

## Reducing Bias in Nonparametric Estimation of the Distribution Function of Nonstandard Mixtures

E. D. McCune, Regents Professor, Stephen F. Austin State University, mccune@sfasu.edu

Sandra L. McCune, Regents Professor, Stephen F. Austin State University, smccune@sfasu.edu

### Abstract

Nonstandard mixtures occur when a random variable behaves in a continuous manner except at a countable number of discrete mass points. Polansky (2005) introduced a biased kernel estimator of the distribution function of nonstandard mixtures. In this paper, a new estimator of the distribution function of nonstandard mixtures with less bias than Polansky's (2005) estimator is obtained by applying to Polansky's (2005) estimator a nonparametric data transformation bias-reduction technique introduced by Swanepoel and Van Graan (2005). Statistical properties of the new estimator are determined and presented.

*Keywords:* kernel distribution function estimation; nonstandard mixtures; bias reduction

### 1. Introduction

In practice, nonstandard mixtures occur when a random variable behaves in a continuous manner except at a countable number of discrete mass points. For descriptions of applications to indicate the broad diversity of important situations in which these nonstandard mixtures arise, see the discussion by the Panel on Nonstandard Mixtures of Distributions (1990).

Polansky (2005) introduced a kernel estimator of the distribution function of nonstandard mixtures. In Sec. 2 a review of the asymptotic bias and variance properties of Polansky's (2005) estimator and a generalization of the mean integrated squared error (MISE) used by Polansky (2005) by introducing a weight function are presented. In Sec. 3 a new estimator of the distribution function of nonstandard mixtures is obtained by applying to Polansky's (2005) estimator a nonparametric data transformation bias-reduction technique, which was introduced by Swanepoel and Van Graan (2005). In Sec. 4 asymptotic results in terms of bias reduction and MISE are presented. In Sec. 5 a simulated data example and discussion of the results in terms of pointwise bias and pointwise mean squared error (MSE) are presented. In Sec. 6 topics for further research are discussed.

### 2. Preliminaries

Let  $F$  be a distribution function of a nonstandard mixture random variable  $X$  expressed as

$$F(x) = \alpha F_1(x) + (1 - \alpha) F_2(x) \quad (1)$$

where for all  $x \in R$ ,

(i)  $F_1(x) = F_d(x)/\alpha$  with  $F_d$  being a nondecreasing step function having a countable set, denoted by  $A$ , of jump points;

(ii)  $F_2(x) = F_c(x)/(1 - \alpha)$  with  $F_c$  being a positive, increasing, continuous everywhere function; and

(iii)  $\alpha = \lim_{x \rightarrow \infty} F_d(x)$ .

Also, let  $\Omega$  be the indicator function defined by

$$\Omega(x; B) = \begin{cases} 1 & \text{if } x \in B, \\ 0 & \text{if } x \notin B. \end{cases} \quad (2)$$

Let  $X_1, X_2, \dots, X_n$  be a set of independent and identically distributed random variables with distribution function  $F$ . Polansky (2005) introduced a nonparametric estimator of  $F$  defined as

$$\hat{F}(x; h) = \hat{\alpha} \hat{F}_1(x) + (1 - \hat{\alpha}) \hat{F}_2(x; h) \quad (3)$$

with

$$\hat{\alpha} = n^{-1} \sum_{i=1}^n \Omega(X_i; A), \quad (4)$$

$$\hat{F}_1(x) = \left\{ \sum_{i=1}^n \Omega(X_i; A) \right\}^{-1} \left\{ \sum_{i=1}^n \Omega(x; [X_i, \infty)) \Omega(X_i; A) \right\}, \quad (5)$$

and

$$\hat{F}_2(x; h) = \left\{ \sum_{i=1}^n \Omega(X_i; R \setminus A) \right\}^{-1} \left[ \sum_{i=1}^n K\{(x - X_i)/h\} \Omega(X_i; R \setminus A) \right] \quad (6)$$

