1909 R.

DECONVOLUTING KERNEL DENSITY ESTIMATORS

Leonard A. Stefanski
Department of Statistics
North Carolina State University
Raleigh, NC 27695

Raymond J. Carroll
Department of Statistics
Texas A&M University
College Station, TX 77843

ABSTRACT

This paper considers estimation of a continuous bounded probability density when observations from the density are contaminated by additive measurement errors having a known distribution. Properties of the estimator obtained by deconvolving a kernel estimator of the observed data are investigated. When the kernel used is sufficiently smooth the deconvolved estimator is shown to be pointwise consistent and bounds on its integrated mean squared error are derived. Very weak assumptions are made on the measurement-error density thereby permitting a comparison of the effects of different types of measurement error on the deconvolved estimator.

Key Words and Phrases: Convolution, Deconvolution, Density estimation, Errors-in-variables, Kernel, Measurement Error Models,

AMS 1980 *subject classifications.* Primary 62J05; secondary 62H25, 62G05.

Abbreviated title: Density Deconvolution.

1. INTRODUCTION

1.1 The deconvolution problem

Let U and Z be independent random variables with probability density functions g and h respectively. Then the random variable $X = U + Z$ has the density $f = g * h$ where $*$ denotes convolution. Assuming h is known we consider estimating g from a set of independent observations $\{X_j\}_{j=1}^n$ having the common density f.

The problem arises whenever data are measured with nonnegligible error and knowledge of g is desired. In this case U represents a true value, Z is an error of measurement and X is an observed value. An application is discussed in Mendelsohn and Rice (1983); other applications and related work can be found in Eddington (1913), Trumpler and Weaver (1953), Kahn (1955), Gaffey (1959), Wise et. al. (1977), and Devroye and Wise (1979). Our interest in the problem arises from its potential for application in measurement-error modelling. For example, work is in progress at the Radiation Effects Research Foundation, Hiroshima, Japan, to assess the health effects of radiation exposure. Measured exposures are known to contain substantial measurement errors. Some of the statistical models proposed for the data require estimation of the distribution (density) of true radiation exposures given only data on measured exposures and reasonable assumptions concerning the error distribution. Since the sample sizes involved are very large, nonparametric density estimation with deconvolution seems feasible. This paper establishes the elementary asymptotic theory associated with deconvolving a kernel density estimator and presents results from a simulation study demonstrating the difficulty of nonparametric deconvolution in moderate to large samples.

The deconvolution problem can also be cast in the format of an empirical Bayes problem wherein g represents the prior distribution for a sequence of location parameters. Estimation of prior distributions or mixing distributions have been studied by several authors, for example, Blum and Susarla (1977), Choi and Bulgren (1968), Deely and Kruse (1968), Maritz (1967) and Preston (1971). The estimator proposed in Section 1.2 is a transformation of a kernel density estimator of f. Some papers in which a transformation of a density estimator is of primary interest include Taylor (1982) and Hall and Smith (1986). Recent contributions to the literature on deconvolution include Carroll and Hall (1988), Fan (1988), Liu and Taylor (1988a,b) and Stefanski (1989).

Although many of the estimators proposed in the literature have been shown to be consistent less is known about their rates of convergence. The estimator we propose has the advantage of being analytically and, in some cases, computationally no more complex than an ordinary kernel density estimator, thus facilitating a discussion of its convergence properties. However, a price is

paid for the reduction in complexity in that the resulting estimator is not range preserving, i.e., it assumes negative values with positive probability.

Throughout we make very weak assumptions on h and this allows us to assess the effects of different types of measurement error. A conclusion indicated by the analysis is that the difficulty of the deconvolution problem varies inversely with the smoothness of the measurement-error density. Thus deconvolution is particularly difficult under the common assumption of normally distributed errors.

Some conditions on h are necessary to insure that g is identifiable. We assume that h has a nonvanishing characteristic function, ϕ_h, i.e.,

$$|\phi_h(t)| > 0, \quad \text{for all real } t. \tag{1.1}$$

Although (1.1) is not the weakest assumption insuring identifiability of g it holds in many cases of interest, and in particular at the normal model.

1.2 The estimator

Let K be a bounded even probability density function whose characteristic function, ϕ_K, satisfies, for each fixed $\lambda > 0$,

$$\sup_t |\phi_K(t)/\phi_h(t/\lambda)| < \infty; \qquad \int |\phi_K(t)/\phi_h(t/\lambda)|dt < \infty. \tag{1.2}$$

Implied by (1.2) are the facts that $\phi_K^2/|\phi_h(\bullet/\lambda)|^2, |\phi_K|$ and ϕ_K^2 are all integrable, which in turn implies that ϕ_K is invertible, i.e.,

$$K(x) = (2\pi)^{-1} \int e^{-itx}\phi_K(t)dt. \tag{1.3}$$

Let $\hat{f}$ be an ordinary kernel density estimator of f based on the kernel K,

$$\hat{f}(x) = (n\lambda)^{-1} \sum_{j=1}^{n} K\big((X_j - x)/\lambda\big). \tag{1.4}$$

The characteristic function of $\hat{f}$ is denoted $\phi_{\hat{f}}$ and satisfies $\phi_{\hat{f}}(t) = \hat{\phi}(t)\phi_K(\lambda t)$ where $\hat{\phi}(t) = n^{-1}\sum_{j=1}^{n} e^{itX_j}$ is the empirical characteristic function of $\{X_j\}_{j=1}^n$. Under (1.2) $\phi_{\hat{f}}/\phi_h$ is an integrable function and therefore possesses a Fourier transform. The estimator we propose is $\hat{g}$ given by

$$\hat{g}(x) = (2\pi)^{-1} \int e^{-itx}\big\{\phi_{\hat{f}}(t)/\phi_h(t)\big\}dt. \tag{1.5}$$

Note that it is not possible to replace ϕ_f with $\hat{\phi}$ in (1.5) since the resulting integral does not exist. By using ϕ_f in place of $\hat{\phi}$ we are able to force integrability of the integrand in (1.5) by suitable choice of ϕ_K.

Define the function K_λ^* as

$$K_\lambda^*(t) = (2\pi)^{-1} \int e^{ity} \{\phi_K(y)/\phi_h(y/\lambda)\} dy. \tag{1.6}$$

Then $\hat{g}$ has the representation

$$\hat{g}(x) = (n\lambda)^{-1} \sum_{j=1}^{n} K_\lambda^*((X_j - x)/\lambda). \tag{1.7}$$

Properties of $\hat{g}$ are best understood in terms of the properties of K_λ^* and the latent variables $\{U_j, Z_j\}_{j=1}^n$.

