# NONPARAMETRIC REGRESSION ESTIMATION FROM DATA CONTAMINATED BY A MIXTURE OF BERKSON AND CLASSICAL ERRORS

Raymond J. Carroll
Department of Statistics, 3143 TAMU, Texas A&M University, College Station, Texas 77843, USA
carroll@stat.tamu.edu

Aurore Delaigle
Department of Mathematics, University of Bristol, Bristol BS8 1TW, UK and
Department of Mathematics and Statistics, University of Melbourne, Parkville, VIC, 3010, Australia
Aurore.Delaigle@bristol.ac.uk

Peter Hall
Department of Mathematics and Statistics, University of Melbourne, Parkville, VIC, 3010, Australia and Department of Statistics, University of California at Davis, Davis, CA 95616, USA
halpstat@maths.anu.edu.au

## ABSTRACT

Estimation of a regression function is a well known problem in the context of errors in variables, where the explanatory variable is observed with random noise. This noise can be of two types, known as classical or Berkson, and it is common to assume that the error is purely of one of these two types. In practice, however, there are many situations where the explanatory variable is contaminated by a mixture of the two errors. In such instances, the Berkson component typically arises because the variable of interest is not directly available and can only be assessed through a proxy, whereas the inaccuracy related to the observation of the latter causes an error of classical type. In this paper we propose a nonparametric estimator of a regression function from data contaminated by a mixture of the two errors. We prove consistency of our estimator, derive rates of convergence and suggest a data-driven implementation. Finite-sample performance is illustrated via simulated and real-data examples.

**Keywords**: Berkson errors, Deconvolution, Errors in variables, Kernel method, Measurement error, Orthogonal series, Radiation dosimetry, Smoothing parameter.

**Short Title**: Nonparametric Regression with Mixed Measurement Errors

# 1 INTRODUCTION

We consider nonparametric estimation of a regression function when the covariate is observed with a mixture of Berkson and classical measurement errors. Contamination by mixed errors arises frequently in toxicologic studies, where, for example, the goal is to relate the occurrence, $Y$, of a disease to the level of exposure, $X$, to a toxic substance. Typically, $X$ cannot be observed directly and can be assessed only by observing another variable, $L$, that is linearly related to it. The observations comprise a sample of independent and identically distributed random vectors $(L_j, Y_j)$, $1 \le j \le n$, generated by a so-called Berkson model

$$Y_j = g(X_j) + \eta_j\,, \quad X_j = L_j + U_{B,j}\,, \tag{1.1}$$

where $U_{B,j}$, $L_j$ and $\eta_j$ are mutually independent, $E(\eta_j \,|\, X_j) = 0$, $\mathrm{var}(\eta_j) < \infty$. In this setting, the variable $L$ is often referred to as a proxy or surrogate for $X$, and $U_B$ is an error of Berkson type. The model at (1.1) was first considered by Berkson (1950), and has been studied mostly in parametric or semiparametric settings. Recent related work includes that of Huwang and Huang (2000), Buonaccorsi and Lin (2002), Stram et al. (2002) and Wang (2003). See Delaigle, Hall and Qiu (2006) for a nonparametric treatment.

In most situations, the surrogate $L$ cannot be observed without measurement error, caused by the inaccuracy of the measurement process (device or experimenter, for example), and what we really observe are contaminated versions $W_j$ of $L_j$, $1 \le j \le n$, generated by the model

$$W_j = L_j + U_{C,j}\,, \tag{1.2}$$

where $U_{C,j}$ and $L_j$ are independent. The variable $U_C$ corresponds to a so-called classical measurement error, a type of error that has been studied extensively in the literature. Nonparametric methods for inference in settings such as this include kernel approaches (e.g. Fan and Masry, 1992; Taupin, 2001; Linton and Whang, 2002) and techniques based on simulation and extrapolation, or SIMEX, arguments (e.g. Cook and Stefanski, 1994; Stefanski and Cook, 1995; Carroll et al., 2006; Carroll et. al., 1999; Kim and Gleser, 2000; Devanarayan and Stefanski, 2002).

The Berkson and classical errors are very different in nature, and most existing methods focus exclusively on cases where the observations are contaminated by errors of only one of the two types. In this paper our interest is in estimating the regression function $g$ when both types of errors are present. In our setting we observe a sample of independent pairs $(W_j, Y_j)$, for $1 \le j \le n$, generated by

$$Y_j = g(X_j) + \eta_j\,, \qquad X_j = L_j + U_{B,j}\,, \qquad W_j = L_j + U_{C,j}\,, \tag{1.3}$$

where $U_{C,j} \sim f_C$, $U_{B,j} \sim f_B$, $L_j \sim f_L$ and $\eta_j$ are mutually independent, $E(\eta_j \,|\, X_j) = 0$, $\mathrm{var}(\eta) < \infty$, and the respective error densities $f_C$ and $f_B$ are known. This model has been studied by Reeves et al. (1998) in a parametric context of radon exposure, and by Mallick et al. (2002) in a semiparametric, Bayesian setting of radiation exposure from nuclear testing, see also Li, et al. (2007). In this paper we consider nonparametric estimation of the regression function $g$, for data generated by the model at (1.3). A good recent discussion of the origins of mixed Berkson and classical errors in the context of radiation dosimetry is given by Schafer and Gilbert (2006).

