*Austral. J. Statist.*, **34**(1), 1992, 19–29

# CORRECTING FOR KURTOSIS IN DENSITY ESTIMATION

D. RUPPERT[1] AND M.P. WAND[2]

*Cornell University and Rice University*

**Summary**

Using a global window width kernel estimator to estimate an approximately symmetric probability density with high kurtosis usually leads to poor estimation because good estimation of the peak of the distribution leads to unsatisfactory estimation of the tails and vice versa. The technique proposed corrects for kurtosis via a transformation of the data before using a global window width kernel estimator. The transformation depends on a "generalised smoothing parameter" consisting of two real-valued parameters and a window width parameter which can be selected either by a simple graphical method or, for a completely data-driven implementation, by minimising an estimate of mean integrated squared error. Examples of real and simulated data demonstrate the effectiveness of this approach, which appears suitable for a wide range of symmetric, unimodal densities. Its performance is similar to ordinary kernel estimation in situations where the latter is effective, e.g. Gaussian densities. For densities like the Cauchy where ordinary kernel estimation is not satisfactory, our methodology offers a substantial improvement.

*Key words:* Convex-concave transformations; kernel estimators; nonparametric density estimation; window width selection.

## 1. Introduction

Consider the problem of estimating an approximately symmetric probability density function $f_X$ from a real-valued random sample $X_1, \ldots, X_n$. The global window width kernel estimator for $f_X$ is

$$\hat{f}_X(x) = n^{-1}h^{-1}\sum_{i=1}^{n} K\{(x - X_i)/h\} \tag{1.1}$$

where the kernel $K$ is a symmetric density and the window width $h$ is a positive number. Silverman (1986) provides a detailed account of this estimator and its applications.

Received June 1990; revised March 1991.

[1]Operations Research & Industrial Engineering, ETC Bldg, Cornell University, Ithaca, NY 14853, U.S.A.

[2]Dept. Statistics, P.O. Box 1892, Houston, TX 77251-1892, U.S.A.

*Acknowledgements.* The work of D. Ruppert was supported by NSF grant DMS-8800294 and by the Army Research Office through the Mathematical Sciences Institute at Cornell University.

The kurtosis of a probability density function $f_X$ can be described in terms of its "peakedness" in the centre and "heaviness" in the tails. It is easier to define *relative* kurtosis than to define kurtosis itself. The support of a density can be divided into the centre, the tails and the "shoulders" which lie between the centre and tails. A transfer of mass from the shoulders to both the centre and the tails is said to increase kurtosis, and a density with a high kurtosis generally has a sharp peak in the centre and long tails.

The presence of high kurtosis has an adverse effect on the performance of (1.1) since the amount of smoothing is uniform across the sample space. For a heavily kurtotic density like the Cauchy, good estimation of the sharp peak requires a relatively small window width which is very likely to induce artificial "bumps" in the long tails. If a larger window width is chosen to smooth out these bumps then the peak will almost certainly be oversmoothed or "missed". To overcome this problem we propose applying the transformation $g_\lambda$ to the data to obtain $Y_1, \ldots, Y_n$ with common density $f_Y(\cdot\,;\lambda)$. The parameter $\lambda$ lies in some finite-dimensional set $\Lambda$. The immediate goal of the transformation is to reduce the kurtosis of $f_Y(\cdot\,;\lambda)$ but the ultimate goal is a density that is easy to estimate using (1.1). The back-transformation by change of variables from $f_Y(\cdot\,;\lambda)$ to $\hat{f}_X(\cdot\,;\lambda)$ is our proposed estimate. We show that the transformation/kernel estimator at a point $x$ can be viewed as a kernel density estimator with a locally adaptive window width which is approximately proportional to $g'_\lambda(x)^{-1}$. Therefore, for estimation of a highly kurtotic density one may choose the window width to estimate the peak well and then choose the transformation parameters to effectively estimate other features of the density such as the shoulder region and the tails. Our proposed estimator may be viewed as an alternative to the variable window width kernel estimator (Breiman *et al.*, 1977; Abramson, 1982; Silverman, 1986, Section 5.3) which permits the window width to vary at each sample point $X_i$ proportionally to $f_X(X_i)^{-\gamma}$ for some $0 < \gamma \leq 1$. However, for effective implementation, their approach requires the specification of a pilot estimator for $f_X$ itself.

The transformation algorithm we use was also employed by Wand, Marron & Ruppert (1991) where the estimation of nonnegative skewed densities was the main concern. Other work on transformations in density estimation may be found in Silverman (1986, Sections 2.9 and 5.3.5) and Devroye & Györfi (1985, Chapter 9).

