Optimal Rates of Convergence for Deconvolving a Density

RAYMOND J. CARROLL and PETER HALL*

Suppose that the sum of two independent random variables X and Z is observed, where Z denotes measurement error and has a known distribution, and where the unknown density f of X is to be estimated. One application is the estimation of a prior density for a sequence of location parameters. A second application arises in the errors-in-variables problem for nonlinear and generalized linear models, when one attempts to model the distribution of the true but unobservable covariates. This article shows that if Z is normally distributed and f has k bounded derivatives, then the fastest attainable convergence rate of any nonparametric estimator of f is only $(\log n)^{-k/2}$. Therefore, deconvolution with normal errors may not be a practical proposition. Other error distributions are also treated. Stefanski–Carroll (1987a) estimators achieve the optimal rates. The results given have versions for multiplicative errors, where they imply that even optimal rates are exceptionally slow.

KEY WORDS: Deconvolution; Density estimation; Errors in variables; Measurement error; Rates of convergence.

1. INTRODUCTION

Suppose that we wish to gain information about the density f of a random variable X, but because of measurement error can only observe $Y = X + Z$, where the measurement error Z is independent of X. Assume Z has a known density function f_Z with characteristic function ϕ_Z. We address the following question: From a sample $Y_1, \ldots, Y_n$, how well can f be estimated?

Applied problems in which knowledge of f is required were discussed by Mendelsohn and Rice (1982) (see also Medgyessy 1977). Nonparametric estimates of f were discussed by Stefanski and Carroll (1987a).

An application of our results is to the nonparametric empirical Bayes problem (see Berger 1980; Maritz 1980). Here f represents the prior distribution for a sequence of location parameters $X_1, \ldots, X_n$. The idea is to estimate the prior nonparametrically, as opposed to the alternative method of specifying a parametric form for the prior with parameters to be estimated. We consider how well a prior can be estimated nonparametrically.

Another application is to the problem of measurement error models (errors in variables) for nonlinear regression and generalized linear models (see Stefanski and Carroll 1987b). Other recent articles include Carroll, Spiegelman, Lan, Bailey, and Abbott (1984), Stefanski and Carroll (1985), Stefanski (1985), and Schafer (1987). In this problem, X is the true predictor, but because of measurement error Z we can observe only $Y = X + Z$. Although Stefanski and Carroll (1985) and Stefanski (1985) use a sensitivity analysis approach, Carroll et al. (1984) and Schafer (1987) assume a specific distributional form for f. This article addresses how well the data can be used in a nonparametric way to suggest a parametric form for f. Schafer (1987) shows that in generalized linear models, the EM algorithm for maximum likelihood requires knowledge of the first two conditional moments of X, given Y and the response variable in the generalized linear model. Other problems require the conditional moments of X, given Y. In either case, how well these conditional moments can be estimated from data depends on how well f can be estimated from data.

The case of normal measurement error is particularly important. We show that if f has k bounded derivatives and errors are normal, then the fastest rate of convergence of any estimator of f is only $(\log n)^{-k/2}$, and that this rate is achieved by a kernel estimator of Stefanski–Carroll (1987a) type. This very slow rate suggests that deconvolution to get precise point estimates of f may not be a practical procedure with normal errors, even if optimal estimators are employed. With $k = 2$, it also follows that the best achievable rate for estimating the distribution function of X can be no faster than $(\log n)^{-3/2}$. Thus even estimating probabilities for X is difficult.

We emphasize that our results pertain to precise point estimation of the density f and its distribution function. It is likely that other quantities may be estimated much more precisely, such as conditional moments of X, given Y or the number of modes of f.

We also show that Stefanski–Carroll estimators attain optimal convergence rates for many other error distributions, such as gamma, exponential, and double-exponential. For example, the optimal achievable rate in the double-exponential case is $n^{-k/(2k+5)}$. Our results indicate that if the error density is compactly supported and infinitely differentiable then the optimal convergence rate is slower than n^{-a} for any $a > 0$. Deconvolving a density with smooth measurement error is intrinsically difficult, with convergence rates much slower than those usually encountered in density estimation.