In Section 2 we introduce a kernel estimator of $g$, involving the characteristic functions of the errors $U_B$ and $U_C$; this methodology is appropriate when these quantities do not vanish. The procedure can also be used as a consistent method in the case of pure Berkson errors, and reduces to the approach of Fan and Truong (1993) when the errors are purely of classical type.

Nonparametric estimation of $g$ necessitates the selection of two bandwidths and a ridge parameter. In Section 3 we propose a cross-validation procedure for choosing these parameters in practice. We implement the fully data-driven method on simulated examples, to illustrate its finite-sample performance. Despite the considerable difficulty of the problem, we show that the results obtained in practice are quite good. We apply the procedure to a real-data example where the goal is to estimate the relation between radiation exposure and incidence of thyroid diseases.

Section 4 discusses theoretical properties of the regression estimator. We obtain upper bounds to a uniform rate of convergence of the estimator under models (1.1) and (1.3). These results emphasize the particular difficulty of the problem, especially when compared

to density estimation in this context: for estimating a density from a sample contaminated by mixed errors, Delaigle (2007) shows that the rates of convergence are the rates for classical errors, multiplied by a factor of improvement proportional to the smoothness of the Berkson error. In the case of regression estimators, however, the upper bound established by the theory indicates that the rates of convergence are the rates for classical errors, multiplied by a "degrading factor" proportional to the smoothness of the Berkson error.

Section 5 suggests an alternative nonparametric orthogonal series estimator, designed for cases where the function $g$ and the densities $f_L$ and $f_B$ are compactly supported. Technical details are collected into an appendix.

# 2 KERNEL METHOD

Assume we observe data $(W_j, Y_j)$, for $1 \le j \le n$, generated by the model (1.3) and define the function

$$a(\ell) \equiv E(Y \,|\, L = \ell) = \int g(\ell - u) f_{-B}(u)\, du\,, \tag{2.1}$$

where $f_{-B}$ denotes the density of $-U_B$. Here and below, unqualified integrals are taken over the whole real line. Write $a = b/f_L$, where

$$b(x) = a(x) f_L(x)\,. \tag{2.2}$$

We shall use the sample $(W_j, Y_j)$, for $1 \le j \le n$, to consistently estimate the functions $b$ and $f_L$, and obtain an estimator of $g$ by deconvolution through equation (2.1).

Given a density $f_Z$, write $f_Z^{\rm Ft}$ for the corresponding characteristic function. Let $K$ be a kernel function, chosen so that its Fourier transform $K^{\rm Ft}$ satisfies $K^{\rm Ft}(0) = 1$ and vanishes outside a compact interval (note that such kernels are fairly standard in deconvolution problems, see for example Fan and Truong (1993)). Given $h > 0$, put

$$K_Z(x) = K_Z(x \,|\, h) = \frac{1}{2\pi} \int e^{-itx}\, K^{\rm Ft}(t) \big/ f_Z^{\rm Ft}(t/h)\, dt\,,$$

where we shall take $Z = C$ or $-B$. Let $h_k > 0$ for $k = 1, 2, 3$. Estimators of $f_L$ and $b$ are given respectively by $\widehat{f}_L$ and $\widehat{b}$, where

$$\widehat{f}_L(x) = \frac{1}{nh_1} \sum_{j=1}^{n} K_C\Big(\frac{x - W_j}{h_1}\Big), \quad \widehat{b}(x) = \frac{1}{nh_2} \sum_{j=1}^{n} Y_j\, K_C\Big(\frac{x - W_j}{h_2}\Big), \tag{2.3}$$

and where $h$ in the formula for $K_C = K_C(\,\cdot\,|\,h)$ is taken as $h_1$ and $h_2$, respectively. In practice one would usually put $h_1 = h_2$. Define $\tilde{f}_L = \max(\widehat{f}_L, 0) + \rho$, where $\rho > 0$ denotes a ridge parameter. Then, $\widehat{a} = \widehat{b}/\tilde{f}_L$ is an estimator of $a$. Hence, by taking the inverse Fourier transform of $\widehat{a}^{\mathrm{Ft}} K^{\mathrm{Ft}}(h_3\,\cdot)/f_{-B}^{\mathrm{Ft}}$ ,

$$\widehat{g}(x) = \frac{1}{h_3} \int \widehat{a}(u)\; K_{-B}\Big(\frac{x - u}{h_3}\Big)\, du \tag{2.4}$$

can be taken to be our estimator of $g$.