Section 2 covers the theory of transformation/kernel density estimation. In Section 3 a suitable two-parameter family of transformations is introduced with the property of reducing the kurtosis of symmetric distributions. This family is reparametrized to allow simple interpretation of the influence of the transformation on the effective window width of the final estimator. Section 4 deals with the implementation of our technique including a simple graphical method and a data-driven choice of the transformation parameters and window width. The latter chooses the transformation parameters and window width,

which together constitute a "generalised smoothing parameter", by minimising an estimate of mean integrated squared error. Examples are presented in the final section.

## 2. Transformation-based Kernel Density Estimators

The transformation technique for density estimation is discussed extensively in Wand *et al.* (1991). In this section we briefly cover the most important ideas.

Let $X$ be a random variable having density $f_X$, and $\{\tilde{g}_\lambda : \lambda \in \Lambda\}$ some parametric family of increasing transformations defined on the support of $f_X$. Put $\tilde{Y} = \tilde{g}_\lambda(X)$. To preserve the scale we take our transformation to be $g_\lambda = (s_X/s_{\tilde{Y}})\tilde{g}_\lambda$ where $s_X$ and $s_{\tilde{Y}}$ are scale estimates of the distribution of $X$ and $\tilde{Y}$ respectively. Our estimate of $f_Y(y;\lambda) = f_X\{g_\lambda^{-1}(y)\}(g_\lambda^{-1})'(y)$ is the usual global window width kernel estimator

$$\hat{f}_Y(y) = n^{-1}h^{-1}\sum_{i=1}^{n} K\{(y - Y_i)/h\}. \tag{2.1}$$

The transformation/kernel density estimator is the back-transform of (2.1):

$$\hat{f}_X(x;h,\lambda) = n^{-1}h^{-1}\sum_{i=1}^{n} g'_\lambda(x)K[\{g_\lambda(x) - g_\lambda(X_i)\}/h]. \tag{2.2}$$

Note that, from the mean value theorem,

$$\hat{f}_X(x;h,\lambda) = n^{-1}\{h/g'_\lambda(x)\}^{-1}\sum_{i=1}^{n} K[(x - X_i)/\{h/g'_\lambda(\xi_i)\}]$$

where $\xi_i$ lies between $x$ and $X_i$. Comparing this formulation with (1.1) we see that the "window width" at $x$ is approximately $h/g'_\lambda(x)$. We call the quantity $h/g'_\lambda(x)$ the effective window width at $x$.

As in Wand *et al.* we aim to choose the parameter $\lambda$ so that $f_Y(\cdot;\lambda)$ can be estimated with the smallest possible error. Under the assumption that $f_Y(\cdot;\lambda)$ possesses two continuous derivatives and that $h = h(n) \to 0$ and $nh \to \infty$ as $n \to \infty$, Wand *et al.* showed that the lowest asymptotic mean integrated squared error of $\hat{f}_Y(\cdot;\lambda)$ can be achieved by taking $\lambda$ to minimise

$$J_Y(\lambda) = \left[\int f''_Y(y;\lambda)^2\,dy\right]^{-}. \tag{2.3}$$

This result is the basis for our data-driven selection of $\lambda$ as described in Section 4. Also, because (2.1) is simply a global window width kernel estimator, any one of the current window width selection procedures (see e.g. Park & Marron, 1990) may be applied to select $h$.

## 3. Transformations for Reducing Kurtosis

In this section we assume that the density $f_X$ is symmetric and unimodal. In the theoretical development below, the centre of symmetry is assumed to be 0; in examples this is achieved by subtraction of the sample median from each observation. The classical measure of kurtosis is the fourth moment kurtosis coefficient given by

$$\kappa_X = \mathrm{E}(X^4)/\sigma_X^4.$$

Ruppert (1987) gives an account of several notions of comparative kurtosis. The notion most relevant to this work is that introduced by van Zwet (1964) who defined $f_Y$ as having no more kurtosis than $f_X$ if $Y = g(X)$ where $g$ is convex on the negative half-line and concave on the positive half-line. Such a convex-concave transformation has the effect of taking probability mass from both the peak and the tails and moving it to shoulders (which reduces peakedness and lightens the tails).