Equation (1.2) implies that $|K_\lambda^*|$ is bounded, thus $|\hat{g}|$ is also bounded and its expectation necessarily exists. Furthermore, an interchange of expectation and integration, justified by Fubini's Theorem and (1.2), shows that

$$E\Big\{K_\lambda^*((X - x)/\lambda)|U\Big\} = K((U - x)/\lambda). \tag{1.8}$$

From (1.8) it follows that

$$E\{\hat{g}(x)\} = \lambda^{-1} E\Big\{K((U - x)/\lambda)\Big\} = \int \lambda^{-1} K((u - x)/\lambda) g(u) du. \tag{1.9}$$

Thus $\hat{g}$ has the same bias as an ordinary kernel density estimator. Formally, this is a consequence of the fact that the linear operations of expectation and Fourier transformation commute. Furthermore if $\hat{g}^*$ is the kernel estimator

$$\hat{g}^*(x) = (n\lambda)^{-1} \sum_{j=1}^{n} K((U_j - x)/\lambda)$$

then it follows from (1.8) that $E\{\hat{g}(x)|U_1, ..., U_n\} = \hat{g}^*(x)$. Thus, conditionally $\hat{g}$ can be viewed as an unbiased estimator of $\hat{g}^*$.

The fact that ϕ_K is even and real implies that K_λ^* is real and thus $\hat{g}$ is also. When h is even, K_λ^* is even. If $\phi_K/\phi_h(\bullet/\lambda)$ possesses m continuous integrable derivatives, then it follows from the Riemann-Lebesgue Lemma that $K_\lambda^*(t) = o(|t|^{-m})$ as $|t| \to \infty$ and for $m \geq 2$ this means that K_λ^* and hence $\hat{g}$ are integrable. Furthermore in this case the Fourier Inversion Formula indicates that

$$\phi_K(\lambda y)/\phi_h(-y) = \lambda^{-1} \int e^{ity} K_\lambda^*(t/\lambda) dt. \tag{1.10}$$

Evaluating (1.10) at $y = 0$ shows that $\int K_\lambda^*(t)dt = 1$ implying that $\int \hat{g}(x)dx = 1$. Since K_λ^* has many properties of an ordinary kernel we call it a *deconvoluting* kernel. The one property it lacks is nonnegativity; the left hand side of (1.10) is not positive definite thus K_λ^* cannot be a true probability density. Since, conditioned on $\{U_j\}_{j=1}^n, \hat{g}$ is an unbiased estimator of $\hat{g}^*$, this problem can be viewed as a failure of unbiased estimators to be range preserving.

In summary, provided $\phi_K/\phi_h(\bullet/\lambda)$ is smooth and (1.2) holds, $\hat{g}$ is continuous, real-valued and integrable with $\int \hat{g}(x)dx = 1$. Although the severity of (1.2) depends on ϕ_h, it is always possible to satisfy these conditions when (1.1) holds by choosing ϕ_K so that it vanishes outside a finite interval. For example, we can take ϕ_K proportional to $U^{(2m)}$ where $U^{(2m)}$ is the $2m$-fold convolution of the uniform density, $\chi(|x| \leq 1)/2$, with itself. The corresponding density is proportional to $\{\sin(t)/t\}^{2m}$. When $m \geq 2, U^{(2m)}$ has two continuous integrable derivatives and the smoothness conditions on $\phi_K/\phi_h(\bullet/\lambda)$ are obtained provided ϕ_h is sufficiently smooth. If ϕ_h is not smooth then $\hat{g}$ need not be integrable although it will still be bounded and square integrable.

For certain measurement-error distributions (1.6) has a closed-form expression. For example, when $h(x) = (1/2)e^{-|x|}, K_\lambda^*(t) = K(t) - \lambda^{-2}K''(t)$. In fact the integral in (1.6) can be evaluated analytically whenever $1/\phi_h$ is a polynomial. Unfortunately, it does not seem possible to obtain K_λ^* in closed form for the normal measurement-error model.

2. ASYMPTOTIC RESULTS

In this section we establish the point-wise consistency of $\hat{g}$ and derive an approximation to its integrated mean square error. Throughout we work under the assumptions that g is continuous and bounded and hence square integrable.

THEOREM 2.1. If ϕ_K and ϕ_h are such that (1.1) and (1.2) hold and g is continuous and bounded then $\hat{g}(x)$ defined by (1.5) is a consistent estimator of $g(x)$ provided $n \to \infty$, $\lambda \to 0$ and $(n\lambda)^{-1} \int \phi_K^2(t)|\phi_h(t/\lambda)|^{-2}dt \to 0$.

PROOF. Since $|\phi_K/\phi_h(\bullet/\lambda)|$ is square integrable we have from (1.6) and Parseval's Identity

$$\int \{K_\lambda^*(x)\}^2 dx = (2\pi)^{-1} \int \phi_K^2(t)|\phi_h(t/\lambda)|^{-2}dt. \tag{2.1}$$

In light of (1.9) we can appeal to known results on kernel density estimators to claim that the bias of $\hat{g}(x)$ converges to zero as $\lambda \to 0$.

Define

$$A(\lambda, a) = \int \{K_\lambda^*(x)\}^2 g(a + \lambda x)dx \left[\int \{K_\lambda^*(x)\}^2 dx\right]^{-1}$$

and note that $A(\lambda, a)$ is bounded by $B_g = \sup_x g(x)$. Now note that

$$E\Big\{K_\lambda^*((X-x)/\lambda)\Big\}^2 = \int\int \Big\{K_\lambda^*((z+u-x)/\lambda)\Big\}^2 g(u)du h(z)dz$$

and after the change of variables $t = (z+u-x)/\lambda$ in the inner integral we get

$$E\Big\{K_\lambda^*((X-x)/\lambda)\Big\}^2 = \lambda \int A(\lambda, x-z)h(z)dz \int \big\{K_\lambda^*(t)\big\}^2 dt \le \lambda B_g \int \big\{K_\lambda^*(t)\big\}^2 dt. \qquad (2.2)$$

Now since $n\lambda^2 Var\{\hat{g}(x)\}$ is bounded by the left hand side of (2.2) we find that

$$Var\{\hat{g}(x)\} \le (n\lambda)^{-1} B_g \int \big\{K_\lambda^*(t)\big\}^2 dt = (2\pi n\lambda)^{-1} B_g \int \phi_K^2(t)|\phi_h(t/\lambda)|^{-2} dt, \qquad (2.3)$$

upon appealing to (2.1). Under the assumptions of the theorem, (1.9) and (2.3) show that $E\{\hat{g}(x)\} \to g(x)$ and $Var\{\hat{g}(x)\} \to 0$ thus concluding the proof. ••