where  $R \setminus A$  denotes the set of elements in  $R$  that are not in  $A$ , and  $K$  is a known distribution function having density  $k$  that is continuous on  $R$  and symmetric about zero such that  $\mu_0(k) = 1$ ,  $\mu_1(k) = 0$ , and  $0 < \mu_2 < \infty$  where

$$\mu_l(k) = \int_{-\infty}^{\infty} t^l k(t) dt. \quad (7)$$

Also, in this paper assume  $k$  is supported on the interval  $[-1, 1]$  and has (with respect to Lebesgue measure) a bounded almost everywhere continuous first derivative. The smoothing parameter  $h$ , called the bandwidth, is assumed to be a function of  $n$  such that as  $n \rightarrow \infty$ ,  $h \rightarrow 0$ , and  $nh \rightarrow \infty$ .

A typical measure of accuracy of an estimator  $F^*$  of a distribution function  $F$  is the mean integrated squared error (MISE), defined as

$$\text{MISE}(F^*) = E \left[ \int_{-\infty}^{\infty} \{F^*(x) - F(x)\}^2 w(x) dF(x) \right] \quad (8)$$

where  $w(\cdot)$  is a weight function (See Swanepoel & Van Graan, 2005). Asymptotic expansions for Eq. (8) were derived by Swanepoel (1988) using  $w(x) = 1$ , by Jones (1990) using  $w(x) = 1/f(x)$ , and by Altman and Léger (1995) using a general weight function  $w(x)$ .

Combining (4), (5), and (6), Polansky (2005) expressed Eq. (3) as

$$\hat{F}(x; h) = n^{-1} \sum_{i=1}^n \hat{\Omega}_i(x; A) \quad (9)$$

where

$$\hat{\Omega}_i(x; A) = \begin{cases} \Omega(x; [X_i, \infty)) & \text{when } X_i \in A, \\ K\{(x - X_i)/h\} & \text{when } X_i \in R \setminus A. \end{cases} \quad (10)$$

Under the conditions that  $F_2$  is differentiable everywhere and that  $F_2'$  is continuous and differentiable with a finite mean and has a square integrable derivative, Polansky (2005) proved that the following holds as  $n \rightarrow \infty$ ,

$$E[\hat{F}(x; h)] = F(x) + (1 - \alpha) \frac{h^2 \mu_2(k)}{2} \{F_2^{(2)}(x)\} + o(h^2) \quad (11)$$

and

$$\text{Var} \left[ \hat{F}(x; h) \right] = n^{-1} \left[ F(x) \{1 - F(x)\} - h(1 - \alpha) F_2'(x) C_1 \right] + o(n^{-1}h) \quad (12)$$

where  $\mu_2(k)$  is defined as in Eq. (7) and

$$C_1 = \int_{-1}^1 K(x) \{1 - K(x)\} dx. \quad (13)$$

Applying results from Altman and Léger (1995) to Eqs. (8), (11), and (12), yields

$$\text{MISE} \left[ \hat{F} \right] = n^{-1} D_1 - n^{-1} h (1 - \alpha) C_1 D_2 + \frac{h^4 (1 - \alpha)^2 C_2}{4} + o(h^4 + n^{-1}h) \quad (14)$$

with  $C_1$  defined as in Eq. (13),

$$D_1 = \int_{-\infty}^{\infty} F(x) (1 - F(x)) w(x) dF(x), \quad (15)$$

$$D_2 = \int_{-\infty}^{\infty} F_2'(x) w(x) dF(x), \quad (16)$$

and

$$C_2 = \mu_2^2(k) \int_{-\infty}^{\infty} \left\{ F_2^{(2)}(x) \right\}^2 w(x) dF(x) \quad (17)$$

provided  $D_1 < \infty$ ,  $D_2 < \infty$ , and  $C_2 < \infty$ . When  $w(x) = 1$ , Eq. (14) reduces to the expression for  $\text{MISE} \left[ \hat{F} \right]$  given by Polansky (2005) for the condition that  $k$  is supported on the interval  $[-1, 1]$ .

