nh)^{-1/2}(\log n)^{(1+\eta)/2}\Delta_n) = o((nh)^{-1/2}(\log n)^{-\epsilon'}).$$

for some $\epsilon' > 0$ if we take η small enough.

Proof of Theorem 3. This result is easily established by differentiating the Jacobian formula for transformed densities.

Proof of Theorem 4. Let $g_0 = g_{\theta_0}$ and $g_n = g_{\theta_n}$. We know that $f_Y^{(1)}(y; F_X) \equiv 0$ so by Theorem 3, $f_Y^{(1)}(y; g_0) = 0$ if $y = 0$ or 1. Since $f_X^{(2)}$ and $g_0^{(3)}$ are continuous, $f_Y^{(2)}$ is continuous and hence bounded on $[0,1]$. Therefore,

$$f_Y^{(1)}(y_n; g_0) = O(h) \text{ for } y_n = Ch \text{ or } y_n = (1 - Ch),\ C \in [0,1]. \tag{A.3}$$

It follows from (2.2), a similar result for the right boundary, (2.3), and (A.3) that

$$\mathbf{E}\hat{f}_Y(y_n; g_0) = f_Y(y_n; g_0) + O(n^{-2/5}) \quad \forall \text{ sequences } \{y_n\} \subset [0,1],$$

and therefore that

$$\mathbf{E}\hat{f}_X(x_n; g_0) = f_X(x_n) + O(n^{-2/5}) \quad \forall \text{ sequences } \{x_n\} \subset [0,1]. \tag{A.4}$$

By standard results

$$\hat{f}_Y(y_n; g_0) = \mathbf{E}\hat{f}_Y(y_n; g_0) + O_P(n^{-2/5}) \quad \forall \text{ sequences } \{y_n\} \subset [0,1], \tag{A.5}$$

and so by (A.4) $\hat{f}_X(x_n; g_0) = f_X(x_n) + O_P(n^{-2/5})$ for all sequences $\{x_n\} \subset [0,1]$. This proves the first assertion of the theorem. To simplify notation, for any random transformation g_n, let $\mathbf{E}\hat{f}_Y(y; g_n) = [\mathbf{E}\hat{f}_Y(y;\ g)]\big|_{g=g_n}$ and define $\mathbf{E}\hat{f}_X(x; g_n)$ similarly. If (2.3) holds, then by (A.3) and Condition 1, $f_Y^{(1)}(y_n; g_n) = O_P(h)\ \forall\ \{y_n\} \subset [0,h) \cup (1-h, 1]$, so by (2.2)

$$\mathbf{E}\hat{f}_X(x_n; g_n) = f_X(x_n) + O_P(n^{-2/5}) \quad \forall\ \{x_n\} \subset [0,1]. \tag{A.6}$$

By (2.1), (A.5), and Condition 1 applied to $g_n^{(1)}$

$$\begin{aligned} \hat{f}_X(x_n; g_n) - \mathbf{E}\hat{f}_X(x_n; g_n) &= \hat{f}_X(x_n; g_0) - \mathbf{E}\hat{f}_X(x_n; g_0) + o_P(n^{-2/5}) \\ &= O_P(n^{-2/5}) \quad \forall\ \{x_n\} \subset [0,1]. \end{aligned} \tag{A.7}$$

By (A.6) and (A.7), (2.4) holds.

Proof of Theorem 5. The result follows from (2.4) if we verify the conditions of Theorem 4. Note that G here is G_4 and satisfies Condition 1 for all x. We can parametrize G by defining $g_\theta(x) = \sum_{k=0}^{3} a_k x^k$, where $\theta = (a_0, \ldots, a_3)$. θ_0 satisfies

$$\frac{\int_0^x p_0(u)\,du}{\int_0^1 p_0(u)\,du} = g_0(x) \tag{A.8}$$

and θ_n satisfies $g_{\theta_n} = g_n$, where g_n is defined by Step 4. To complete the proof, it suffices to show that $\theta_n = \theta_0 + O_P(n^{-1/5})$. Suppose that the local least-squares fits satisfy

$$(b_0^*, b_1^*, b_2^*, b_3^*) = (f_x(0), f_x^{(1)}(0), f_x(1), f_x^{(1)}(1)) + O_P(n^{-1/5}). \tag{A.9}$$

Then by Condition 3, $Pr\{b_k = b_k^*, k = 0, \ldots, 3 \text{and} p_n^* = p_n\} \to 1$, and then comparing g_n to (A.8), we see that $\theta_n = \theta_0 + O_P(n^{-1/5})$.