It is important that our family of convex-concave transformations is sufficiently smooth so the $f_Y(\cdot\,;\lambda)$ inherits the smoothness properties of $f_X$. Putting $\lambda = (\alpha, \sigma)$, a class of transformations having the required properties is given by

$$\tilde{g}_{\alpha,\sigma}(x) = \alpha x + (1-\alpha)\sigma(2\pi)^{1/2}\{\Phi(x/\sigma) - \tfrac{1}{2}\}$$

for $0 \leq \alpha \leq 1$ and $\sigma > 0$ where $\Phi$ is the distribution function of the standard normal distribution. Decreasing both $\sigma$ and $\alpha$ strengthens the kurtosis reduction of the transformation, but in different manners. When $\alpha = 1$, $\tilde{g}_{\alpha,\sigma}$ has no effect on distributional shape, since it is the identity transformation. For fixed $\alpha$ less than 1, a large value of $\sigma$ makes $\tilde{g}_{\alpha,\sigma}$ close to the identity transformation through the centre region and shoulders, and only the extreme tail region is affected significantly by the transformation. Three parameters, $h$, $\sigma$, and $\alpha$ are necessary to allow independent adjustment of the amount of smoothing at the centre, shoulders, and tails. We call $(h, \alpha, \sigma)$ a "generalised smoothing parameter". The family $\tilde{g}_{\alpha,\sigma}$ has proved sufficiently rich to handle the examples we have examined, both real and simulated, but it has only been tested on densities that are approximately symmetric and unimodal.

Our parametization of $\tilde{g}_\lambda$ ensures $\tilde{g}'_{\alpha,\sigma}(0) = 1$ which implies that the effective window width of (2.2) at the origin is $h$. To understand better the influence of the transformation on (2.2), it is necessary to consider the effective window width at other parts of the density. Corresponding to the example in Section 5.2, Figure 3 shows a peaked, nearly symmetric, density with vertical lines placed at (roughly) the points of differing amounts of curvature of the density. Let $q_0 = 0$ be the centre of symmetry, $q_1$ be a point near the right shoulder of the density and $q_2$ be a point further into the right shoulder or in the right tail. Clearly we would like the effective window width at $q_0$ to be relatively small because of the large amount of curvature there. At $q_2$ we would like a relatively large effective

window width since the curvature there is quite low. An intermediate situation arises at $q_1$. As pointed out in Section 2, the effective window width at a point $q$ is $h/\tilde{g}'_\lambda(q)$ so we set

$$d_i = \tilde{g}'_{\alpha,\sigma}(q_i) = \alpha + (1-\alpha)\exp\{q_i^2/(2\sigma^2)\} \qquad (i = 0, 1, 2). \tag{3.1}$$

Clearly $d_0 = 1$ but the pair $(d_1, d_2)$ is determined by $(\alpha, \sigma)$. Suppose $h$ is chosen to perform the correct amount of smoothing at the peak. Then $h/d_1$ and $h/d_2$ is approximately the amount of smoothing at $q_1$ and $q_2$. Therefore we should select $(d_1, d_2)$ to properly smooth the "shoulder region" and "tail region"and then determine the corresponding value of $(\alpha, \sigma)$ from these.

There is a simple graphical method for choosing $\alpha$ and $\sigma$. First, examine ordinary kernel estimates with various values of $h$ until $h_0$, $h_1$, and $h_2$ are found giving the proper amount of smoothing at the centre $q_0$, and at points $q_1$ amd $q_2$ chosen in the shoulder and in the tail respectively. One should have $h_0 < h_1 < h_2$, for otherwise kurtosis correction will probably not be helpful. Then let $h = h_0$ and $d_i = h/h_i$ for $i = 1, 2$. Knowing $d_1$ and $d_2$, one can then solve (3.1) for the proper $\alpha$ and $\sigma$.

For this graphical method and for the data-driven method of the next section, it is important to know for what pairs $(d_1, d_2)$ (3.1) has a solution. Since $\tilde{g}_{\alpha,\sigma}$ is concave on $[0, \infty)$, the $(d_1, d_2)$ pairs lie in the triangle

$$T = \{(d_1, d_2) : 0 \le d_2 \le d_1 < 1\}.$$

However, for given $(q_1, q_2)$, not all points in $T$ are possible. In fact, straightforward analysis shows that $(d_1, d_2)$ is constrained to lie in the set $D \subset T$ given by

$$D = \{(d_1, d_2) : 0 \le d_1^{(q_2/q_1)^2} < d_2 \le d_1 \le 1\}.$$

## 4. Data-driven Implementation

For complete automatic implementation of our transformation/kernel estimate we require reliable choices of the scale estimates $s_X$ and $s_{\hat{Y}}$, the transformation parameters $(\alpha, \sigma)$ and the window width $h$. For the scale estimates we recommend the standardised interquartile range given by (for the $X$-space data)

$$s_X = \{F^{-1}_{X,n}(\tfrac{3}{4}) - F^{-1}_{X,n}(\tfrac{1}{4})\}/\{\Phi^{-1}(\tfrac{3}{4}) - \Phi^{-1}(\tfrac{1}{4})\}$$