Asymptotically minimizing Eq. (14) with respect to  $h$  yields the optimal value of the bandwidth to be

$$\hat{h}_0 = \left\{ \frac{C_1 D_2}{C_2} \right\}^{\frac{1}{3}} [n(1 - \alpha)]^{\frac{1}{3}} \quad (18)$$

and evaluating  $\text{MISE} \left[ \hat{F} \right]$  in Eq. (14) using  $h = \hat{h}_0$  yields

$$\text{MISE}_0 \left[ \hat{F} \right] = n^{-1} D_1 - \left[ 3(C_1 D_2)^{\frac{4}{3}} C_2^{-\frac{1}{3}} (1 - \alpha)^{\frac{2}{3}} n^{-\frac{4}{3}} \right] / 4 + o \left( n^{-\frac{4}{3}} \right). \quad (19)$$

### 3. New estimator

Consider a new estimator of the nonstandard distribution function  $F$  defined as

$$\tilde{F}(x; h, g) = \hat{\alpha} \hat{F}_1(x) + (1 - \hat{\alpha}) \tilde{F}_2(x; h, g) \quad (20)$$

with  $\hat{\alpha}$  and  $\hat{F}_1$  defined as in Eqs. (4) and (5), respectively, and

$$\tilde{F}_2(x; h, g) = \left\{ \sum_{i=1}^n \Omega(X_i; R \setminus A) \right\}^{-1} \left[ \sum_{i=1}^n K \left( \frac{\hat{F}_2(x; g) - \hat{F}_2(X_i; g)}{h} \right) \Omega(X_i; R \setminus A) \right] \quad (21)$$

where  $\hat{F}_2$  is defined as in Eq. (6) and  $g$  is a bandwidth that might be different from  $h$ .

Note that  $\tilde{F}_2$ , which estimates the continuous part of the mixture distribution function  $F$ , is the estimator of an absolutely continuous distribution function as introduced by Swanepoel and Van Graan (2005). In the present case, the estimate is computed only on the sample values in  $R \setminus A$ . Combining Eqs. (4), (5) and (21) yields

$$\tilde{F}(x; h, g) = n^{-1} \sum_{i=1}^n \tilde{\Omega}_\tau(x; A) \quad (22)$$

where

$$\tilde{\Omega}_\tau(x; A) = \begin{cases} \Omega(x; [X_i, \infty)) & \text{when } X_i \in A, \\ K \left( \frac{\hat{F}_2(x; g) - \hat{F}_2(X_i; g)}{h} \right) & \text{when } X_i \in R \setminus A. \end{cases} \quad (23)$$

Theorem 1

Assume  $F_2$  is four times continuously differentiable in a neighborhood of  $x$  and that  $F_2'(x) > 0$ . Also, assume  $(g/h) \rightarrow c$  as  $n \rightarrow \infty$ , for some constant  $c$ ,  $0 \leq c < \infty$ . Then when  $\tilde{F}$  is the estimator in Eq. (22), it can be concluded that

(i) if  $ngh^3 \rightarrow \infty$  and  $ng^2 \rightarrow \infty$  as  $n \rightarrow \infty$ ,

$$E[\tilde{F}(x; h, g)] = F(x) + (1-\alpha) \frac{g^2 h^2 \mu_2^2(k)}{4} \left[ \frac{F_2^{(2)}(x) F_2^{(3)}(x)}{(F_2'(x))^3} - \frac{F_2^{(4)}(x)}{(F_2'(x))^2} \right] + o(g^2 h^2) \quad (24)$$

(ii) if  $ngh^3 \rightarrow \infty$  and  $(h^3/g) \rightarrow 0$  as  $n \rightarrow \infty$ ,

$$\text{Var}[\tilde{F}(x; h, g)] = n^{-1} [F(x)\{1-F(x)\} - h(1-\alpha)C_1] + o(n^{-1}h) \quad (25)$$

where  $C_1$  is defined by Eq. (13).