Thus, it suffices to prove (A.9). We shall prove that

$$(b_0^*, b_1^*) = (f_x(0), f_x^{(1)}(0)) + O_P(h) \tag{A.10}$$

since the result for (b_2^*, b_3^*) is similar. The design matrix for the simple linear regression least squares problem in (3.1) is

$$\mathbf{X} = \begin{pmatrix} 1 & (1-1/2)/m \\ \vdots & \vdots \\ 1 & (N-1/2)/m \end{pmatrix}.$$

We will consider the cases $N \to \infty$ and N fixed. (Other cases are not very interesting, but could be treated by similar arguments.) If $N \to \infty$, then as $n \to \infty$

$$\mathbf{X}^T\mathbf{X} \sim \begin{pmatrix} N & N^2/2m \\ N^2/2m & N^3/3m^2 \end{pmatrix}$$

and

$$\begin{aligned} (\mathbf{X}^T\mathbf{X})^{-1} &\sim [N^4/3m^2 - N^4/4m^2]^{-1} \begin{pmatrix} N^3/3m^2 & -N^2/2m \\ -N^2/2m & N \end{pmatrix} \\ &= 12 \begin{pmatrix} 1/3N & -m/2N^2 \\ -m/2N^2 & m^2/N^3 \end{pmatrix}. \end{aligned} \tag{A.11}$$

If $N \geq 2$ is fixed, then X^TX is a fixed nonsingular matrix. The remainder of the proof holds for either N is fixed or $N \to \infty$. Note that $\text{Var}(\frac{md_i}{n}) \sim \frac{m}{n} f_x(0)$, so it follows from (A.11) that

$$\text{Var}(b_k^*) = O\left(\frac{m^3}{N^3 n}\right) = O(n^{-2/5}) \text{ for } k = 0, 1. \tag{A.12}$$

$\mathbf{E}\left(\frac{md_i}{n}\right) = f_x(z_i) + O(h_1^2) = f_x(0) + f_x^{(1)}(0)z_i + f_x^{(2)}(\xi_i)z_i^2 + O(h_1^2) = f_x(0) + f_x^{(1)}(0)z_i + u_i$, say. Let $u = (u_1, \ldots, u_N)^T$ and note that $\max_{i \leq N} |u_i| = O(h_1^2)$ since $f_x^{(2)}$ is bounded. Therefore,

$$\mathbf{E}\begin{pmatrix} b_0^* \\ b_1^* \end{pmatrix} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\left[\mathbf{X}\begin{pmatrix} f_x(0) \\ f_x^{(1)}(0) \end{pmatrix} + u\right].$$

Since, $\mathbf{X}^T u = O((Nh_1^2 \quad N^2h_1^2m^{-1})^T)$,

$$\mathbf{E}b_0^* - f_x(0) = O(h_1^2) \quad \text{and} \quad \mathbf{E}b_1^* - f_x^{(1)}(0) = O(NN^{-1}mh_1^2) = O(h_1). \tag{A.13}$$

By (A.12) and (A.13), (A.10) holds.