Ideally we would like to choose $(\alpha, \sigma)$ to minimise (2.3). However, since that quantity is unknown we instead choose $(\alpha, \sigma)$ to minimise the "diagonals-in" kernel estimate of $J_Y(\alpha, \sigma)$ (Jones & Sheather, 1991)

$$\hat{J}_Y(\alpha, \sigma) = \left[n^{-2}a^{-5}\sum_{i=1}^{n}\sum_{j=1}^{n} K^{(4)}\{(Y_i - Y_j)/a\}\right]^{1/5},$$

where $K$ is the Gaussian kernel. Result 2 of Jones & Sheather (1991) shows that, in terms of minimising asymptotic mean squared error, the optimal window width for this estimator is $a^* = [3/\{(2\pi)^{1/2} \int (f_Y''')^2 n\}]^{1/7}$. In a sequel (Sheather & Jones, 1991), these authors recommend further kernel estimation of the functional $\int (f_Y''')^2$. This is so that the window width selection rule being proposed there achieves the optimal rate of convergence. In our implementation we used instead a normal model scale rule for estimating this functional which involves replacing $f_Y$ by a normal density with standard deviation estimated by $s_Y$, the standardised interquartile range as defined above. This strategy yields the window width $a = s_Y\{16(2^{1/2})/(5n)\}^{1/7}$ which gives reasonable performance without being computationally expensive. In the examples below, $\hat{J}_Y(\alpha, \sigma)$ was minimized over the grid

$$(\alpha, \sigma) \in \{0, 0.1, \ldots, 0.9, 1\} \times \{0.1, 0.2, \ldots, 1.5\}.$$

Once we have chosen $(\alpha, \sigma)$ we propose using the plug-in window width

$$h_{\mathrm{PI}} = \left[ \frac{\int K^2}{\{\int x^2 K(x)\, dx\}^2 \hat{J}_Y(\alpha, \sigma; a)^5 n} \right]^{1/5} \tag{4.1}$$

to use in the kernel estimators (2.1) and (2.2). This is based on the formula for the asymptotically optimal window width for the estimator (2.1) (Silverman, 1986, p.40) with $J_Y(\alpha, \sigma)^5$ used to estimate $\int f_Y''(\cdot\,; \alpha, \sigma)^2$. As with the choice of $a$ above this is quite a simple and quick rule for choosing the window width $h$ and it could be replaced by other more elaborate rules (Park & Marron, 1990; Sheather & Jones, 1991).

## 5. Examples

### 5.1. Simulation Results

We tested the effectiveness of our method for estimating high-kurtosis densities by applying the transformation/kernel density estimator to 50 samples each of size $n = 200$ from the standard Cauchy distribution and from the normal mixture density $\frac{2}{3}N(0,1) + \frac{1}{3}N(0, 0.01)$. Also, to see how the method performed for cases where the kurtosis is not high we did the same for 50 samples of $N(0,1)$ data again with $n = 200$. The data were generated using the GAUSS programming language and all kernel estimates were obtained using a linear binning algorithm with a Gaussian kernel.

For each simulated data set, we calculated $\mathrm{ISE}_{\mathrm{TR}}$ and $\mathrm{ISE}_{\mathrm{UN}}$, the integrated mean square error with transformation and untransformed, respectively. Using the differences between the two ISEs, we constructed 95% confidence intervals for the difference between the two MISEs, ($\mathrm{MISE}_{\mathrm{TR}} - \mathrm{MISE}_{\mathrm{UN}}$). The confidence intervals are given in Table 1. In the case of the Cauchy or normal mixture,