When the distribution of $U_B$ is degenerate at zero, i.e. when the errors-in-variables are of classical type, $a = g$ and so our estimator $\widehat{g}$ is simply $\widehat{a} = \widehat{b}/\tilde{f}_L$. This is the well-known Fan and Truong (1993) kernel estimator in classical errors-in-variables regression, modified here only to include a ridge parameter. The latter is introduced so as to avoid problems with the denominator of $\widehat{a}$ at points $x$ where $\widehat{f}_L(x)$ is too close to zero.

When the distribution of $U_C$ is degenerate at zero, i.e. when the errors-in-variables are solely of Berkson type, $\widehat{f}_L$ and $\widehat{b}$ are standard kernel estimators, and in particular,

$$\widehat{f}_L(x) = \frac{1}{nh_1} \sum_{j=1}^{n} \mathcal{K}\Big(\frac{x - W_j}{h_1}\Big), \quad \widehat{b}(x) = \frac{1}{nh_2} \sum_{j=1}^{n} Y_j\, \mathcal{K}\Big(\frac{x - W_j}{h_2}\Big), \tag{2.5}$$

where $\mathcal{K}$ can be taken to be a conventional kernel. Using these alternative definitions of $\widehat{f}_L$ and $\widehat{b}$ we may continue to define $\widehat{g}$ by (2.4).

# 3 NUMERICAL PROPERTIES

## 3.1 A Data-Driven Method

We sought a cross-validation approach to choosing the three parameters $h_1$, $h_3$ and $\rho$. In our setting, the smoothing-parameter selection problem is made especially difficult by the fact that the variables $X_i$ and $L_i$ are not observable. Additionally, calculating $\widehat{g}$ is

a computationally intensive operation. We split the problem into two parts, selecting $(h_1, \rho)$ and $h_3$ separately, as follows.

Define

$$S_{k1}(\ell) = K_C\biggl(\frac{\ell - W_k}{h_1}\biggr) \Big/ \biggl\{ \sum_{k'=1}^n K_C\biggl(\frac{\ell - W_{k'}}{h_1}\biggr) + \rho \biggr\},$$
$$S_{k2}(x) = h_3^{-1} \int S_{k1}(\ell)\, K_{-B}\biggl(\frac{x - \ell}{h_3}\biggr)\, d\ell\,.$$

Ideally we would use a cross-validation (CV) approach, selecting $(h_1, \rho)$ as

$$(\widehat{h}_1, \widehat{\rho}) = \text{argmin}_{(h_1,\rho)} \sum_{j=1}^n \biggl\{ \frac{Y_j - \widehat{a}(L_j)}{1 - S_{j1}(L_j)} \biggr\}^2, \tag{3.1}$$

and then estimating $h_3$ by

$$\widehat{h}_3 = \text{argmin}_{h_3} \sum_{j=1}^n \biggl\{ \frac{Y_j - \widehat{g}(X_j)}{1 - n^{-1} \sum_{k=1}^n S_{k2}(X_k)} \biggr\}^2. \tag{3.2}$$

(Here we use a GCV procedure in order to reduce computational labour.) However, $L_j$ and $X_j$ are unobservable, and so we cannot calculate $S_{k1}(L_j)$, $\widehat{a}(L_j)$, $S_{k2}(X_j)$ and $\widehat{g}(X_j)$ directly. We suggest two ways of estimating the unknown quantities, and combine the two ideas to define our final procedure.

The first approach, motivated by the case where the error variances are small, is to simply ignore all error present in the data, i.e. replace all $L_j$'s and $X_j$'s by $W_j$'s, and replace $f_B^{\rm Ft}$ and $f_C^{\rm Ft}$ (in the definitions of $K_C$ and $K_{-B}$) by 1. The second possibility is to replace $e^{-itL_j/h_1}$ (respectively, $e^{-itX_j/h_3}$) in $K_C\{(L_j - W_k)/h_1\}$ (respectively, $K_{-B}\{(X_j - \ell)/h_3\}$) by $e^{-itW_j/h_1} K^{\rm Ft}(t) \,/\, f_C^{\rm Ft}(t/h_1)$ (respectively, $e^{-itW_j/h_3} f_B^{\rm Ft}(-t/h_3)/f_C^{\rm Ft}(-t/h_3)$), which has, asymptotically, the same expected value.

To gain more intuition, let $(Z, f, r, V, h)$ denote $(C, a, 1, L, h_1)$ or $(-B, g, 2, X, h_3)$. Then the $\nu$th procedure, $\nu = 1, 2$, just described amounts to replacing $S_{kr}(V_j)$ by $S_{kr;\nu}(W_j)$, this being the version of $S_{kr}(V_j)$ obtained by replacing $K_Z\{(V_j - \cdot)/h\}$ by $\widehat{K}_{Z,\nu}\{(W_j - \cdot)/h\}$, and $\widehat{f}(V_j)$ by $\widehat{f}_\nu(W_j) = \sum_{k=1}^n Y_k\, S_{kr;\nu}(W_j)$, where $\widehat{K}_{Z,1}\{(W_j - \cdot)/h\} = K\{(W_j - \cdot)/h\}$, $\widehat{K}_{-B,2}\{(W_j - \ell)/h_3\} = K_C\{(W_j - \ell)/h_3\}$ and

$$\widehat{K}_{C,2}\biggl(\frac{W_j - W_k}{h_1}\biggr) = (2\pi)^{-1} \int \exp\{-it\,(W_j - W_k)/h_1\}\, \frac{K^{\rm Ft}(t)}{f_C^{\rm Ft}(t/h_1)^2}\, dt\,.$$