Proof of Theorem 6. First,

$$\begin{aligned}
\mathbf{E}\hat{f}_x(x;g) &= \frac{g'(x)}{2h}\int\left\{K\left(\frac{g(y)-g(x)}{h}\right)+K\left(\frac{g(y)+g(x)}{h}\right)\right\}f_x(y)dy \\
&= \frac{g'(x)}{2h}\left\{\int_0^{(h/B+ch/B)^{1/(\xi+1)}} Ay^\alpha dy + \int_0^{(h/B-ch/B)^{1/(\xi+1)}} Ay^\alpha dy\right\} \\
&= \frac{g'(x)A}{2h(\alpha+1)}\left\{\left(\frac{h(1+c)}{B}\right)^{\frac{\alpha+1}{\xi+1}}+\left(\frac{h(1-c)}{B}\right)^{\frac{\alpha+1}{\xi+1}}\right\} \\
&= Ax^\xi h^{\frac{\alpha-\xi}{\xi+1}}K(\alpha,\xi,c,B), \qquad (A.14)
\end{aligned}$$

where

$$K(\alpha,\xi,c,B)=\left(\frac{\xi+1}{\alpha+1}\right)\frac{B}{2}\left[\left(\frac{1+c}{B}\right)^{\frac{\alpha+1}{\xi+1}}+\left(\frac{1-c}{B}\right)^{\frac{\alpha+1}{\xi+1}}\right].$$

Now $K(\alpha,\xi,c,B)$ is continuous in ξ and equals 1 when $\xi=\alpha$. so

$$h^{\frac{\alpha-\xi}{\xi+1}}=\left[1+\log(h)\left(\frac{\alpha-\xi}{\xi+1}\right)\right](1+o(1)),$$

and

$$Ax^\xi = f_x(x)[1+\log(x)(\xi-\alpha)](1+o(1)),$$

so by (A.14)

$$\begin{aligned}
\mathbf{E}\hat{f}_x(x;g) &= f_x(x)\left[1+\log(h)\left(\frac{\alpha-\xi}{\xi+1}\right)\right][1+\log(x)(\xi-\alpha)](1+o(1)) \\
&= f_x(x)\left[1+(\xi-\alpha)\left(\log(x)-\frac{\log(h)}{\xi+1}\right)\right](1+o(1)).
\end{aligned}$$

Proof of Theorem 7. In this proof we write $(\hat{A},\hat{\alpha})$ in place of $(\hat{A}_1,\hat{\alpha}_1)$. We write $u_n \approx v_n$ if $u_n = v_n(1+o_P(1))$. $(\hat{A},\hat{\alpha})$ solves

$$\sum_1^N\left\{z_j^{-\hat{\alpha}/2}\left(\frac{md_j}{n}-\hat{A}z_j^{\hat{\alpha}}\right)\right\}\begin{pmatrix} z_j^{\hat{\alpha}/2} \\ A(\log(z_j))z_j^{\hat{\alpha}/2}\end{pmatrix}=0,$$

and

$$\mathrm{Var}\left\{z_j^{-\alpha/2}\left(\frac{md_j}{n}-Az_j^\alpha\right)\right\}\approx\frac{m}{n}A.$$

Therefore,

$$\mathrm{Var}\begin{pmatrix}\hat{A}\\ \hat{\alpha}\end{pmatrix}\approx\Sigma^{-1},$$

where

$$\Sigma = \frac{n}{Am}\sum_{j=1}^{N}\begin{pmatrix} z_j^\alpha & A(\log z_j)z_j^\alpha \\ A(\log z_j)z_j^\alpha & A^2(\log z_j)^2 z_j^\alpha \end{pmatrix}.$$

Since, $N/m \approx h_1$

$$\sum_{j=1}^{N} z_j^\alpha \approx m\int_0^{h_1} x^\alpha dx \approx \frac{m}{\alpha+1}h_1^{\alpha+1},$$

$$\sum_{j=1}^{N}(\log(z_j))z_j^\alpha \approx m\int_0^{h_1}(\log(x))x^\alpha dx \approx \frac{m}{\alpha+1}h_1^{\alpha+1}\left(\log(h_1) - \frac{1}{\alpha+1}\right), \quad \text{and}$$

$$\sum_{j=1}^{N}(\log(z_j))^2 z_j^\alpha \approx m\int_0^{h_1}(\log(x))^2 x^\alpha dx \approx \frac{m}{\alpha+1}h_1^{\alpha+1}\left\{\left(\log(h_1) - \frac{1}{\alpha+1}\right)^2 + (\frac{1}{\alpha+1})^2\right\}.$$