Now we derive an approximation to the integrated mean squared error of $\hat{g}$. Using Parseval's Identity, (2.1) and the change of variables $t = (X-y)/\lambda$ we have that as $n \to \infty$ and $\lambda \to 0$,

$$\begin{aligned}
\int Var\{\hat{g}(y)\}dy &= n^{-1} \int \lambda^{-2} E\Big\{K_\lambda^*((X-y)/\lambda)\Big\}^2 dy - n^{-1} \int \Big[E\Big\{\lambda^{-1} K_\lambda^*((X-y)/\lambda)\Big\}\Big]^2 dy \\
&= n^{-1} E \int \lambda^{-2}\Big\{K_\lambda^*((X-y)/\lambda)\Big\}^2 dy - (2n\pi)^{-1} \int |\phi_g(t)|^2 \phi_K^2(\lambda t)dt \\
&= (\lambda n)^{-1} \int \big\{K_\lambda^*(t)\big\}^2 dt - (2n\pi)^{-1} \int |\phi_g(t)|^2 \phi_K^2(\lambda t)dt \\
&= (2\pi n\lambda)^{-1} \int \phi_K^2(t)|\phi_h(t/\lambda)|^{-2} dt + o\{(n\lambda)^{-1}\} \\
&\sim (2\pi n\lambda)^{-1} \int \phi_K^2(t)|\phi_h(t/\lambda)|^{-2} dt. \qquad (2.4)
\end{aligned}$$

If in addition to previous assumptions, g possesses two bounded integrable derivatives then as $\lambda \to 0$,

$$\int \Big[E\{\hat{g}(x)\} - g(x)\Big]^2 dx \sim (\lambda^4/4)\mu_{K,2}^2 \int \big\{g''(x)\big\}^2 dx \qquad (2.5)$$

when $\mu_{K,2} = \int y^2 K(y)dy < \infty$. Combining (2.4) and (2.5) we have that to a first-order approximation

$$\int E\{\hat{g}(x) - g(x)\}^2 dx \sim (2\pi n\lambda)^{-1} \int \phi_K^2(t)|\phi_h(t/\lambda)|^{-2} dt + (\lambda^4/4)\mu_{K,2}^2 \int \big\{g''(x)\big\}^2 dx. \qquad (2.6)$$

The first term in (2.6) can be much larger than the variance component of the integrated mean squared error of an ordinary kernel density estimator. This is the price paid for not measuring $\{U_j\}_{j=1}^n$ precisely. The rate at which

$$V_{K,h}(\lambda) = \int \phi_K^2(t)|\phi_h(t/\lambda)|^{-2} dt \qquad (2.7)$$

diverges as λ decreases is dictated by the tail behavior of $|\phi_h|$, which in turn is related to the smoothness of h. Suppose that ϕ_K is strictly positive on $(-B, B)$ and vanishes off this interval. Then considering (2.7) when h is standard normal, Cauchy and double-exponential we have respectively that for each $0 < \epsilon < B$ there exist positive constants $c_1, ..., c_5$ such that

$$c_1 e^{(B-\epsilon)^2/\lambda^2} \leq V_{K,h}(\lambda) \leq c_2 e^{B^2/\lambda^2} \quad \text{(normal)}$$

$$c_3 e^{(B-\epsilon)/\lambda} \leq V_{K,h}(\lambda) \leq c_4 e^{2B/\lambda} \quad \text{(Cauchy)}$$

$$V_{K,h}(\lambda) \sim c_5 \lambda^{-4} \quad \text{(double exponential).}$$

Thus in these cases in order for (2.4) to converge to zero, λ must approach zero at a rate no faster than $\{\log(n)\}^{-1/2}$ for the normal model, no faster than $\{\log(n)\}^{-1}$ for the Cauchy model, and no faster than $n^{-1/5}$ for the double-exponential model. In other words, these are necessary conditions on λ if the first term on the right hand side of (2.6) is to be asymptotically negligible. By considering these rates in the right hand side of (2.5), it follows that the best possible rates on the integrated mean squared errors of $\hat{g}$ are $\{\log(n)\}^{-2}$, $\{\log(n)\}^{-1}$ and $n^{-4/9}$ for the normal, Cauchy and double-exponential cases respectively. Here we have considered only the order of the bandwidth. A more complete discussion of bandwidth selection can be found in Stefanski(1989).

The normal, Cauchy and double-exponential densities share the same ordering with respect to their peakedness at the origin; the normal is least peaked and the double-exponential is most peaked. The relationship between the variance of $\hat{g}$ and the peakedness of h is intuitive if we think of peakedness as a measure of the closeness of h to a delta function for which measurement error is identically zero and the deconvolution problem disappears. This analogy can be pushed a little further by considering a model wherein only $100p\%(0 < p < 1)$ of the data are measured with error and the remaining data are error free. In this case we have $X = U + Z^*$ where $P(Z^* = 0) = 1 - p$ and $P(Z^* = Z) = p$. The characteristic function, ϕ^* of Z^* is $\phi^*(t) = (1 - p) + p\phi_h(t)$. Nothing in the previous analysis required Z to have an absolutely-continuous distribution. If Z^* and not Z is the measurement-error variable then the variance term in (2.7) becomes

$$(2\pi n\lambda)^{-1} \int \{\phi_K^2(t)|1 - p + p\phi_h(t/\lambda)|^{-2}\}dt \sim (2\pi n\lambda)^{-1} \int \phi_K^2(t)dt/(1-p)^2$$

which is the same order of magnitude as the variance term for ordinary kernel density estimation. Thus for the model in which some data are error free we get convergence of the integrated mean-squared error at the usual rates. A similar model for discrete data was studied by Devroye and

Wise (1979). We do not know of any instances in which this model has been studied for continuous variates.

3. NORMAL MEASUREMENT ERROR

We now argue that the poor performance of $\hat{g}$ at the normal measurement error model is intrinsic to the deconvolution problem and that convergence rates like $\{\log(n)\}^{-p}$, $p > 0$, are to be expected.

Let $\bar{g}$ be any estimator of g which is continuous, bounded and integrable. Then $\bar{g}$ determines an estimator of f, namely $\bar{f} = h * \bar{g}$ where h is the standard normal density. It follows that $\bar{f}$ and all of its derivatives are continuous, bounded and integrable.

Let $\mathcal{C}$ be the convolution operator corresponding to standard normal measurement error and let $\mathcal{D}$ be the differential operator. Expanding $e^{t^2/2}$ in a Maclaurin series we get that formally $\mathcal{C}^{-1} = \sum_{j=1}^{\infty}(-1/2)^j\mathcal{D}^{2j}/j!$. Thus when $\bar{f} = h * \bar{g}$, $\bar{g}(x) = \sum_{j=1}^{\infty}(-1/2)^j\bar{f}^{(2j)}(x)/j!$ provided the series is convergent. Therefore we cannot expect to estimate g any better than we can estimate arbitrarily high derivatives of f. In contrast, when h is double-exponential, $\phi_h(t) = 1/(1+t^2)$, and deconvolution corresponds to a differential operator of order 2. Thus the rate of convergence for deconvolving double-exponential errors is the same as for estimating a second derivative.