We noticed in our simulations that the first procedure tended to select smoothing parameters that were too small, while the second tended to select too large values. The following approach combines the two approaches in a way which tends to remove this problem. (1) Choose

$$(\widehat{h}_1, \widehat{\rho}) = \text{argmin}_{(h_1,\rho)} \sum_{j=1}^{n} \left\{ w_1 \, \frac{Y_j - \widehat{a}_1(W_j)}{1 - S_{j1;1}(W_j)} + w_2 \, \frac{Y_j - \widehat{a}_2(W_j)}{1 - S_{j1;2}(W_j)} \right\}^2 ;$$

then, (2) with $w_2 = 0.8 \log\{1+0.695 \, (\sigma_Z/\widehat{\sigma}_L)^{0.2}\}$ and $w_1 = 1-w_2$, where $\sigma_Z^2$ is the variance of $U_Z$, for $Z = B$ or $C$, and $\widehat{\sigma}_L^2 = \widehat{\sigma}_W^2 - \sigma_C^2$ with $\widehat{\sigma}_W^2$ the empirical variance of $W$, put

$$\bar{h}_3 = \text{argmin}_{h_3} \sum_{j=1}^{n} \left\{ w_1 \, \frac{Y_j - \widehat{g}_1(W_j)}{1 - n^{-1} \sum_{k=1}^{n} S_{k2;1}(W_k)} + w_2 \, \frac{Y_j - \widehat{g}_2(W_j)}{1 - n^{-1} \sum_{k=1}^{n} S_{k2;2}(W_k)} \right\}^2 ;$$

and finally, (3) select $(\widehat{h}_1, \widehat{\rho}, \widehat{h}_3) = (\widehat{h}_1, \widehat{\rho}, (\sigma_B/\sigma_C)^{2/3} \, \bar{h}_3)$. The weight functions $w_1$ and $w_2$ were chosen empirically and are such that, when the error variance tends to zero, we select the smoothing parameters via the first procedure only, which, for errors tending to zero, is the same as the CV procedure that would be used in the error-free case. The correction applied to $\bar{h}_3$ at the third step of the procedure allowed us to improve the results in cases where one of the two errors was much larger than the other one.

## 3.2 Simulations

We applied the kernel method by generating samples $(W_1, Y_1), \ldots, (W_n, Y_n)$ according to model (1.3), where the regression function $g$ was one of the following curves: (i) $g(x) = (50x^2 + 10x + 25)^{-1}$ (sharp unimodal), (ii) $g(x) = \phi_{0,1.5}(4x) + \phi_{1,2}(4x) + \phi_{2,5}(4x)$ (asymmetric), and (iii) $g(x) = 5 \sin(2x) \exp(-16x^2/50)$ (sinusoidal), where $\phi_{\mu,\sigma}$ is the density of an $\text{N}(\mu, \sigma^2)$ variable.

For each example, the variable $L$ was either a centered normal or a multiple of $T_8$, a Student-t distributed variable with eigth degrees of freedom, and the error variables $U_B$ and $U_C$ were normal or Laplace variables, centered at zero and having a variance equal to 10% and 20%, respectively, of the variance of $L$. The variable $\eta$ was $\text{N}(0, \sigma_\eta^2)$, where $\sigma_\eta^2 = 0.1 \times \text{var}|g|$. Here, $\text{var}|g|$ was defined by $\text{var}(|g|) = \int_{q_{0.01}}^{q_{0.99}} (|g| - E|g|)^2/(q_{0.99} - q_{0.01})$,