Therefore, $\det(\frac{nh_1^{\alpha+1}}{A(\alpha+1)}\Sigma) \approx A^2/(\alpha+1)^2$, so that

$$\Sigma^{-1} \approx \frac{(\alpha+1)^3 m h_1^{-(\alpha+1)}}{An}\begin{pmatrix} A^2\left\{\left(\log(h_1) - (\frac{1}{\alpha+1})\right)^2 + (\frac{1}{\alpha+1})^2\right\} & -A\left(\log(h_1) - \frac{1}{\alpha+1}\right) \\ -A\left(\log(h_1) - \frac{1}{\alpha+1}\right) & 1 \end{pmatrix},$$

which completes the proof.

REFERENCES

Cleveland, W. (1979), "Robust locally weighted regression and smoothing of scatterplots," *Journal of the American Statistical Association*, **74** 829–836.

Cline, D. B. H., and Hart, J. D., "Kernel estimation of densities with discontinuities or discontinuous derivatives," *Statistics*, **22**, 69–84.

Devroye, L. and Györfi, L. (1985), *Nonparametric Density Estimation: The L_1 View.* Wiley, New York.

Diggle, P., and Marron, J. S. (1987), "Equivalence of smoothing parameter selectors in density and intensity estimation," *Journal of the American Statistical Association*, **83** 793–800.

Hall, P. G. "On estimating the endpoint of a distribution," *Annals of Statistics*, **10**, 556–569.

Müller, H-G. (1987), "Weighted local regression and kernel methods for nonparametric curve fitting," *Journal of the American Statistical Association*, **82** 231–238.

Pollard, D. (1984), *Convergence of Stochastic Processes.* Springer, Berlin.

Rice, J. (1984), "Boundary modifications for kernel regression," *Communications in Statistics, Series A*, **13** 893–900.

Ruppert, D., and Cline, D. (1991), "Transformation-kernel density estimation—bias reduction by empirical transformations," Manuscript.

Ruppert, D., and Wand, M. (1991), "Correcting for kurtosis in density estimation," *Australian Journal of Statistics* (to appear).

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

Schuster, E. F. (1985), "Incorporating support constraints into nonparametric estimators of densities," *Communications in Statistics, Part A, Theory and Methods*, **14** 1123–1136.

Smith, R. L. (1987), "Estimating tails of probability distributions," *Annals of Statistics*, **15**, 1174–1207.

van Eeden, C. (1985), "Mean integrated squared error of kernel estimators when the density and its derivatives are not necessarily continuous," *Annals of the Institute of Statistical Mathematics*, **37**, *Part A*, 461–472.

Wand, M., Marron, J. S., and Ruppert, D. (1991), "Transformations in density estimation," with discussion, *Journal of the American Statistical Association*, **86**, 343–361.

CAPTIONS

Figure 1. Density estimates (solid lines) for 8 samples of $n = 500$ observations from the parabolic density. The true density is a dotted line. For Algorithm D $h_1 = .4$.

Figure 2. Density estimates (solid lines) for 8 samples of $n = 500$ observations from the U^2 (uniform squared) density. The true density is a dotted line. For Algorithm D, $h_1 = .25$. For Algorithm P, $h_1 = .5$.

Figure 3. Density estimates (solid lines) for 8 samples of $n = 500$ observations from the mixture density. The true density is a dotted line. For Algorithm D, $h_1 = .2$. For Algorithm D-LB, $h_1 = .25$.