Theorem 3.1 below shows that under certain regularity conditions, if the performance of $\bar{f}$ deteriorates sufficiently upon *one* differentiation, then the rate of convergence of the integrated mean squared error of $\bar{g}$ can be no better than inverse powers of $\log(n)$ for a large class of estimators.

In the theorem it is assumed that $\bar{g}, \bar{f}$ and $\bar{f}'$ are bounded, integrable estimators of g, f and f' respectively. For an integrable random function, $\bar{s}(\bullet)$, with an integrable expectation, $E\{s(\bullet)\}$, let $\phi_{\bar{s}}(t)$ and $\phi_{E\{\bar{s}\}}(t)$ denote $\int e^{itx}\bar{s}(x)dx$ and $\int e^{itx}E\{\bar{s}(x)\}dx$ respectively. With this notation we have that $\phi_{\bar{f}} = \phi_h\phi_{\bar{g}}$ and $\phi_{\bar{f}'}(t) = -it\phi_{\bar{f}}(t)$ under the conditions stated above.

THEOREM 3.1. If for $\bar{s} = \bar{g}, \bar{f}$ and $\bar{f}'$,

$$E\{\phi_{\bar{s}}(t)\} = \phi_{E\{\bar{s}\}}(t); \tag{3.1}$$

$$\int Var\{\bar{s}(x)\}dx = (2\pi)^{-1}\int E\{|\phi_{\bar{s}}(t) - \phi_{E\{\bar{s}\}}(t)|^2\}dt; \tag{3.2}$$

$$\int Var\{\bar{f}(x)\}dx = n^{-1}c_{1,n}a_n; \tag{3.3}$$

$$\int \left[E\{\bar{f}(x)\} - f(x)\right]^2 dx = c_{2,n}a_n^{-r}; \tag{3.4}$$

$$\int Var\{\bar{f}'(x)\}dx = n^{-1}c_{3,n}a_n^{1+\epsilon}; \tag{3.5}$$

where r and ϵ are positive constants; the sequence $\{a_n\} \to \infty$; and $\{c_{1,n}\}, \{c_{2,n}\}$ and $\{c_{3,n}\}$ are convergent sequences with positive limits; then the integrated mean squared error of $\bar{g}$ exceeds $c_{2,n}\{(c_{1,n}/c_{3,n})\log(n)\}^{-r/\epsilon}$ for large n.

PROOF. For convenience drop the subscript n on $a_n, c_{1,n}, ..., c_{3,n}$. The relationship $\phi_{\bar{f}} = \phi_{\bar{g}}\phi_h$, (3.1), (3.2), Jensen's inequality, (3.3) and (3.5) are used to show that

$$\int Var\{\bar{g}(x)\}dx = (2\pi)^{-1}\int e^{t^2}E\{|\phi_{\bar{f}}(t) - \phi_{E\{\bar{f}\}}(t)|^2\}dt$$

$$\geq \int Var\{\bar{f}(x)\}dx \exp\left(\left[\int Var\{\bar{f}'(x)\}dx\right]\left[\int Var\{\bar{f}(x)\}dx\right]^{-1}\right)$$

$$= (n^{-1}ac_1)\exp\{(c_3/c_1)a^\epsilon\}.$$

If $(c_3/c_1)a^\epsilon \geq \log(n)$, then $(c_1a/n)\exp\{(c_3/c_1)a^\epsilon\}$ diverges and thus the integrated mean squared error of $\bar{g}$ diverges unless $a < \{(c_1/c_3)\log(n)\}^{1/\epsilon}$. But (3.4), (3.1) and two simple inequalities together imply that

$$c_2a^{-r} = \int\left[E\{\bar{f}(x)\} - f(x)\right]^2dx = (2\pi)^{-1}\int e^{-t^2}|\phi_{E\{\bar{g}\}}(t) - \phi_g(t)|^2dt$$

$$\leq (2\pi)^{-1}\int |\phi_{E\{\bar{g}\}}(t) - \phi_g(t)|^2dt$$

$$\leq E\left[\int\{\bar{g}(x) - g(x)\}^2dx\right].$$

Thus when $a < \{(c_1/c_3)\log(n)\}^{1/\epsilon}$ the integrated mean squared error of $\bar{g}$ exceeds

$$c_2\{(c_1/c_3)\log(n)\}^{-r/\epsilon},$$

which concludes the proof. ••

When $\bar{f}$ is a kernel density estimator with nonnegative kernel, a^{-1} is the bandwidth, $r = 4, \epsilon = 2$ and the theorem shows that the mean integrated squared error of $\bar{g}$ can converge at a rate no faster than $\{\log(n)\}^{-2}$.

The significance of Theorem 3.1 lies in the fact that most nonparametric density estimators used in applications show a significant drop in performance upon differentiation. The theorem indicates that if such an estimator is deconvolved to remove a normal component, then the resulting

estimator will have a mean integrated squared error that conveges no faster than negative powers of $\log(n)$.

In a paper that has appeared since submission of this paper, (Carroll and Hall, 1988), it is established that the best *point-wise* convergence rates for deconvolving normal measurement error are proportional to $\{\log(n)\}^{-d/2}$ when g has d bounded derivatives. This result and Theorem 3.1 suggest that deconvolving normal measurement error is generally not likely to be very successful except in very large samples or when the density being estimated is extremely smooth and its smoothness is exploited. This often means resorting to estimators of f and g which are neither positive or integrable or both. Smoothness assumptions on g are not unreasonable in some applications and if we are interested primarily in determining the gross structure of g, e.g., presence and location of modes (Silverman, 1981), nonnegativity and integrability are not crucial.

4. A PROCEDURE FOR NORMAL ERRORS

Due to the difficulties inherent in deconvolving normal errors we investigated use of the so-called sinc kernel, $K(x) = (\pi x)^{-1}\sin(x)$, (Davis, 1975; Tapia and Thompson, 1978) with characteristic function $\phi_K(t) = \chi(|t| \leq 1)$. This kernel takes full advantage of smoothness properties of g by allowing bias to decrease at rates dictated by the tail behavior of $|\phi_g|$. Lighter tails of $|\phi_g|$ correspond to better convergence rates for $\hat{g}$. The improved asymptotic performance is obtained at the expense of nonnegativity of $\hat{f}$ and integrability of both $\hat{f}$ and $\hat{g}$. These are properties which might reasonably be sacrificed for the sake of determining shape.