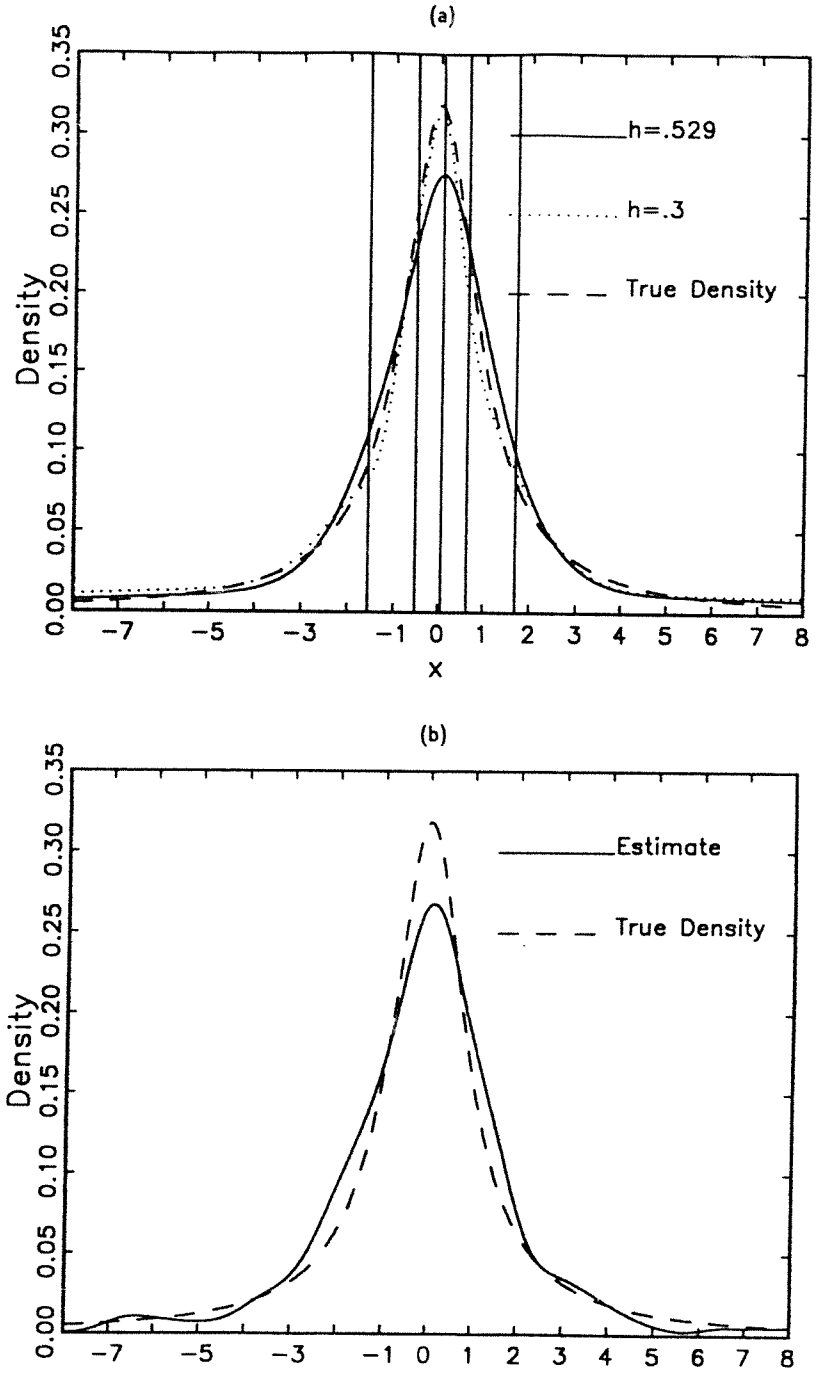


Fig. 1.—Simulated Cauchy data.
(a) Kernel estimates with transformation. $(\alpha, \sigma) = (0.1, 1.5)$. The solid curve is with automatically selected window width $h = 0.529$ giving $\mathrm{ISE}_{\mathrm{TR}} = 2.26 \times 10^{-3}$. The dotted curve is with window width $h = 0.3$ chosen by eye giving $\mathrm{ISE}_{\mathrm{TR}} = 1.71 \times 10^{-3}$.
(b) Ordinary kernel estimate. $h = 0.506$ and $\mathrm{ISE}_{\mathrm{UN}} = 4.25 \times 10^{-3}$.