Figure 1:

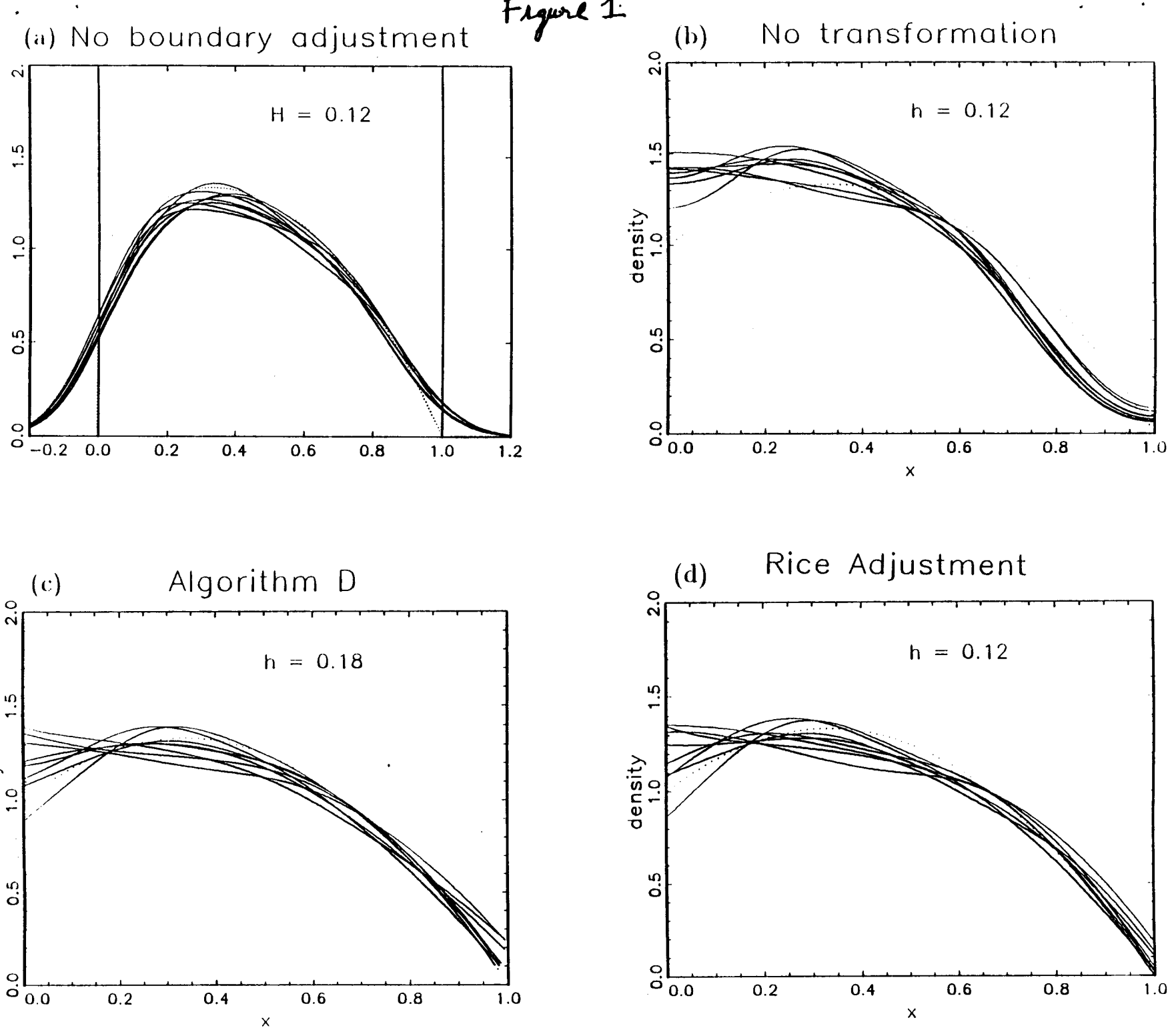

(a) No boundary adjustment
H = 0.12
(b) No transformation
h = 0.12
density
x
(c) Algorithm D
h = 0.18
density
x
(d) Rice Adjustment
h = 0.12
density
x