The estimator proposed in Section 1 requires a bandwidth-selection rule for implementation. We now show that a cross-validation approximation to the integrated squared error of $\hat{g}$, $AISE_g(\lambda)$, is given by

$$AISE_g(\lambda) = \int_0^{1/\lambda} \frac{2 - (n+1)|\hat{\phi}(t)|^2}{(n-1)\pi|\phi_h(t)|^2} dt. \tag{4.1}$$

Let $\{X\}_{(j)}$ denote the sequence $\{X_j\}$ with X_j removed and let $\hat{\phi}_{(j)}$ be the empirical characteristic function of $\{X\}_{(j)}$. Then the fact that

$$E\Big[\, e^{-itX_j}\hat{\phi}_{(j)}(t)|U_j, \{X\}_j\Big] = \hat{\phi}_{(j)}(t)e^{-itU_j}\phi_h(-t),$$

motivates the approximation via Parseval's relation

$$\int \hat{g}g dx \sim (2\pi n)^{-1} \int \frac{\sum_{j=1}^n e^{-itX_j}\hat{\phi}_{(j)}(t)\phi_K(\lambda t)}{|\phi_h(t)|^2} dt.$$

Thus for the purpose of minimization

$$\int (\hat{g}-g)^2 dx \sim \int \frac{n|\phi_K(\lambda t)|^2|\hat{\phi}(t)|^2 - 2\sum_{j=1}^n e^{-itX_j}\hat{\phi}_{(j)}\phi_K(\lambda t)}{(2\pi n)|\phi_h(t)|^2} dt = AISE_g(\lambda).$$

The last equality follows from (4.1) upon invoking the relationship

$$\sum_{j=1}^n e^{-itX_j}\hat{\phi}_{(j)}(t) = (n-1)^{-1}\{n^2|\hat{\phi}(t)|^2 - n\},$$

substituting $\chi(|\lambda t| \le 1)$ for $\phi_K(\lambda t)$ and noting that $|\hat{\phi}(\bullet)|^2$ is even.

The cross-validation approximation to the integrated squared error of $\hat{f}$, $AISE_f(\lambda)$, is given by the right side of (4.1) upon setting $\phi_h(t) \equiv 1$. Differentiating (4.1) with respect to λ shows that extreme points of $AISE_g$ and $AISE_f$ both satisfy $\hat{I}_n(\lambda) = 0$, where

$$\hat{I}_n(\lambda) = n(n-1)^{-1}\{2 - (n+1)|\hat{\phi}(1/\lambda)|^2\}. \tag{4.2}$$

Equation (4.2) indicates that the optimal cross-validation bandwidths for estimating g and f respectively are identical except possibly when (4.2) has multiple solutions. This is a consequence of using the sinc kernel and the fact that f is infinitely differentiable. A sufficient condition for the *mean* integrated squared errors of $\hat{g}$ and $\hat{f}$ to have identical minima is given in the Appendix.

5. SIMULATION RESULTS

We conducted a Monte Carlo study to determine if $\hat{g}$ is capable of revealing features of g which are masked by convolution with h. In particular we took g to be a $50-50$ mixture of normal densities with means $\pm(2/3)^{1/2}$ and common variances $1/3$. Normal measurement error with variance $1/3$ was added to g so that f is a $50-50$ normal mixture with means, $\pm(2/3)^{1/2}$, and common variances, $2/3$. For this parameterization, g is bimodal, f is unimodal and the measurement error variance is $1/3$ the variance of g, although it equals the variance of each normal component of g.

Observations were generated from f, and $\hat{g}$ was computed according to (1.5) with $\phi_K(t) = \chi(|t| \le 1)$. This required numerical integration of

$$\hat{g}(x) = \pi^{-1}\int_0^{1/\lambda} q_{n,\sigma}(t)dt$$

where $q_{n,\sigma}(t) = n^{-1}\sum_{j=1}^n \cos\{t(X_j - x)\}\, e^{t^2\sigma^2/2}$. The integral was evaluated using Simpson's rule with sequential doubling of the grid size until successive iterations of $\hat{g}(x)$ differed by less than 10^{-5}. The computational procedure is accurate although it is slow.

In our study estimates were calculated over a 65-point grid spanning $[-3.5, 3.5]$. Sample size was set at $n = 2500$. With samples this large $AISE_g(\lambda)$ and $AISE_f(\lambda)$ are well-behaved and estimated bandwidths can be reliably computed by solving (4.2). Optimal bandwidths were also determined by minimizing the integrated squared error of g. Due to the large number of computations involved only 25 repetitions were performed. Figure 1 summarizes the findings of the simulation. Of the 25 density estimators, two were of significantly poorer quality than the rest due to estimated bandwidths which were much too small. Figure 1a contains an overlay plot of the remaining 23 estimates and gives a good idea of the variability inherent to the estimators.

Figure 1b displays the mean of the density estimates $\hat{f}_i$ and $\hat{g}_i$ over the same 23 observations alluded to above. The mean density estimates are very similar to their population counterparts, apart from the negativity in $\bar{g}$.

Figure 1c graphs the three most extreme estimated densities. These densities correspond to the largest and two smallest estimated bandwidths.

Figure 1d contains a scatter plot of estimated versus optimal bandwidths. The latter were determined by minimizing $\int(\hat{g} - g)^2 dx$. The discrete nature of the data is an artifact of the optimization procedures employed, which searched over the grid

$$(1.5, 1.475, ..., 0.75)\sigma/\left\{\log(2\pi\sigma^2 n^{1/2})\right\}^{1/2}, \quad \sigma^2 = 1/3, \quad n = 2500.$$

All bandwidths, estimated and optimal, fell within the boundaries of this grid. The correlation coefficient for these data is -0.458. Removing the maximum and minimum estimated bandwidths changes the correlation to -0.670. The negative correlation between estimated and optimal bandwidths is typical of bandwidth selection procedures.

Of the 25 estimates, 13 showed clear evidence of bimodality, 4 showed questionable evidence and the remaining 8 gave little or no evidence of bimodality. The marginal performance of $\hat{g}$ is consistent with the asymptotic results of Sections 2 and 3 even though the latter pertain specifically to nonnegative kernel estimators. The model is artificial only with regards to the loss of bimodality in the presence of measurement error. The signal-to-noise ratio is not extreme. Thus a reasonable conclusion is that deconvolution is generally going to be a viable technique only with very large sample sizes. And in these cases computational efficiency may dictate the choice of estimator, at least to some extent.

6. APPLICATIONS

The theoretical results and simulation evidence in the previous sections indicate that decon-

volution with normal errors will generally be feasible only with very large sample sizes and this limits its applicability. Furthermore, the need to specify the error density and the fact that our asymptotic results indicate widely varying performance under different error models, suggest that deconvolution may not be robust to choice of error model. We now examine the extent of these limitations in a particular application.

We consider estimating the density of long-term log daily saturated fat intake in women, using data from a study on the relationship of breast cancer incidence and dietary fat, see Jones, *et al.* (1987). We use the same 2888 observations on women of age less than 50 employed by Stefanski and Carroll (1989). Estimates of the error variance for these data suggest that as much as $50-75\%$ of the variability in the observed data may be due to measurement error. The simulation results suggest that for a sample of size 2888, deconvolving this much noise is problematic. However, it is possible to gain some insight into the data by deconvolving lesser amounts of noise.