Proof: Motivated by the proof of Eqs. (11) and (12), presented by Polansky (2005), let  $X = WD + (1-W)C \sim F$ , where  $D$ ,  $C$ , and  $W$  are mutually independent random variables such that

$$W = \begin{cases} 1 & \text{with probability } \alpha, \\ 0 & \text{with probability } 1-\alpha \end{cases}$$

with  $D \sim F_1$  and  $C \sim F_2$ . To establish the proof of part (i),

$$\begin{aligned} E[\tilde{F}(x; h, g)] &= E[\tilde{\Omega}_\tau(x; A)] \\ &= \alpha E[\tilde{\Omega}_\tau(x; A)|W=1] + (1-\alpha) E[\tilde{\Omega}_\tau(x; A)|W=0] \\ &= \alpha E[\hat{F}_1(x)] + (1-\alpha) E[\hat{F}_2(x)] \\ &= \alpha F_1(x) + (1-\alpha) \left\{ F_2(x) + \frac{g^2 h^2 \mu_2^2(k)}{4} \left[ \frac{F_2^{(2)}(x) F_2^{(3)}(x)}{(F_2'(x))^3} - \frac{F_2^{(4)}(x)}{(F_2'(x))^2} \right] \right\} \\ &= F_d(x) + F_c(x) + (1-\alpha) \frac{g^2 h^2 \mu_2^2(k)}{4} \left[ \frac{F_2^{(2)}(x) F_2^{(3)}(x)}{(F_2'(x))^3} - \frac{F_2^{(4)}(x)}{(F_2'(x))^2} \right] + o(g^2 h^2) \\ &= F(x) + (1-\alpha) \frac{g^2 h^2 \mu_2^2(k)}{4} \left[ \frac{F_2^{(2)}(x) F_2^{(3)}(x)}{(F_2'(x))^3} - \frac{F_2^{(4)}(x)}{(F_2'(x))^2} \right] + o(g^2 h^2). \end{aligned} \quad (26)$$

For the variance, note that

$$\text{Var}[\tilde{F}(x; h, g)] = n^{-1} \left\{ E[\tilde{\Omega}_i^2(x; A)] - E^2[\tilde{\Omega}_i(x; A)] \right\} \quad (27)$$

where

$$\begin{aligned} E[\tilde{\Omega}_i^2(x; A)] &= \alpha E[\tilde{\Omega}_i^2(x; A)|W=1] + (1-\alpha) E[\tilde{\Omega}_i^2(x; A)|W=0] \\ &= \alpha F_1(x) + (1-\alpha)(F_2(x) - hC_1) + o(h) \\ &= \alpha F_1(x) + (1-\alpha)F_2(x) - h(1-\alpha)C_1 + o(h) \\ &= F_d(x) + F_c(x) - h(1-\alpha)C_1 + o(h) \\ &= F(x) - h(1-\alpha)C_1 + o(h) \end{aligned} \quad (28)$$

where  $C_1$  is defined as in Eq. (13). Now Eq. (26) implies that

$$E^2[\tilde{\Omega}_i(x; A)] = (F(x))^2 + o(gh). \quad (29)$$

Hence, from Eqs. (26), (28), and (29) it follows that

$$\text{Var}[\tilde{F}(x; h, g)] = n^{-1} [F(x)\{1-F(x)\} - h(1-\alpha)C_1] + o(n^{-1}h), \quad (30)$$

establishing the proof of part (ii).

#### Theorem 2

Under the assumptions of Theorem 3.1,

$$\text{MISE}[\tilde{F}] = n^{-1}\tilde{D}_1 - n^{-1}h(1-\alpha)C_1\tilde{D}_2 + \frac{(1-\alpha)^2 g^4 h^4 \tilde{C}_2}{16} + o(g^4 h^4 + n^{-1}h) \quad (31)$$

where

$$\tilde{D}_1 = D_1, \quad (32)$$

$$\tilde{D}_2 = \int_{-\infty}^{\infty} w(x) dF(x), \quad (33)$$

and

$$\tilde{C}_2 = \mu_2^4(k) \int_{-\infty}^{\infty} \left[ \frac{F_2^{(2)}(x)F_2^{(3)}(x)}{(F_2'(x))^3} - \frac{F_2^{(4)}(x)}{(F_2'(x))^2} \right]^2 w(x) dF(x) \quad (34)$$

provided that  $\tilde{D}_1 < \infty$ ,  $\tilde{D}_2 < \infty$ , and  $\tilde{C}_2 < \infty$ .