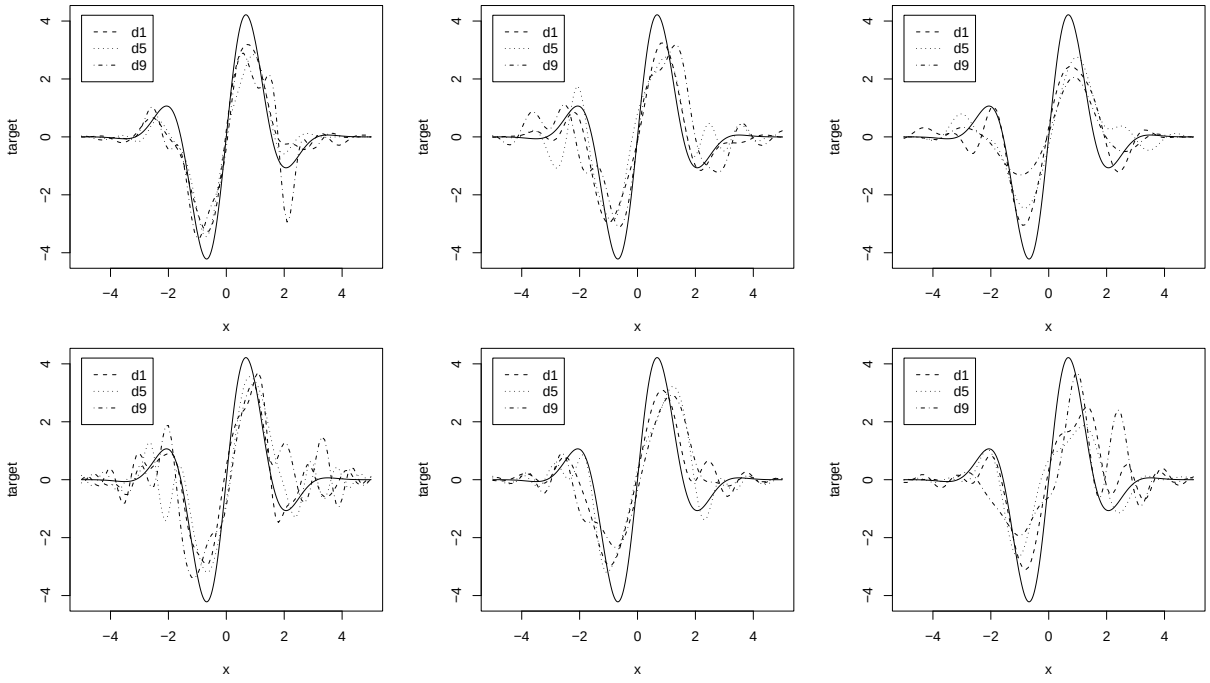


Figure 1: Estimation of function (iii) for samples of size $n = 250$, when $L \sim \sqrt{0.75}\, T_8$, $U_B$ and $U_C$ are Laplace (row 1) or normal (row 2), when $(\sigma_B^2/\sigma_L^2, \sigma_C^2/\sigma_L^2)$ is, from left to right, $(0.1, 0.1)$, $(0.1, 0.2)$, $(0.2, 0.2)$. The solid curve is the target curve.

where $E(|g|) = \int_{q_{0.01}}^{q_{0.99}} |g|/(q_{0.99} - q_{0.01})$ and $q_\alpha$ was the $\alpha$th quantile of $|g|$ rescaled to integrate to 1.

In each case, we considered samples of size $n = 100$ or 250, we generated 200 replicated samples from the random vector $(W, Y)$, and we constructed the corresponding estimator $\widehat{g}$, using the data-driven method of Section 3.1 and the kernel $K$ with Fourier transform $K^{\mathrm{Ft}}(t) = (1-t^2)^3\, 1_{[-1,1]}(t)$, which is commonly used in deconvolution problems. We report the Integrated Squared Error, $\mathrm{ISE}(x) = \int \{\widehat{g}(x) - g(x)\}^2\, dx$. In all figures, the estimates shown correspond to the first ($d1$), fifth ($d5$) and ninth ($d9$) deciles of the ordered values of ISE. We present only a portion of the results; the conclusions are also supported by the simulations not presented here.

In deconvolution problems it is rather common to consider two classes of errors, called ordinary-smooth and supersmooth errors. Roughly, an error of the first (respectively, second) type has a characteristic function behaving like a negative polynomial (respectively, exponential) in the tails. Rates of convergence in errors-in-variables problems are