In the example we used the sinc kernel and bandwidth selection procedure described in Section 4. Recall that for the sinc kernel, bandwidth selection is independent of the error density asymptotically, see Section 4 and the appendix. The deconvolved density estimator was computed under three different assumptions on the error density, normal (N), double exponential (DE) and hyperbolic cosine (HC) $\left((2/\pi)(e^t + e^{-t})^{-1}\right)$. These densities were chosen for their qualitatively different behaviour at the origin and in the tails, and because of their analytical tractability. In each case the densities were scaled to have common variance.

Because the estimators are not range preserving, in applications we suggest employing the positive projections of the estimators renormalized to integrate to one over the range of the data. Figures 2a and 2b display the resulting estimators $\hat{f}$, $\hat{g}_N$, $\hat{g}_{DE}$, and $\hat{g}_{HC}$ assuming that $\sigma_Z^2 = (1/5)\sigma_X^2$ and $(1/3)\sigma_X^2$ respectively. The density estimates are graphed over the range of the observed data. For the case $\sigma_Z^2 = (1/5)\sigma_X^2$, the three deconvolved densities are nearly identical. For the case of larger measurement error, distinctions between the three deconvolved densities are more noticeable. However, differences between the three estimates of g are still small relative to the differences between $\hat{f}$ and any one estimate of g.

Assuming that the additive symmetric error model is reasonable, both figures suggest the interpretation that the long left tail of $\hat{f}$ is due to an underlying bimodal g smoothed by convolution. However, the data are 24-hour recall measurements of log saturated fat intake (Stefanski and Carroll, 1989) and it seems prudent not to blindly accept the assumption of symmetric measurement error and the interpretations it renders. For example, it may be that a proportion of subjects systematically underreport foods high in saturated fats, resulting in a skewed or bimodal error

density. This would also account for the long left tail in $\hat{f}$. In fact, if the density, g, of "true" log saturated fat intake is approximately normal and the error distribution, h, is bimodal, then the deconvolved density estimates, calculated under the assumption of symmetric errors, would be approximating h not g.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support of the U. S. National Science Foundation, the U. S. Air Force Office of Scientific Research, the National Institutes of Health and the helpful comments of two referees.

REFERENCES

Blum, T. and Susarla, V. (1977). Estimation of a mixing distribution function. *Annals of Probability*, **5** 200-209.

Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association*, **83** 1184-1186.

Choi, K. and Bulgren, W. (1968). An estimation procedure for mixtures of distributions, *Journal of the Royal Statistical Society, Series B*, **30** 444-460.

Davis, K. B. (1975). Mean square error properties of density estimates. *Annals of Statistics*, **3** 1025-1030.

Deely, J. J. and Kruse, R. L. (1968). Construction of sequences estimating the mixing distribution. *Annals of Mathematical Statistics*, **39** 286-288.

Devroye, L. P. and Wise, G. L. (1979). On the recovery of discrete probability densities from imperfect measurements. *Journal of the Franklin Institute*, **307** 1-20.

Eddington, A. S. (1913). On a formula for correcting statistics for the effects of a known probable error of observation. *Mon. Not. R. Astrom. Soc.* **73** 359-360.

Fan, J. (1988). On the optimal rates of convergence for nonparametric deconvolution problem. Technical Report No. 157, Department of Statistics, University if California, Berkeley.

Gaffey, W. R. (1959). A consistent estimator of a component of a convolution. *Annals of Mathe-*

matical Statistics, **30** 198-205.

Hall, P. and Smith, R. L.(1986). Unfolding a nonparametric density estimate. Preprint.

Kahn, F. D. (1955). The correction of observational data for instrumental bandwidth. *Proceedings of the Cambridge Philosophical Society.* **51** 519-525.

Liu, M. C. and Taylor, R. L. (1989a). A consistent nonparametric density estimator for the deconvolution problem. Technical Report STA 73, University of Georgia.

Liu, M. C. and Taylor, R. L. (1989b). Simulations and computations of nonparametric density estimates for the deconvolution problem. Technical Report STA 74, University of Georgia.

Maritz, J. S. (1970). *Empirical Bayes Methods.* Methuen, London.

Mendelsohn, J. and Rice, J. (1982). Deconvolution of microfluorometric histograms with B splines. *Journal of the American Statistical Association*, **77** 748-753.

Preston, P. F. (1971). Estimating the mixing distribution by piecewise polynomial arcs. *Australian Journal of Statistics* **13** 64-76.

Silverman, B. W. (1981). Using kernel density estimators to investigate multi-modality. *Journal of the Royal Statistical Society, Series B*, **43** 97-99.

Stefanski, L. A. (1989). Rates of convergence of some estimators in a class of deconvolution problems. *Statistics & Probability Letters*, to appear.

Stefanski, L. A. and Carroll, R. J. (1989). Score tests in generalized linear measurement error models. Preprint.

Tapia, R. A. and Thompson, J. R. (1978). *Nonparametric Density Estimation.* John Hopkins University Press, Baltimore.

Taylor, C. C. (1982). A new method for unfolding sphere size distributions. *Journal of Microscopy*, **132** 57-66.

Trumpler, R. J. and Weaver, H. F. (1953). *Statistical Astronomy*, University of California Press, Berkeley.

Wise, G. L., Traganitis, A. P. and Thomas, J. B. (1977). The estimation of a probability density

function from measurements corrupted by Poisson noise. *IEEE Transactions on Information Theory*, **IT-23** 764-766.

APPENDIX

Using Parseval's Identity, the fact that $\phi_K(\bullet)$ is an indicator function, eveness of $|\phi_f(\bullet)|^2$ and $|\phi_h(\bullet)|^2$ and the relationship $E\{|\hat{\phi}(t) - \phi_f(t)|^2\} = \{1 - |\phi_f(t)|^2\}/n$ it can be shown that the mean integrated squared error of $\hat{g}$, $MISE_g(\lambda)$, is given by

$$MISE_g(\lambda) = c + \pi^{-1} \int_0^{1/\lambda} \frac{1 - (n+1)|\phi_f(t)|^2}{n|\phi_h(t)|^2} dt \qquad (A.1)$$

where c is a constant depending on f and h but not λ. The mean integrated squared error of $\hat{f}$, $MISE_f(\lambda)$, is obtained from (A.1) by setting $\phi_h(t) \equiv 1$.

Suppose λ_g and λ_f minimize $MISE_g$ and $MISE_f$ respectively. By the Fundamental Theorem of Calculus, $I_n(\lambda_g) = 0$ and $I_n(\lambda_f) = 0$ where

$$I_n(\lambda) = 1 - (n+1)|\phi_f(1/\lambda)|^2.$$

Assume that $|\phi_f(t)|^2$ is strictly decreasing on $[B, \infty)$ for some $B > 0$. Since λ_g and λ_f necessarily converge to zero, λ_g^{-1} and λ_f^{-1} are contained in the interval $[B, \infty)$ for sufficiently large n. However, for $\lambda \in (0, 1/B]$, $I_n(\lambda)$ is one-to-one under the assumption on $|\phi_f(t)|^2$ and thus the condition $I_n(\lambda) = 0$ uniquely determines λ. It follows that λ_g and λ_f are equal for sufficiently large n.