TABLE 1

*Summary statistics for simulated datasets*

| | 95% conf. int. for $MISE_{UN} - MISE_{TR} \times 10^3$ | 95% conf. int. for E($R$) | # of samples with $R \leq 1$ | 1st quartile of $R$ | 3rd quartile of $R$ |
|---|---|---|---|---|---|
| Normal | (-0.74, 40) | (0.918, 1.03) | 36* | 0.95 | 1.03 |
| Cauchy | (1.57, 2.19) | (0.532, .707) | 48 | 0.49 | 0.64 |
| Normal mixture | (39.0, 49.4) | (0.116, 0.274) | 49 | 0.077 | 0.22 |

* For 6 of the 50 data sets $\alpha = 1$ so the transformation/kernel estimate and the ordinary kernel estimate were identical.

transformation was clearly superior to ordinary kernel estimation, so to further compare the two estimators we computed the ratio $R = \text{ISE}_{\text{TR}}/\text{ISE}_{\text{UN}}$. Table 1 gives the sample mean of $R$, a 95% confidence interval for its expected value, the proportion of cases where $R \leq 1$, and the first and third sample quartiles of $R$.

It is clear from Table 1 that kernel and transformation/kernel estimation are equally effective for normal data, with $.95 \leq R \leq 1.03$ for half the cases. However, for the other two densities, transformation/kernel estimation is clearly superior to ordinary kernel estimation: transformation/kernel estimation had a smaller ISE in 49 out of 50 samples for the normal mixture case, and for 48 out of 50 samples for the Cauchy case.

The transformation/kernel and ordinary kernel estimates are given in Figures 1a and 1b, respectively, for a "typical" Cauchy sample, the sample with the 26th smallest choice of $\text{ISE}_{\text{TR}}$. Both estimates are a bit oversmoothed at the peak, but the ordinary kernel estimate is also undersmoothed in the shoulders and tails, whereas the transformation/kernel estimate appears to have the proper amount of smoothing in the shoulders and tails. The transformation/kernel estimator in Figure 1a had a data-driven window width of 0.53 and an ISE of $2.26 \times 10^{-3}$, while the same quantities for the ordinary kernel estimator in Figure 1b were 0.51 and $4.25 \times 10^{-3}$. In Figure 1a, vertical lines are drawn at $2q_0 - q_2 = -1.56$, $2q_0 - q_1 = -0.51$, $q_0 = 0.07$ (sample median), $q_1 = 0.65$, and $q_2 = 1.70$. For these values of $q_1$ and $q_2$, the data-driven value of $(\alpha, \sigma)$ corresponds to $q_1 = 0.62$ and $q_2 = 1.67$. From the values of $q_1$ and $q_2$ — that although $q_2$ is only at the transition from the shoulder to the tail, the amount of smoothing at $q_2$ is already substantially greater than at the centre.

Since window width selection rule (4.1) performed a little disappointingly for Cauchy data we also include a plot of a density estimate based on the sample and transformation parameters, but with the window width chosen by eye. This is the dotted line of Figure 1a which is seen to be quite a satisfactory estimate of the Cauchy density with a sample size of $n = 200$. While this estimate was