These results have obvious implications for models with multiplicative error, $Y = XZ$, that may be expressed additively by taking logs. The density of $\log Z$ is infinitely differentiable in many important cases, such as when Z is gamma or lognormal, so convergence rates are extremely

* Raymond J. Carroll is Professor and Head, Department of Statistics, Texas A&M University, College Station, TX 77843. Peter Hall is Professor, Department of Statistics, Australian National University, Canberra, Australian Capital Territory 2601, Australia. Carroll's work was supported by the U.S. Air Force Office of Scientific Research and undertaken during a visit to the Australian National University. The authors thank Len Stefanski and Cliff Spiegelman for helpful conversations.

Journal of the American Statistical Association
December 1988, Vol. 83, No. 404, Theory and Methods

slow. Hence deconvolution is difficult when errors are multiplicative.

Of course, our lower bounds to convergence rates continue to apply when error distributions are known imperfectly—for example, when errors are normal with unknown variance. In such cases where the error distribution is specified up to estimable parameters, the distribution can often be estimated $n^{1/2}$ consistently by replication. Since estimators of the X-density f converge at rates considerably slower than $n^{-1/2}$, replacing the true error distribution by its estimated version does not measurably affect convergence rates of Stefanski–Carroll estimators. Hence both our lower and upper bounds to convergence rates apply when error distributions are imperfectly specified, up to a parametric form.

Section 2 gives details of our calculations in the case of normal measurement errors. In Section 3 we briefly discuss other error distributions.

2. DECONVOLUTION WHEN ERRORS ARE NORMAL

Write $C_k(B)$ for the class of k-times differentiable densities f having $\sup f \le B$ and $\sup|f^{(k)}| \le B$. Let X have density f, Z be normal $N(0, 1)$ independent of X, and $Y = X + Z$. The following theorem provides bounds to the accuracy with which $f \in C_k(B)$ can be estimated from an n sample of Y's.

Let x_0 be any real number, and $\hat{f}(x_0)$ be any nonparametric estimator of $f(x_0)$, based on an n sample of Y's.

Theorem 1. Assume that the error distribution is normal $N(0, 1)$. If for some sequence of positive constants $\{a_n, n \ge 1\}$ we have

$$\liminf_{n\to\infty} \inf_{f\in C_k(B)} P_f\{|\hat{f}(x_0) - f(x_0)| \le a_n\} = 1 \quad (2.1)$$

for each $B > 0$, then

$$\lim_{n\to\infty} (\log n)^{k/2} a_n = \infty. \quad (2.2)$$

Theorem 1 (see the Appendix for proof) declares that the rate of convergence of $\hat{f}$ to f cannot be faster than $(\log n)^{-k/2}$, over densities in $C_k(B)$. Kernel estimators attaining this rate of convergence were constructed by Stefanski and Carroll (1987a) and are given as follows. Let G be a symmetric function vanishing outside $(-1, 1)$, having $k + 2$ bounded derivatives on $(-\infty, \infty)$ and satisfying $G(t) = 1 + O(|t|^k)$ as $t \to 0$. Put $h \equiv (2/\log n)^{1/2}$, $G(w, h) \equiv (2\pi)^{-1} \int \cos(tw/h) G(t) \exp\{(t/h)^2/2\}\, dt$, and $\hat{f}(x) \equiv (nh)^{-1} \sum_j G(Y_j - x, h)$, where $\{Y_1, \ldots, Y_n\}$ is a random sample from the distribution of Y. The following result is an easy generalization to $k > 2$ of a result of Stefanski and Carroll (1987a).

Theorem 2. Assume that the error distribution is normal $N(0, 1)$. If the constants a_n satisfy (2.2) and $\hat{f}$ is the kernel estimator just defined, then (2.1) holds for each real number x_0 and each $B > 0$.