Proof: It follows from Eq. (26) that

$$\text{Bias}[\tilde{F}(x; h, g)] = (1-\alpha) \frac{g^2 h^2 \mu_2^2(k)}{4} \left[ \frac{F_2^{(2)}(x)F_2^{(3)}(x)}{(F_2'(x))^3} - \frac{F_2^{(4)}(x)}{(F_2'(x))^2} \right] + o(g^2 h^2). \quad (35)$$

Therefore,

$$\text{Bias}^2[\tilde{F}(x; h, g)] = (1-\alpha)^2 \frac{g^4 h^4 \mu_2^4(k)}{16} \left[ \frac{F_2^{(2)}(x)F_2^{(3)}(x)}{(F_2'(x))^3} - \frac{F_2^{(4)}(x)}{(F_2'(x))^2} \right]^2 + o(g^4 h^4). \quad (36)$$

The proof of Eq. (31) follows immediately by using Eqs. (25) and (36) and by applying Eq. (8).

#### 4. Asymptotic results

4.1 Bias comparison between the Polansky (2005) estimator  $\hat{F}(x; h)$  as defined in Eq. (9) and the new estimator  $\tilde{F}(x; h, g)$  as defined in Eq. (22)

Select

$$g = \beta h^\gamma \quad (37)$$

for some constant  $\beta$  such that  $0 < \beta < \infty$  and  $1 \leq \gamma < 3$  so that Eqs. (24) and (25) will hold. From Eq. (11),

$$\begin{aligned} \text{Bias}[\hat{F}(x; h)] &= (1-\alpha) \frac{h^2 \mu_2(k)}{2} (F^{(2)}(x)) + o(h^2) \\ &= O(h^2) \end{aligned} \quad (38)$$

and from Eq. (24)

$$\begin{aligned} \text{Bias}[\tilde{F}(x; h, \beta h^\gamma)] &= (1-\alpha) \frac{\beta^2 h^{2\gamma+2} \mu_2^2(k)}{4} \left[ \frac{F_2^{(2)}(x) F_2^{(3)}(x)}{(F_2'(x))^3} - \frac{F_2^{(4)}(x)}{(F_2'(x))^2} \right] + o(h^{2\gamma+2}) \\ &= O(h^{2\gamma+2}). \end{aligned} \quad (39)$$

Hence, from Eqs. (38) and (39), the bias of  $\tilde{F}(x; h, \beta h^\gamma) = O(h^{2\gamma+2})$ , which is less than the  $O(h^2)$  bias of the Polansky (2005) estimator,  $\hat{F}(x; h)$ .

4.2 MISE comparison between Polansky (2005) estimator  $\hat{F}$  as defined in Eq. (9) and the new estimator  $\tilde{F}$  as defined in Eq. (22)

Now, in Eq. (31) select  $g$  as in Eq. (37) obtaining

$$\text{MISE}[\tilde{F}] = n^{-1} \tilde{D}_1 - n^{-1} h (1-\alpha) C_1 \tilde{D}_2 + \frac{(1-\alpha)^2 \beta^4 h^{4\gamma+4} \tilde{C}_2}{16} + o(h^{4\gamma+4} + n^{-1} h). \quad (40)$$

Asymptotically minimizing Eq. (40) with respect to  $h$ , gives the optimal value as

$$\tilde{h}_0 = \left[ \frac{4C_1 \tilde{D}_2}{\beta^4 (\gamma+1) \tilde{C}_2} \right]^{\frac{1}{4\gamma+3}} \left( [n(1-\alpha)]^{\frac{-1}{4\gamma+3}} \right). \quad (41)$$