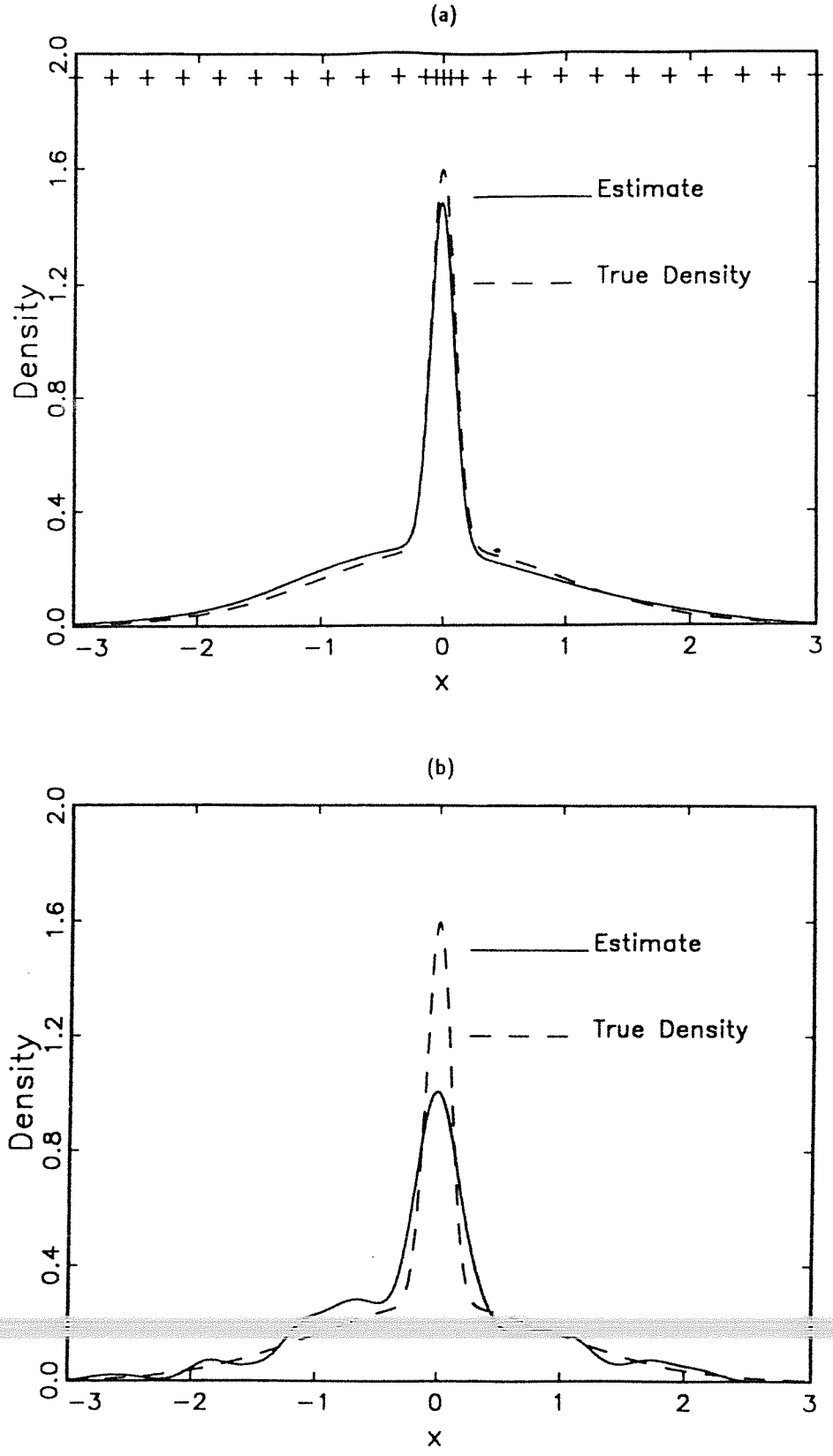


Fig. 2.—Simulated normal mixture data, $\frac{2}{3}N(0,1) + \frac{1}{3}N(0,.1^2)$.
(a) Kernel estimate with transformation. $(\alpha, \sigma) = (0.2, 0.1)$, $h = 0.097$ and $\mathrm{ISE}_{\mathrm{TR}} = 6.00 \times 10^{-3}$. (b) Ordinary kernel estimate. $h = 0.142$ and $\mathrm{ISE}_{\mathrm{UN}} = 57.24 \times 10^{-3}$.

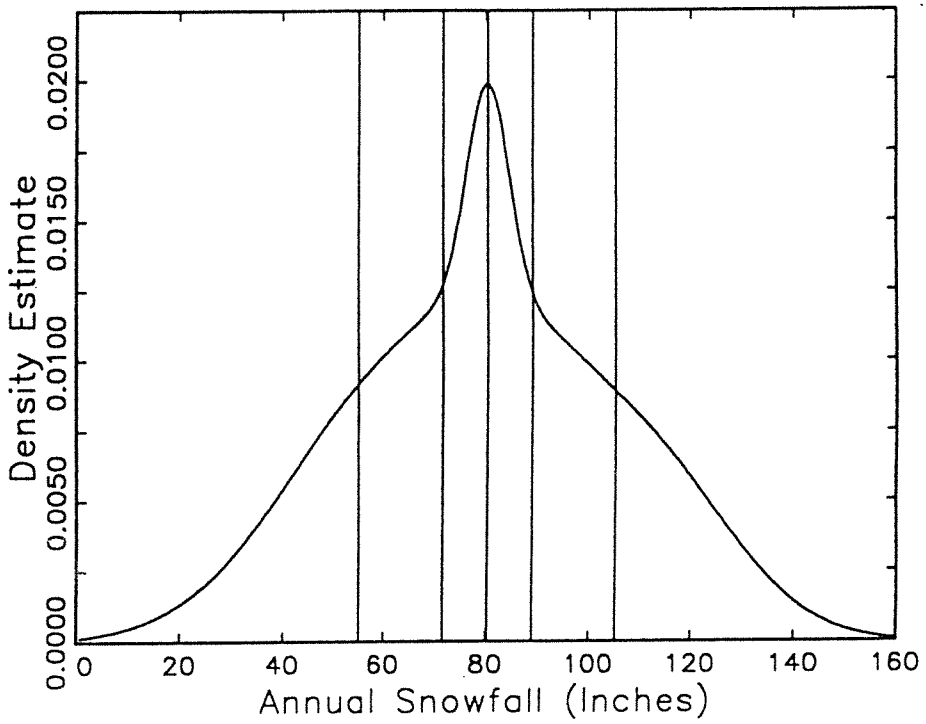


Fig. 3.—Annual snowfall in Buffalo, New York.
Kernel density estimate with transformation. $(\alpha, \sigma) = (0.64, 4.05)$, $h = 9.03$.

not automatically generated we feel that it demonstrates the flexibility of the proposed transformation/kernel estimator.