A referee has commented that the minimax nature of Theorem 1 may be unduly pessimistic. Further work with the Stefanski–Carroll estimator will be the judge of this concern. Cliff Spiegelman has conjectured that much better rates of convergence can be obtained if we limit consideration to smaller classes of densities, such as those confined to a known interval. Len Stefanski has also suggested that the small error approach of Stefanski and Carroll (1985) and Stefanski (1985) could be used to good effect in small samples.

The basic method of proof of Theorem 1 (see the Appendix) can be used to show that if $k = 2$, the distribution function of X can be estimated at a rate no faster than $(\log n)^{-3/2}$. Let F_n and F_0 be the distribution functions for f_n and f_0 in the proof of Theorem 1, and evaluate them at εx_0, where $x_0 > 0$. The calculations rely on an approximation to $H_{2l-1}(x_0)$ given by Magnus, Oberhettinger, and Soni (1966, p. 254), and various integral identities (p. 251). We omit the details. Using slightly different techniques, the same result has been obtained independently by Y. Ritov in an as-yet unpublished paper.

3. DECONVOLUTION FOR GENERAL ERRORS

There are versions of Theorems 1 and 2 for a variety of different types of error distributions. The general principle is: the smoother the residual distribution, the slower the optimal achievable rate of convergence. It is convenient to consider this principle in the Fourier domain, bearing in mind that smoother distributions have characteristic functions with thinner tails. If X, Y, and Z have respective characteristic functions ϕ_X, ϕ_Y, and ϕ_Z, and if $Y = X + Z$ where X and Z are independent, then the characteristic function of X is recoverable from that of Y via the formula $\phi_X = \phi_Y/\phi_Z$. Any data-based form of this inversion becomes increasingly difficult as the tails of ϕ_Z become thinner. For example, if Z has a gamma distribution with shape parameter α, then the tails of $\phi_Z(t)$ decrease like $|t|^{-\alpha}$ as $|t| \to \infty$, so deconvolution is difficult for large α. In fact, the fastest achievable rate of convergence over densities in $C_k(B)$ is $n^{-k/(2k+2\alpha+1)}$. This is made clear by the following analog of Theorem 1. Again, $\hat{f}(x_0)$ is a nonparametric estimator of $f(x_0)$.

Theorem 3. Assume that the error distribution is gamma with shape parameter $\alpha > 0$. If for some sequence of positive constants $\{a_n, n \ge 1\}$ we have $\liminf_{n\to\infty} \inf_{f\in C_k(B)} P_f\{|\hat{f}(x_0) - f(x_0)| \le a_n\} = 1$ for each $B > 0$, then

$$\lim_{n\to\infty} n^{k/(2k+2\alpha+1)} a_n = +\infty. \quad (3.1)$$

The "double gamma" case, where Z is symmetric and $|Z|$ is gamma(α), is similar. There, Theorem 3 continues to hold for integer α, provided 2α in (3.1) is changed to $4(\alpha - [\alpha/2])$, where $[\alpha/2]$ denotes the largest integer not exceeding $\alpha/2$. In particular, the optimal rate of convergence when errors have a double-exponential distribution is $n^{-k/(2k+5)}$.

Proofs of results such as Theorem 3, where algebraic rates are available, run as follows. Let $\varepsilon \to 0$ as $n \to \infty$, and fix a k-times differentiable density f_0 that is bounded away from 0 in a neighborhood of the origin. Let H be a

bounded, compactly supported function with at least k bounded derivatives, satisfying $H(0) \neq 0$ and $\int x^j H(x)\,dx = 0$ for $0 \leq j < \alpha + 1$. Put $f_n(x) \equiv f_0(x) + \varepsilon^k H(x/\varepsilon)$, and let g_n and g_0 be the convolution densities for f_0 and f_n, respectively. It may be shown that if $\varepsilon = n^{-1/(2k+2\alpha+1)}$, then I, defined at (A.6) (see the Appendix), satisfies $I = O(n^{-1})$. Then, arguing much as in the proof of Theorem 1, the best attainable rate of convergence emerges as being no faster than ε^k. Similar techniques show that for smooth, infinitely differentiable error densities such as the Cauchy, the optimal convergence rate is slower than n^{-a} for any $a > 0$.