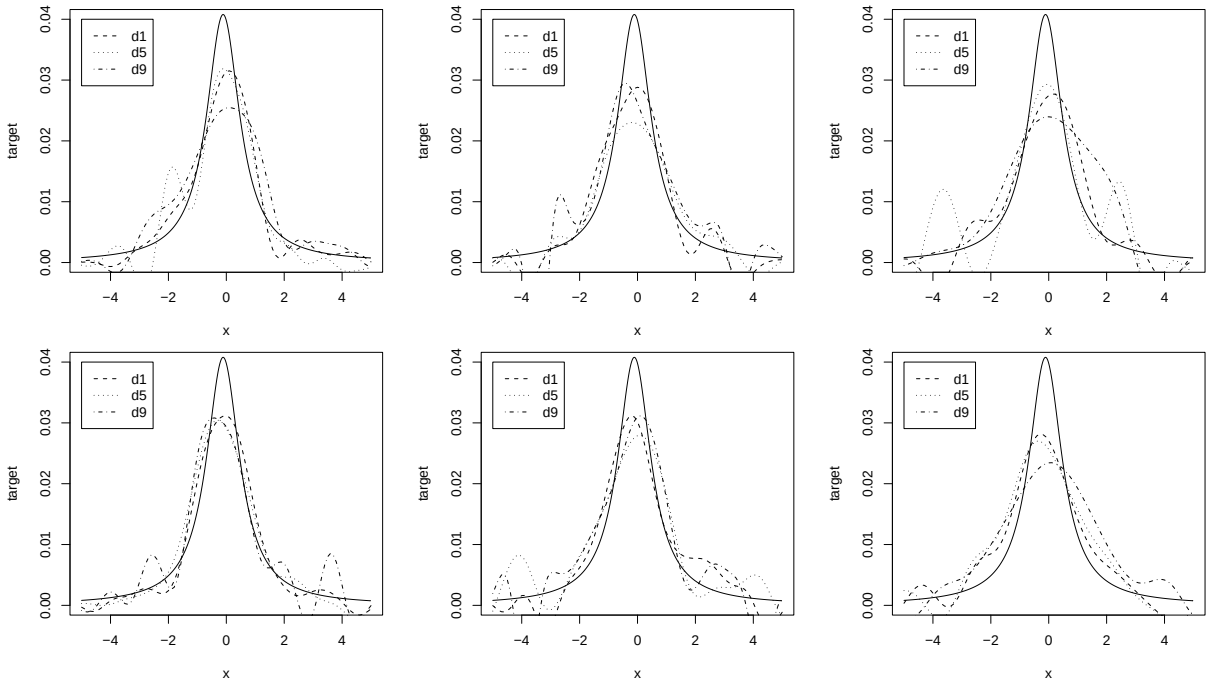


Figure 2: Estimation of function (i) for samples of size $n = 100$ (row 1) or $n = 250$ (row 2), when $L \sim \text{Normal}(0, 2)$, $U_B$ and $U_C$ are Laplace with $(\sigma_B^2/\sigma_L^2, \sigma_C^2/\sigma_L^2)$ equal to, from left to right, $(0.1, 0.1)$, $(0.1, 0.2)$, $(0.2, 0.2)$. The solid curve is the target curve.

typically algebraic in the first case, and logarithmic in the second, and the results of Section 4 can be extended to show that such rates also hold in our case. We illustrate this fact by comparing the results obtained when estimating curve (iii), in the case where $L \sim \sqrt{0.75}\, T_8$, and $U_B$ and $U_C$ are both Laplace (ordinary-smooth) or both normal (supersmooth), for samples of size $n = 250$. The pair of variance ratios, $(\sigma_B^2/\sigma_L^2, \sigma_C^2/\sigma_L^2)$, equals $(0.1, 0.1)$, $(0.1, 0.2)$ or $(0.2, 0.2)$. The graphs in Figure 1 indicate that, although the method also works in the case of normal errors, the results are more variable than for Laplace errors. Other simulation results, not reported here, show that the Laplace error case systematically outperforms the normal error case, which occasionally performs very poorly, especially when $\sigma_B^2$ is large.

Figure 2 illustrates the way in which the estimator improves as sample size increases. We compare the results obtained when estimating curve (i) for samples of size $n = 100$ or $n = 250$. Here, $L \sim \text{Normal}(0, 2)$, $U_B$ and $U_C$ are Laplace, and we consider several values of $(\sigma_B^2/\sigma_L^2, \sigma_C^2/\sigma_L^2)$. As expected, the graphs show a clear improvement in the quality of