Figures 2a and 2b are analogous to Figures 1a and 1b for the normal mixture density. Notice that the transformation/kernel estimator appears to have the right amount of smoothing throughout the range of $X$, whereas the ordinary kernel estimate is noticeably oversmoothed at the peak and seriously undersmoothed in the tails. Using $(q_1, q_2, q_3) = (0.00, 0.14, 0.68)$, we have $(d_1, d_2) = (0.5, 0.2)$. Thus the effective window widths at 0.14 and 0.68 are, respectively, 2 and 5 times larger than at 0. In Figure 2a, one sees that 0.14 is not at all far from the centre and 0.68 is hardly into the tail region. This extreme change in effective window width is needed to estimate the sharp spike at the centre properly. The estimate in Figure 2a is remarkably close to the true density — one should recall that the estimate is completely data-driven and that this is the "median case" out of 50 simulations. The window width and ISE are 0.097 and $6.00 \times 10^{-3}$ for the transformation/kernel estimator in Figure 2a and 0.142 and $57.24 \times 10^{-3}$ for the ordinary kernel estimator in Figure 2b.

## 5.2 The Buffalo Snowfall Data

The Buffalo snowfall data consist of amounts of annual snowfall in inches at Buffalo, New York, for 63 consecutive years between 1910/11 and 1972/73 (the data are listed in Scott, 1985). Using the data-driven selection algorithms described in the previous section we obtained $(\alpha, \sigma) = (0.64, 4.05)$ and $h = 9.03$. Figure 3 shows the corresponding transformation/kernel density estimate, with

vertical lines drawn at $2q_0 - q_2 = 54.5$, $2q_0 - q_1 = 71.0$, $q_0 = 79.6$, $q_1 = 88.2$, and $q_2 = 104.7$. With the data-driven choice of $(\alpha, \sigma)$, we have $d_1 = 0.68$ and $d_2 = 0.64$, so the effective window widths are 9.03, 13.3, and 14.1 at $q_0$, $q_1$, and $q_2$, respectively. It is interesting to compare this with the global window width kernel estimates obtained by Silverman (1986, p.45) in his Figures 3.2 (a) and (b) where he uses window widths of 6.0 and 12.0 respectively. In the former, the estimate is a "trimodal curve suggesting a mixture of three populations in the ratio 1 : 3 : 1", while in the latter Silverman obtains a much less pronounced peak with the smaller modes replaced by some slight shoulder structure. The transformation/kernel estimate gives a different interpretation, with the mixture of two equally located populations being suggested.

## *References*


ABRAMSON, I.S. (1982). On bandwidth variation in kernel estimates — a square root law. *Ann. Statist.* **9**, 168–176.

BREIMAN, L., MEISEL, W. & PURCELL, E. (1977). Variable kernel estimates of probability density estimates. *Technometrics* **19**, 135–144.

DEVROYE, L. & GYÖRFI, L. (1985). *Nonparametric Density Estimation: The $L_1$ View.* New York: Wiley.

JONES, M.C. & SHEATHER, S.J. (1991). Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. *Statist. Probab. Letters* **11**, 511–514.

PARK, B.U. & MARRON, J.S. (1990). Comparison of data-driven bandwidth selectors. *J. Amer. Statist. Assoc.* **85**, 66–72

RUPPERT, D. (1987). What is kurtosis?: an influence function approach. *Amer. Statistician*, **41**, 1–5.

SCOTT, D.W. (1985). Average shifted histograms: effective nonparametric density estimators in several dimensions. *Ann. Statist.* **13**, 1024–1040.

SHEATHER, S.J. & 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.* New York: Chapman and Hall.

VAN ZWET, W.R. (1964). *Convex Transformations of Random Variables.* Amsterdam: Mathematisch Centrum.

WAND, M.P., MARRON, J.S. & RUPPERT, D. (1991). Transformations in density estimation. *J. Amer. Statist. Assoc.* **86**, 343–366.