Stefanski–Carroll–type kernel estimators achieve optimal rates in the normal, gamma, and double-gamma cases.

4. DISCUSSION

Deconvolution problems are important in their own right, as well as in nonparametric estimation of priors. In measurement error models, deconvolution arises if one wishes to use data either to suggest models for the unobservable predictors or to estimate conditional moments useful in likelihood calculations. When the measurement errors are normally distributed, our results are pessimistic, suggesting that it is difficult to deconvolve effectively over a wide class of distributions for X if one is interested in precise estimates of the true density f. Other functions of f may be estimated better, such as the conditional moments of X given Y or the general shape of f.

APPENDIX: PROOF OF THEOREM 1

To simplify notation, we relocate so that $x_0 = 0$ and rescale so that Z is normal $N(0, \frac{1}{2})$, with density $\psi(z) \equiv \pi^{-1/2} e^{-z^2}$. Let $\sigma \geq 1$, and write f_0 for the $N(0, \sigma^2)$ density; l for the integer part of $\log n$; $b_j \equiv 2^{-j}\{(2j)!\}^{-1/2} j^{1/4}$; $\eta \equiv l^{-k/2}\varepsilon^k \delta B$, where ε and $\delta \in (0, \frac{1}{2}]$ are fixed; and $H_0, H_1, \ldots$ for Hermite polynomials orthogonal with respect to ψ. The following properties are obtainable from Magnus et al. (1966, p. 252) and Sansone (1959, p. 324): $H_j(-x) = (-1)^j H_j(x)$;

$$\exp\{2x\varepsilon y - (\varepsilon y)^2\} = \sum_{j=0}^{\infty} H_j(x)(\varepsilon y)^j / j!; \tag{A.1}$$

$$\int H_i(x) H_j(x) e^{-x^2}\,dx = \pi^{1/2} 2^i i! \quad \text{if } i = j, \text{ 0 otherwise;} \tag{A.2}$$

$$\int H_{2i}(x) x^{2j} \psi(x)\,dx = (2j)!/\{4^{j-i}(j-i)!\}; \tag{A.3}$$

$$|b_l H_{2j}(x)\psi(x)| \leq C(1 + |x|^{5/2}) e^{-x^2/2}; \tag{A.4}$$

$$\eta \sup|(d/dx)^k b_l H_{2l}(x/\varepsilon)\psi(x/\varepsilon)| \leq C\,\delta B, \tag{A.5}$$

where C depends only on k.

Put $f_n(x) \equiv f_0(x) + \eta b_l H_{2l}(x/\varepsilon)\psi(x/\varepsilon)$. By (A.4), and since $\eta(n) \to 0$ and $\varepsilon < 1 \leq \sigma$, f_n is a density for large n. If X has density f_0 or f_n then $Y = X + Z$ has density g_0 or g_n, respectively, where g_0 is the $N(0, \sigma^2 + \frac{1}{2})$ density, $g_n(x) \equiv g_0(x) + \eta b_l h_l(x)$, and

$$\begin{aligned} h_l(x) &\equiv \int H_{2l}(y/\varepsilon)\psi(y/\varepsilon)\psi(x - y)\,dy \\ &= \varepsilon\psi(x) \sum_{j=l}^{\infty} H_{2j}(x)\varepsilon^{2j}\{4^{j-l}(j-l)!\}^{-1}, \end{aligned}$$

using (A.1) and (A.3). Since $\psi(x)^2/g_0(x) \leq C_1 e^{-x^2}$,

$$\begin{aligned} I &\equiv \int (g_n - g_0)^2 (g_0)^{-1} \leq C_1(\eta b_l)^2 \\ &\quad \times \int h_l(x)^2 e^{x^2}\,dx \\ &= C_2 \varepsilon^{2k+2}\,\delta^2 2^{2l}\{(2l)!\}^{-1} l^{1/2-k} \\ &\quad \times \sum_{j=l}^{\infty} (\varepsilon^4/4)^j (2j)!\{(j-l)!\}^{-2}, \end{aligned} \tag{A.6}$$