Using  $h = \tilde{h}_0$  in Eq. (41), the value of  $\text{MISE}[\tilde{F}]$  as given in Eq. (40) becomes

$$\begin{aligned} \text{MISE}[\tilde{F}] &= n^{-1} \tilde{D}_1 \\ &- \left\{ (1-\alpha) C_1 \tilde{D}_2 \left[ \frac{4C_1 \tilde{D}_2}{(1-\alpha)(\gamma+1)\beta^4 \tilde{C}_2} \right] + \frac{(1-\alpha)^2 \beta^4 \tilde{C}_2 (\gamma+1)}{8} \left[ \frac{4C_1 \tilde{D}_2}{(1-\alpha)(\gamma+1)\beta^4 \tilde{C}_2} \right]^{\frac{4\gamma+4}{4\gamma+3}} \right\} \left( n^{\frac{-(4\gamma+4)}{4\gamma+3}} \right) \\ &+ o\left( n^{\frac{-(4\gamma+4)}{4\gamma+3}} \right). \end{aligned} \quad (42)$$

One can observe from Eq. (19) that  $\text{MISE}_0[\hat{F}] = O(n^{-1})$  and from Eq. (42) that  $\text{MISE}_0[\tilde{F}] = O(n^{-1})$ , indicating MISE of the new estimator asymptotically converges at the same rate as the Polansky (2005) estimator.

#### 4.3 Remark

Note that, from an asymptotic perspective, the new estimator  $\tilde{F}$  has the improved quality of less bias than the Polansky (2005) estimator without producing the undesirable result of an increase in MISE.

### 5. Finite sample size results

#### 5.1 Example

Consider the example of estimating  $F$  when the random sample  $X_1, X_2, \dots, X_n$  is from a nonstandard mixture random variable  $X$  with distribution function

$$F(x) = (0.25)F_1(x) + (0.75)F_2(x) \quad (43)$$

where  $F_1$  is the discrete distribution function

$$F_1(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \quad (44)$$

and  $F_2$  is a Pareto continuous distribution function, see Hogg and Klugman (1984), expressed as

$$F_2(x) = \begin{cases} 1 - \frac{1}{(x+1)^3} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases} \quad (45)$$

To implement the estimators, choose the Epanechnikov kernel

$$k(z) = \begin{cases} \left(\frac{3}{4}\right)(1-z^2) & \text{if } |z| \leq 1 \\ 0 & \text{if } |z| > 1 \end{cases} \quad (46)$$

from which

$$K(z) = \int_{-\infty}^z k(t) dt = \begin{cases} 1 & \text{if } z > 1 \\ \frac{1}{2} + \frac{3}{4}z - \frac{z^3}{4} & \text{if } -1 \leq z \leq 1 \\ 0 & \text{if } z < -1. \end{cases} \quad (47)$$

As in Swanepoel (1988),  $w(x) = 1$  in Eq. (8).

For the Polansky estimator  $\hat{F}(x; h)$  defined in Eq. (9), replace the bandwidth  $h$  by the asymptotically optimal bandwidth  $h_0$  expressed in Eq. (18). Likewise, for the new estimator  $\tilde{F}(x; h, g)$  defined in Eq. (22), replace the bandwidth  $h$  by the asymptotically optimal value  $\tilde{h}_0$  expressed in Eq. (41); and for Eq. (37),  $g = \beta h^\gamma = \beta(\tilde{h}_0)^\gamma$ , select  $\beta = 2.5$  and  $\gamma = 1$ .

Table 1 displays pointwise bias and pointwise mean squared error (MSE) estimates for the two estimators (with asymptotically optimal bandwidths) based on a Monte Carlo simulation of 10,000 random samples generated using SAS/IML (See SAS Institute Inc., 2002) from  $F$  for each sample size  $n=50, 100,$  and  $400$ . The standard errors of the averages of the two estimators (with optimal bandwidths) from the 10,000 trials were less than  $6 \times 10^{-3}$  for  $n=50$ , less than  $4 \times 10^{-3}$  for  $n=100$ , and less than  $2 \times 10^{-3}$  for  $n=400$ .