Figure 2

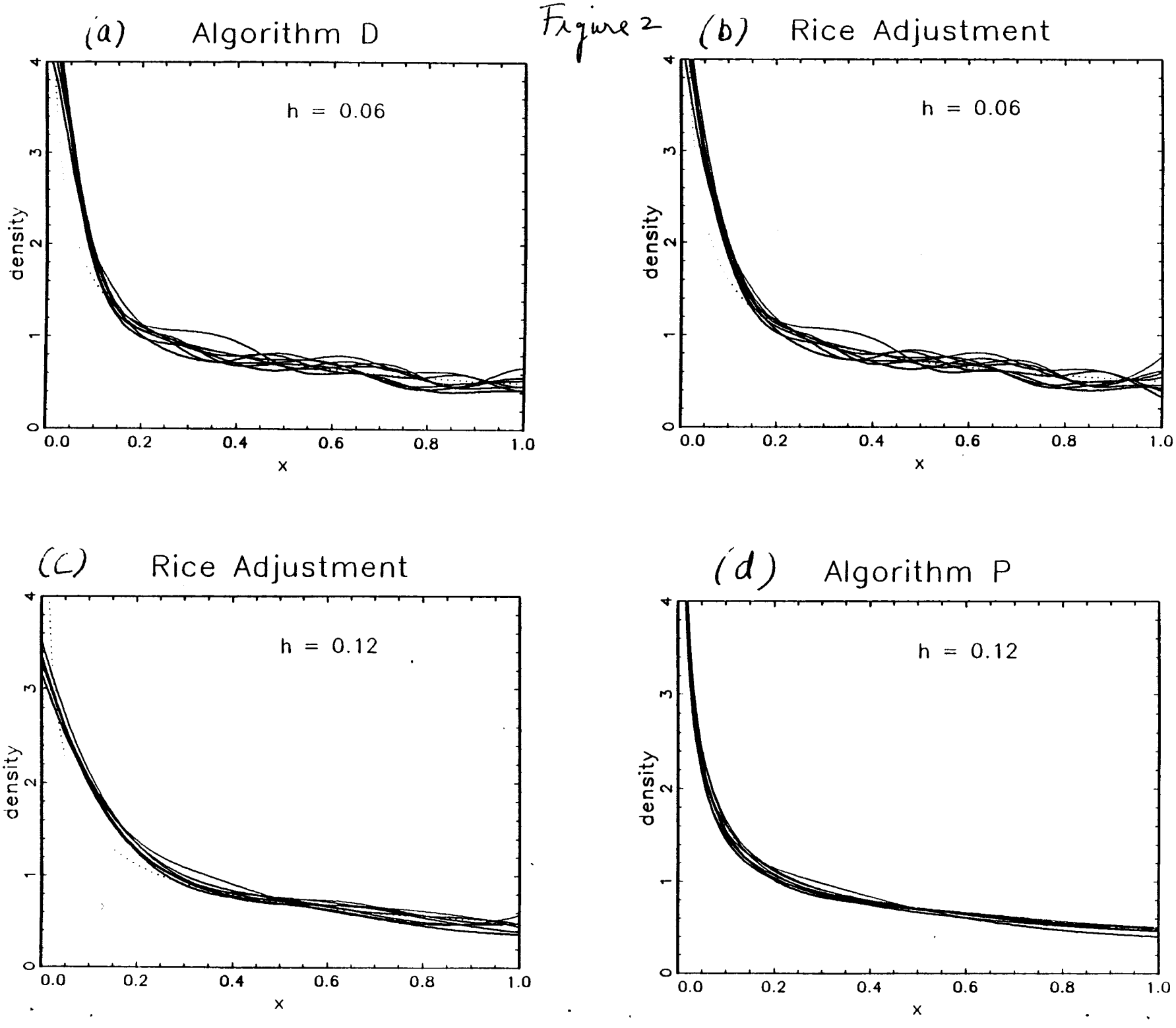
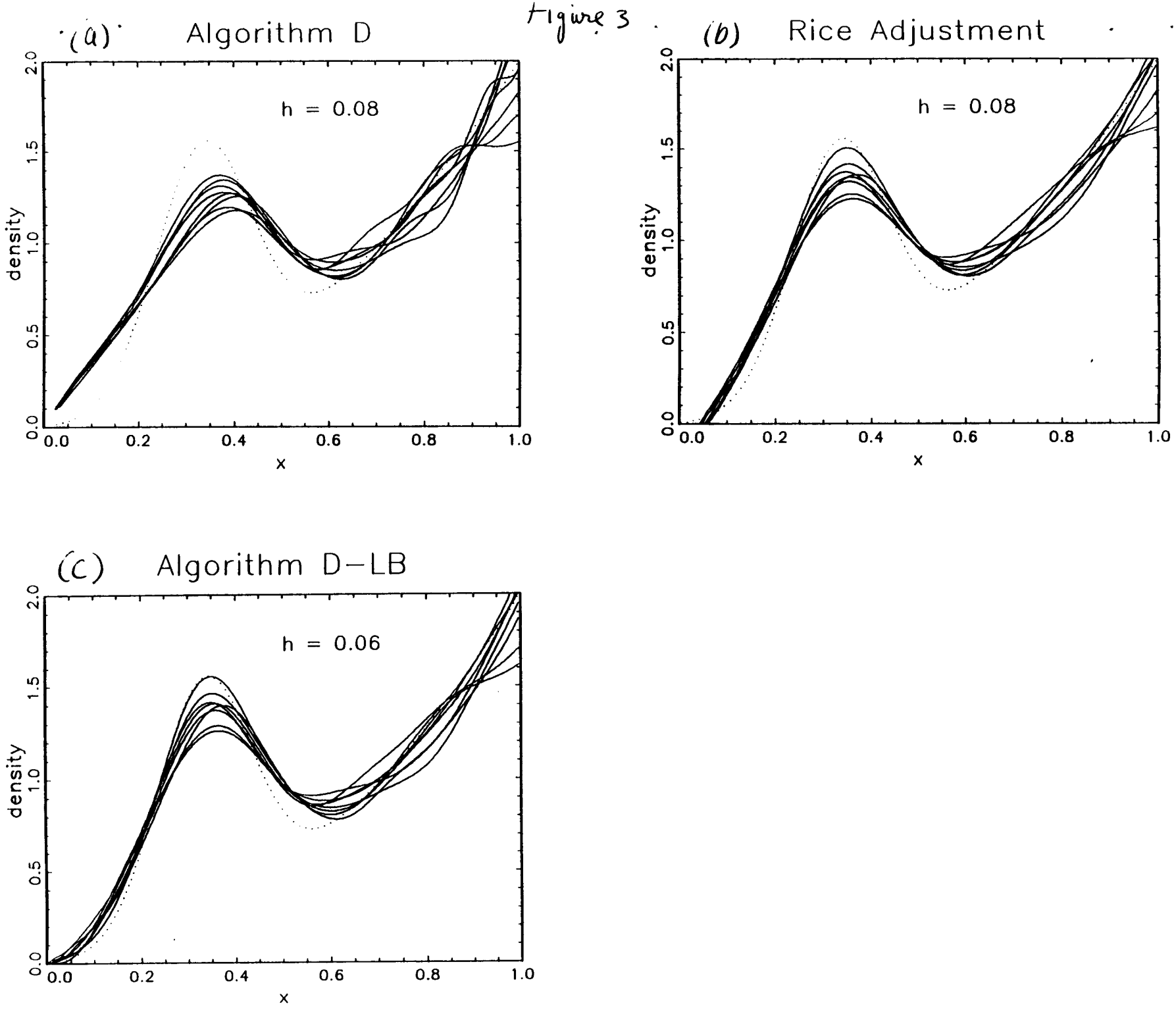

(a) Algorithm D
h = 0.06
density
x
(b) Rice Adjustment
h = 0.06
density
x
(c) Rice Adjustment
h = 0.12
density
x
(d) Algorithm P
h = 0.12
density
x

Figure 3


(a) Algorithm D
h = 0.08
density
x
(b) Rice Adjustment
h = 0.08
density
x
(c) Algorithm D−LB
h = 0.06
density
x