Note that $\hat{I}_n(\lambda)$ defined in (4.2) can be regarded as an unbiased estimating equation for λ_g and λ_f in the sense that $E\{\hat{I}_n(\lambda)\} = I_n(\lambda)$.

FIGURE LEGENDS

FIGURE 1. Simulation results: **1a.** Twenty-three estimates $\hat{g}$; **1b.** Means of twenty-three estimates of $\hat{f}$ (unimodal) and $\hat{g}$ (bimodal); **1c.** Three worst estimates $\hat{g}$; **1d.** Scatterplot of estimated bandwidths versus optimal bandwidths.

FIGURE 2. Saturated-fat example: **2a.** $\sigma_Z^2 = (1/5)\sigma_X^2$; **2b.** $\sigma_Z^2 = (1/3)\sigma_X^2$; $\hat{f}$, solid line; $\hat{g}_{DE}$, dashed line; $\hat{g}_{HC}$, dotted line; $\hat{g}_N$, solid line; Ordering at the primary mode, $\hat{f} < \hat{g}_{DE} < \hat{g}_{HC} < \hat{g}_N$, both cases.

Figure 1a

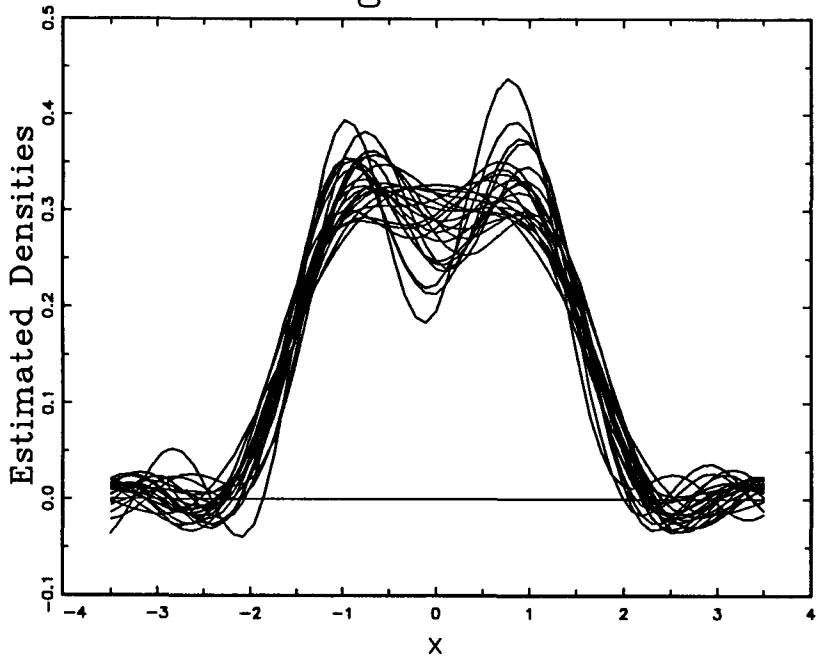

Figure 1b

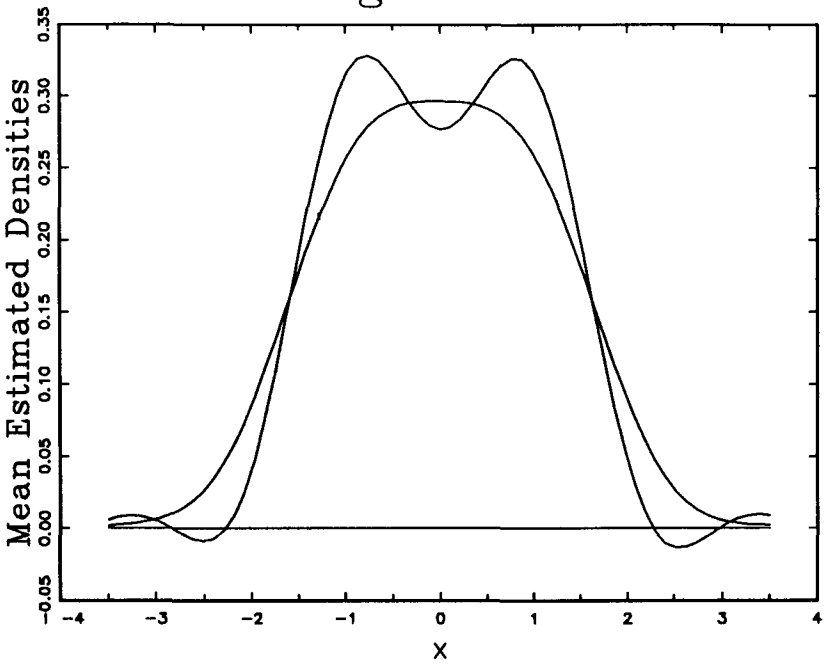

Figure 1c

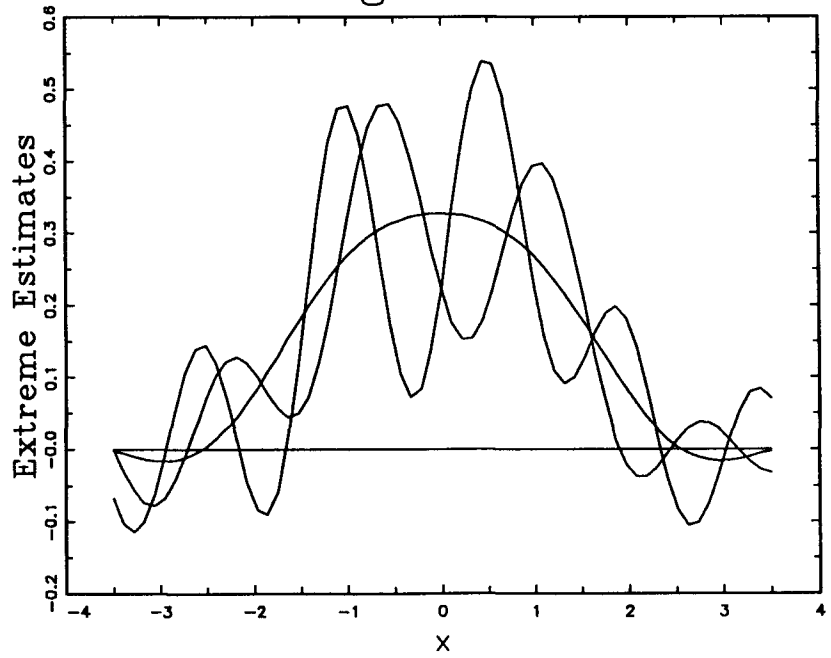

Figure 1d

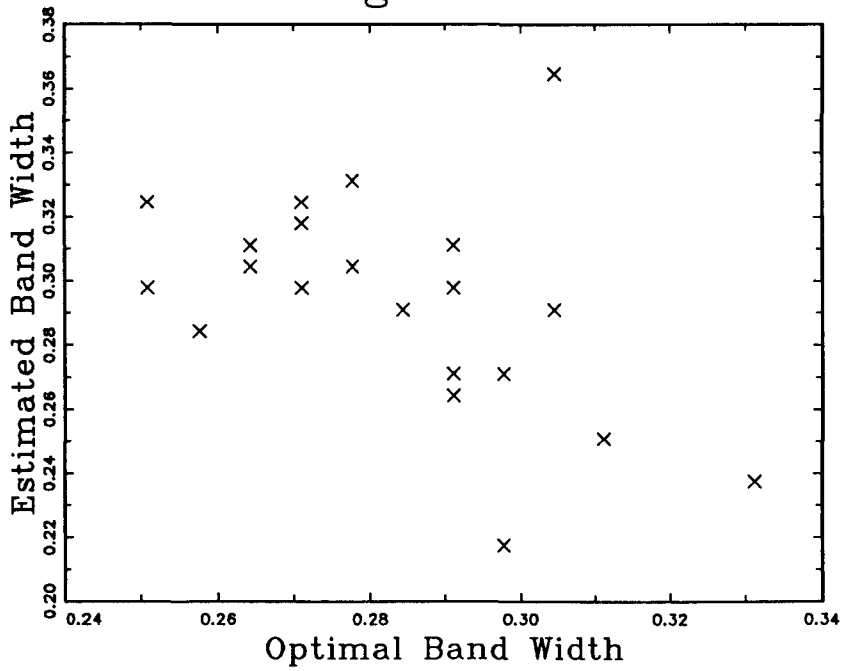

Figure 2a

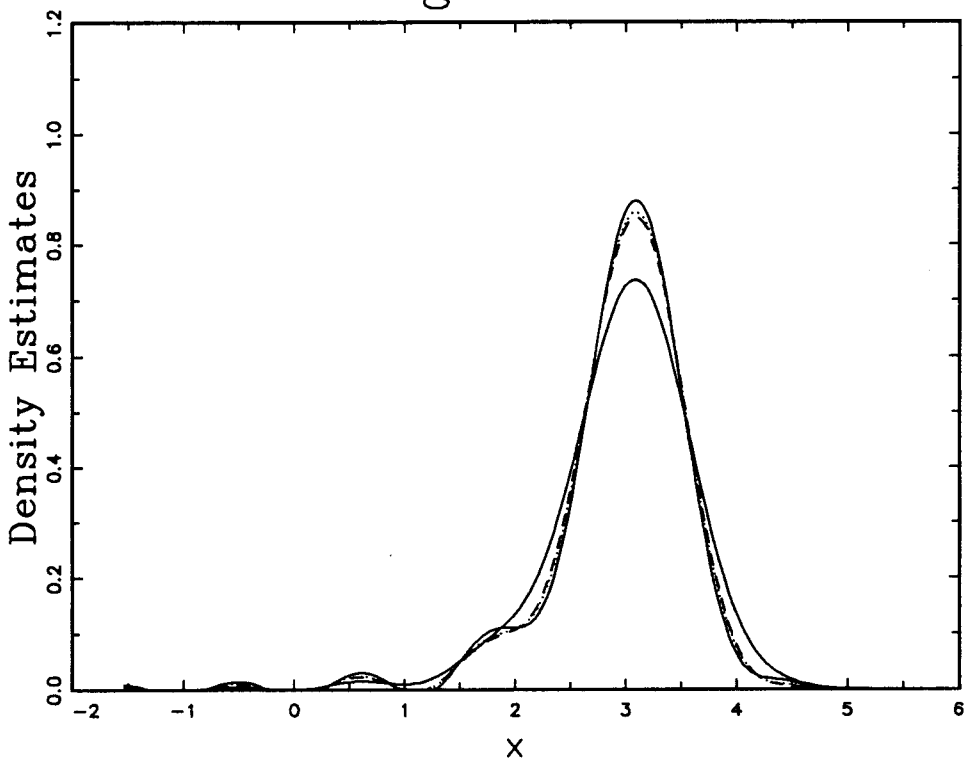
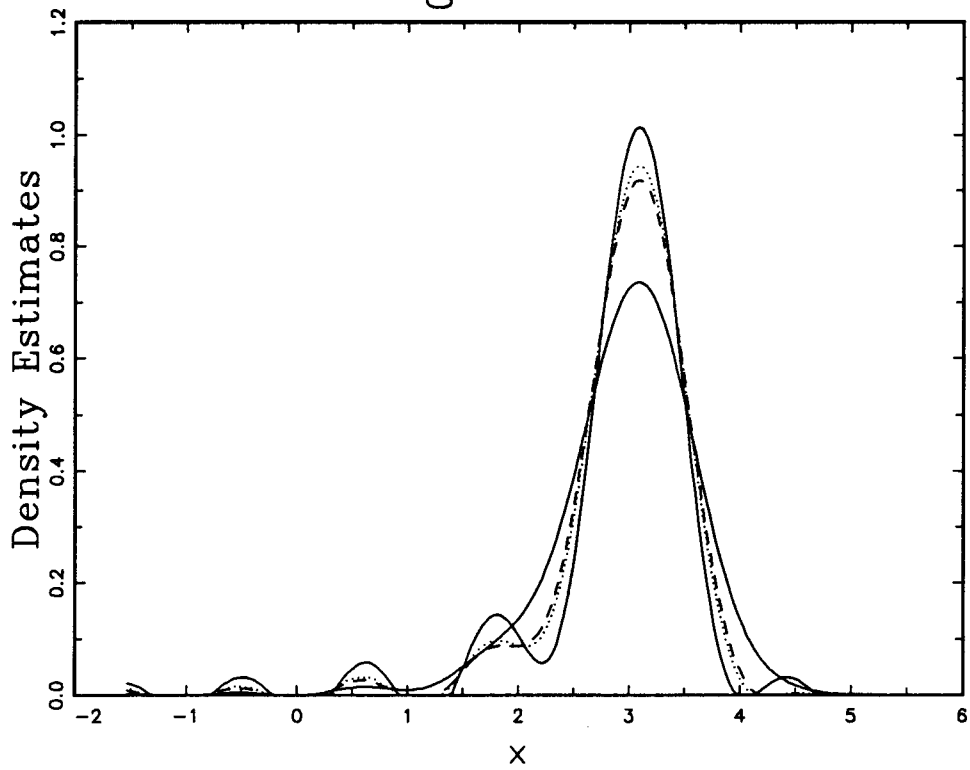
Figure 2b