$n$	$x$	$F(x)$	$\widehat{Bias}(\widehat{F})$	$\widetilde{Bias}(\widetilde{F})$	$\widehat{MSE}(\widehat{F})$	$\widetilde{MSE}(\widetilde{F})$
50	0	.250000	.083016	.005599	.010208	.003706
	0.5	.777778	-.007129	.000548	.003011	.003438
	1.0	.906250	-.001397	-.000252	.001510	.001658
	1.5	.952000	-.000552	-.000241	.000822	.000845
100	0	.250000	.068416	.005271	.006363	.001855
	0.5	.777778	-.004352	.000361	.001543	.001721
	1.0	.906250	-.001296	-.000210	.000776	.000835
	1.5	.952000	-.000527	-.000185	.000435	.000452
400	0	.250000	.046312	.005025	.002583	.000486
	0.5	.777778	-.002102	-.000300	.000398	.000424
	1.0	.906250	-.000508	-.000077	.000203	.000213
	1.5	.952000	-.000180	-.000043	.000110	.000113

Table 1: Bias Reduction with Asymptotically Optimal Bandwidths

Note that an examination of Table 1 reveals that, even for finite sample sizes,  $\widetilde{F}$  has the improved quality of less bias than the Polansky (2005) estimator without producing the undesirable result of appreciably increasing MSE. Unfortunately, the optimal bandwidths depend on  $F$  and would not be obtainable in practice; therefore, data-based choices for the bandwidths need to be derived.

A practical implementation of the Polansky estimator  $\widehat{F}(x; h)$  defined in Eq. (9) is achieved by deriving a simple plug-in normal reference bandwidth selector, a strategy supported in the literature (see, e.g., Altman & Léger, 1995; Bowman *et al.*, 1998). Replace  $F_2$  in Eq. (18) with an  $N(0, \sigma^2)$  reference distribution with  $\sigma^2$  replaced by an appropriate estimator,  $\widehat{\sigma}^2$ . Hence, with  $w(x)=1$  and  $K$  as in Eq. (46) the data-driven bandwidth is

$$\tilde{h} = 3.9\widehat{\sigma}[(n(1-\widehat{\alpha}))^{(-1/3)}] \quad (48)$$

where  $\widehat{\alpha}$  is defined as in Eq. (4). Similarly, for the new estimator  $\widetilde{F}(x; h, g)$  defined in Eq. (22), replace  $F_2$  in Eq. (41) with an  $N(0, \sigma^2)$  reference distribution except with weight function  $w(x) = [\phi(x/\sigma)/\sigma]^2$ , where  $\phi$  is the standard normal density (following the lead of Swanepoel and Van Graan, 2005). Thus, the data driven bandwidths are

$$\tilde{h} = \left[ \frac{375\sqrt{3}}{28\pi(\beta^4)} \right]^{(1/7)} \widehat{\sigma}^{(4/7)} [n(1-\widehat{\alpha})]^{(-1/7)} \quad (49)$$

and

$$g = \beta\tilde{h} \quad (50)$$

where, for this example,  $\beta = 2.5$  is chosen. In Eq. (48) and Eq. (49) use  $\hat{\sigma} = \min\{S, IQR/1.349\}$ , a measure suggested by Silverman (1986), with  $S$  being the sample standard deviation and  $IQR$  the interquartile range.

Table 2 displays pointwise bias and pointwise mean squared error (MSE) estimates for the two estimators (with data-driven bandwidths) based on a Monte Carlo simulation of 10,000 random samples generated using SAS/IML from  $F$  for each sample size  $n = 50, 100, \text{ and } 400$ .

$n$	$x$	$F(x)$	$\widehat{Bias}(\widehat{F})$	$\widetilde{Bias}(\widetilde{F})$	$\widehat{MSE}(\widehat{F})$	$\widetilde{MSE}(\widetilde{F})$
50	0	.250000	.122226	.006016	.018215	.003703
	0.5	.777778	-.022814	.000581	.003019	.003395
	1.0	.906250	-.005226	-.000343	.001470	.001689
	1.5	.952000	-.001754	-.000313	.000829	.000892
100	0	.250000	.103722	.005842	.012492	.001900
	0.5	.777778	-.014001	.000418	.001558	.001694
	1.0	.906250	-.003158	-.000241	.000776	.000862
	1.5	.952000	-.001202	-.000189	.000422	.000456
400	0	.250000	.072424	.005279	.005683	.000492
	0.5	.777778	-.005313	-.000344	.000405	.000432
	1.0	.906250	-.001272	-.000085	.000206	.000219
	1.5	.952000	-.000438	-.000053	.000118	.000115