using (A.2). But $\{(2l)!\}^{-1} \leq C_3(l!)^{-2} 2^{-2l} l^{1/2}$, $(2j)! \leq C_3(j!)^2 2^{2j} j^{-1/2}$, and $j!/(j-l)! = \binom{j}{l} l! \leq 2^l l!$. Hence, remembering that $\varepsilon \leq \frac{1}{2}$,

$$\begin{aligned} I &\leq C_4 \varepsilon^{2k+2}\,\delta^2 l^{1-k} \sum_{j=l}^{\infty} (4\varepsilon^4)^j j^{-1/2} \\ &\leq C_5(\varepsilon, \delta) l^{1/2-k}(4\varepsilon^4)^l = o(n^{-1}). \end{aligned} \tag{A.7}$$

Given $B > 0$, we see from (A.4) and (A.5) that by choosing σ large and δ small, not depending on B, we may ensure that f_0 and $f_n \in C_k(B)$ for large n. For an event A, let $P_n(A)$ and $P_0(A)$ denote the probability of A under f_n and f_0, respectively. If $\{a_n\}$ satisfies (2.1), then by (A.7) and the Cauchy–Schwarz inequality

$$\begin{aligned} &[P_n\{|\hat{f}(0) - f_n(0)| \leq a_n\}]^2 \\ &\quad \leq P_0\{|\hat{f}(0) - f_n(0)| \leq a_n\}(1 + I)^n \\ &\quad = \{1 + o(1)\}P_0\{|\hat{f}(0) - f_n(0)| \leq a_n\}, \end{aligned}$$

so both $P_0\{|\hat{f}(0) - f_n(0)| \leq a_n\}$ and $P_0\{|\hat{f}(0) - f_0(0)| \leq a_n\}$ converge to 1 as $n \to \infty$. Hence $|f_n(0) - f_0(0)| \leq 2a_n$ for large n. But $|f_n(0) - f_0(0)| = \eta b_l (2l)!/l!\pi^{1/2} \geq 2CB(\log n)^{-k/2}$, where C does not depend on B. Therefore, $a_n \geq CB(\log n)^{-k/2}$ for large n. Since this is true for each $B > 0$, then $(\log n)^{k/2} a_n \to \infty$, completing the proof.

[*Received July 1987. Revised March 1988.*]

REFERENCES

Berger, J. O. (1980), *Statistical Decision Theory,* New York: Springer-Verlag.

Carroll, R. J., Spiegelman, C. H., Lan, K. K., Bailey, K. R., and Abbott, R. D. (1984), "On Errors-in-Variables for Binary Regression Models," *Biometrika,* 71, 19–25.

Magnus, W., Oberhettinger, F., and Soni, R. P. (1966), *Formulas and Theorems for the Special Functions of Mathematical Physics,* Berlin: Springer-Verlag.

Maritz, J. S. (1980), *Empirical Bayes Methods,* London: Methuen.

Medgyessy, P. (1977), *Decomposition of Superpositions of Density Functions and Discrete Distributions,* New York: John Wiley.

Mendelsohn, J., and Rice, R. (1982), "Deconvolution of Microfluorometric Histograms With *B* Splines," *Journal of the American Statistical Association,* 77, 748–753.

Sansone, G. (1959), *Orthogonal Functions,* London: Wiley Interscience.

Schafer, D. W. (1987), "Covariate Measurement Error in Generalized Linear Models," *Biometrika,* 74, 385–391.

Stefanski, L. A. (1985), "The Effects of Measurement Error on Parameter Estimation," *Biometrika,* 72, 583–592.

Stefanski, L. A., and Carroll, R. J. (1985), "Covariate Measurement Error in Logistic Regression," *The Annals of Statistics,* 13, 1335–1351.

——— (1987a), "Deconvoluting Kernel Density Estimators," preprint.

——— (1987b), "Conditional Scores and Optimal Scores for Generalized Linear Measurement-Error Models," *Biometrika,* 74, 703–716.