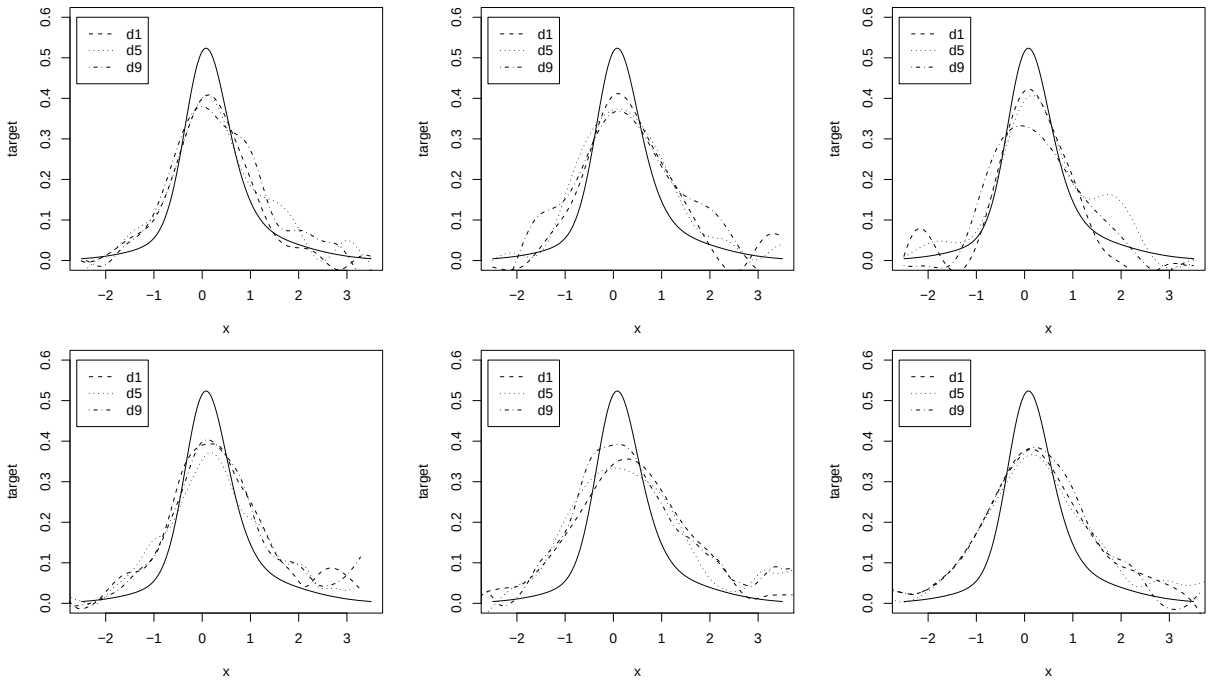


Figure 3: Estimation of function (ii) for samples of size $n = 250$, when $L \sim \text{Normal}(0, 1)$, $U_B$ is Laplace and $U_C$ is normal, with $(\sigma_B^2/\sigma_L^2, \sigma_C^2/\sigma_L^2)$ equal to, from left to right, $(0.1, 0.1)$, $(0.1, 0.2)$, $(0.2, 0.1)$. Row 1 shows the estimator (2.4) and row 2 shows the local linear estimator that ignores the error in the data. The solid curve is the target curve.

the estimators, in all cases, as $n$ increases from 100 to 250.

Figure 3 illustrates the performance of the estimator in a case where the classical error is smoother than the Berkson error — a situation encountered very often in real-data applications. We compare the results obtained when estimating curve (ii) for different values of $(\sigma_B^2/\sigma_L^2, \sigma_C^2/\sigma_L^2)$, when the classical error $U_C$ is normal, i.e. supersmooth, and the Berkson error $U_B$ is Laplace, i.e. ordinary-smooth. Here, $L \sim \text{Normal}(0, 1)$ and the 200 generated samples are of size $n = 250$. Here, and also in all other cases we considered, the best results are clearly in the case of the lowest error variance, i.e. $\sigma_B^2 = \sigma_C^2 = 0.1\sigma_L^2$. In the figure we also illustrate the effect of the errors present in the data; we show the local-linear estimators obtained when ignoring the error, i.e. when using the procedure with plug-in bandwidth described by Fan and Gijbels (1995). The graphs show that ignoring the error leads to severely biased estimators.

Of course, as for any nonparametric method in the usual 'error-free' regression prob-

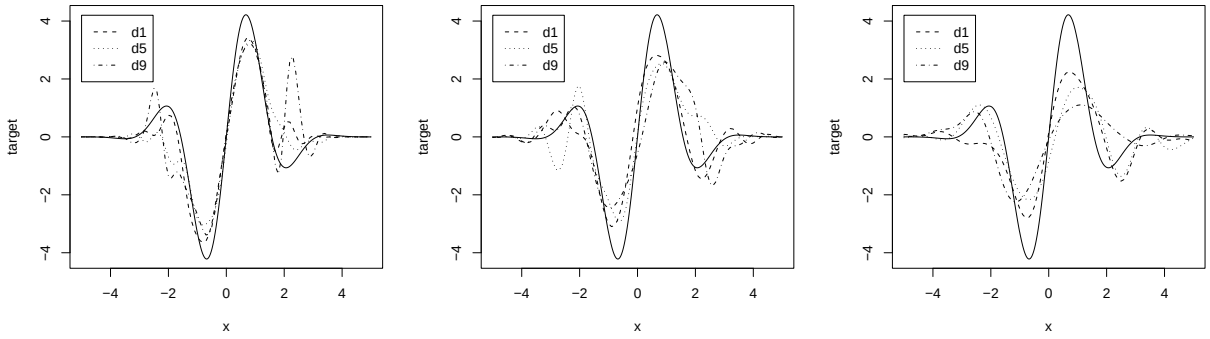


Figure 4: Estimation of function (iii) when $n = 250$, $L \sim \text{Normal}(0, \sigma_L^2)$ and $Z \sim \text{Laplace}\,(\sqrt{0.5}\,\sigma_Z)$, with $Z = U_B$ or $U_C$, for $(\sigma_L^2, \sigma_B^2, \sigma_C^2)$ equal to, from left to right, $(0.5, 0.05, 0.1)$, $(1, 0.1, 0.2)$ or $(2, 0.2, 0.4)$. The solid curve is the target curve.