Table 2: Bias Reduction with Data Driven Bandwidths

Again, the standard errors of the averages of the two estimators (with data-driven bandwidths) from the 10,000 trials were less than  $6 \times 10^{-3}$  for  $n = 50$ , less than  $4 \times 10^{-3}$  for  $n = 100$ , and less than  $2 \times 10^{-3}$  for  $n = 400$ . An examination of the values exhibited in Table 2 reveals that, even for finite sample sizes,  $\widetilde{F}$  appears to have a resounding improved quality of less bias than the Polansky (2005) estimator without an appreciable increase in MSE.

### 5.2. Remarks

In this paper no attempt has been made to do an exhaustive Monte Carlo simulation of the effectiveness of the new estimator. It is hoped that the example that has been included will be sufficient to convince the reader that the new estimator warrants further study as it applies to specific problems.

## 6. Further research

A study comparing bandwidth estimation methods, such as the ones proposed by Bowman et al. (1998) and Polansky and Baker (2000), in terms of which is best suited for the new estimator, should be of interest. In this regard, it should be noted that in his seminal paper on density function estimation, Parzen (1962) revealed that the kernel method he proposed was inspired by the problem of estimating the spectral density function of a stationary time series. Motivated by this disclosure, it would seem worthwhile to investigate whether the proposed adaptive bandwidth selection method for spectral estimation proposed by DiRienzo and Zurbenko (1999) could be modified for use in bandwidth estimation for the proposed new estimator.

## Acknowledgments

This research was supported by a Stephen F. Austin State University Faculty Research Grant. Excellent programming assistance was provided by Shu Li and Lei Sun, graduate assistants. The authors wish to thank the anonymous reviewers for their suggestions and comments that strengthened the paper.

## References

- Altman, N., Leger, C., 1995. Bandwidth selection for kernel distribution function estimation. *J. Statist. Plann. Infer.* 46: 195-214.
- Bowman, A., Hall P., Prvan, T. 1998. Bandwidth selection for the smoothing of distribution functions. *Biometrika.* 85:799-808.
- DiRienzo, G., Zurbenko, I. 1999. Semi-Adaptive nonparametric spectral estimation. *J. Computat. Graph. Statist.* 8: 41-59.
- Hogg, R. V. , Klugman, S. A., 1984. *Loss Distributions.* J. Wiley and Sons, Inc.
- Jones, M. C., 1990. The performance of kernel density functions in kernel distribution function estimation. *Statist. Probab. Lett.* 9: 129-132.
- Panel on Nonstandard Mixtures of Distributions, 1990. Statistical models and analysis in auditing. *Statist. Sci.* 4:2-33.
- Parzen, E. 1962. On estimation of a probability density function and its mode. *Ann. Math. Statist.* 33, 1065-1076.
- Polansky, A. M., 2005. Nonparametric estimation of distribution functions of nonstandard mixtures. *Commun. Statist. Theor. Meth.* 34: 1711-1724.
- Polansky, A. M., Baker E. R. 2000. Multistage plug-in bandwidth selection for kernel distribution function estimates. *J. Statist. Computat. Simul.* 65: 63-80.
- SAS Institute Inc. 2002. *SAS/IML Software: Usage and Reference, Version 9.1*, Cary, NC: SAS Institute Inc.
- Silverman, B. W. 1986. *Density Estimation for Statistics and Data Analysis.* London: Chapman and Hall.
- Swanepoel, J. W. H., 1988. Mean integrated squared error properties and optimal kernels when estimating a distribution function. *Commun. Statist. Theor. Meth.* 17:3785-3799.
- Swanepoel, J. W. H., Van Graan, F. C., 2005. A new kernel distribution function estimator based on a nonparametric transformation of the data. *Scan. J. Statist.* 32:551-562.