lem, the quality of the estimator also depends on the range of the observed sample. In particular, for a given family of densities $f_B$, $f_C$ and $f_L$, and given noise-to-signal ratios $\sigma_B^2/\sigma_L^2$ and $\sigma_C^2/\sigma_L^2$, the performance of the estimator depends on the variance of $U_B$, $U_C$ and $L$. For example, Figure 4 illustrates the results of estimating regression function (iii) in the case where $n = 250$, $L \sim \text{Normal}(0, \sigma_L^2)$ and $Z \sim \text{Laplace}\,(\sqrt{0.5}\,\sigma_Z)$, with $Z = U_B$ or $U_C$, for $(\sigma_L^2, \sigma_B^2, \sigma_C^2) = (0.5, 0.05, 0.1)$, $(1, 0.1, 0.2)$ or $(2, 0.2, 0.4)$. When the variances are smaller, the observations are more concentrated around the centre, and, as a consequence, it is easier to recover the peaks of the curve. As the variances increase the observations become more widespread, and, for a given sample size, it becomes harder to recover the peaks of the regression curve, since the peaks are located around the centre zero.

Finally, we apply our method to a case where the function $g$ is unbounded. We take $g(x) = x^2$, $L \sim \text{N}(0, 1)$ and $U_B$ and $U_C$ are Laplace, with $\sigma_B^2 = \sigma_C^2 = 0.1$. Here, since $|g|$ integrates to infinity, we alter the definition of $q_{0.01}$ and $q_{0.99}$ in $E|g|$ and $\text{var}|g|$, and take $q_{0.99} = -q_{0.01} = 2.5$, corresponding approximately to the 0.99 quantile of the distribution of $L$. In Section 4.3 we shall show that, although it seems quite hard to deal with such unbounded functions $g$, our estimator is able to estimate $g$ on a compact interval, of length growing with the sample size. In Figure 5, we illustrate these results by showing the decile curves obtained for samples of size $n = 100$, 250 and 500. We see clearly that, as the sample size increases, the estimator is able to estimate $g$ correctly on growing intervals.

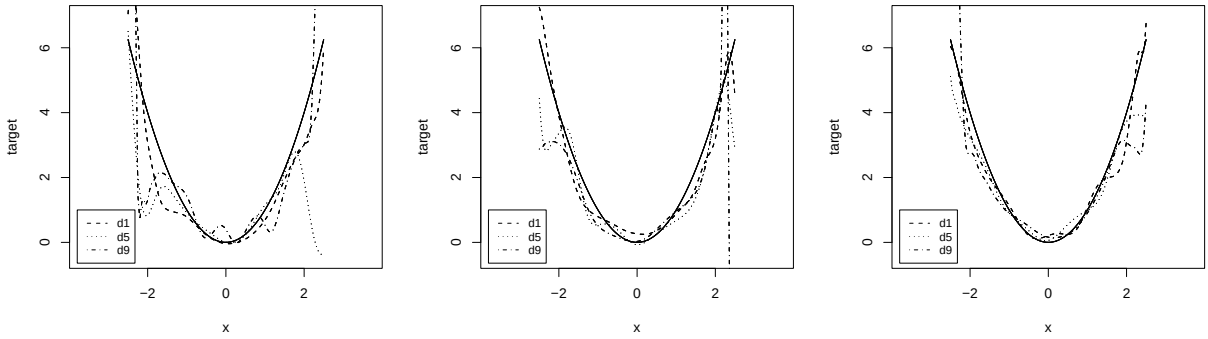


Figure 5: Estimation of the curve $g(x) = x^2$ when $L \sim \text{Normal}(0, 1)$ and $Z \sim \text{Laplace}$ $(\sqrt{0.05})$, with $Z = U_B$ or $U_C$ and sample size is, from left to right, $n = 100$, 250 or 500. The solid curve is the target curve.

Note that we show the estimated curves over a relatively large range, since the interval $[-2.5, 2.5]$ contains $L$ with a probability of 0.988.

## 3.3 Data Example

We applied the kernel method to data from the Nevada Test Site (NTS) Thyroid Disease Study; see, for example, Stevens et al. (1992), Kerber et al. (1993) and Simon et al. (1995). The goal of the study was to relate radiation exposure (largely due to above-ground nuclear testing in the 1950s) to various thyroid disease outcomes. In the Nevada study, over $2,000$ individuals exposed to radiation as children were examined for thyroid disease. The primary radiation exposure came from milk and vegetables. A recent update of the dosimetry is available (Simon et al. 2005), as is a reanalysis of the thyroid disease data (Lyon et al. 2006). We analyze a subset of the revised dosimetry data, namely the 1,278 women in the study, 103 of whom developed thyroiditis.

In this example, $X$ (resp., $W$) is the logarithm of the true (resp., observed) radiation exposure and $Y = 0$ or 1 indicates absence or presence of thyroid disease. As discussed in Mallick, et al. (2002), the uncertainties in this problem are a mixture of classical and Berkson measurement errors. Following the illustrative analysis of Mallick et al. (2002), in this illustration we assume that 50% of the total uncertainty variance is classical, and 50% is Berkson. Also, as in their analysis and those of many others in the area, the Berkson and classical uncertainties in the log-scale are assumed to be normally distributed.

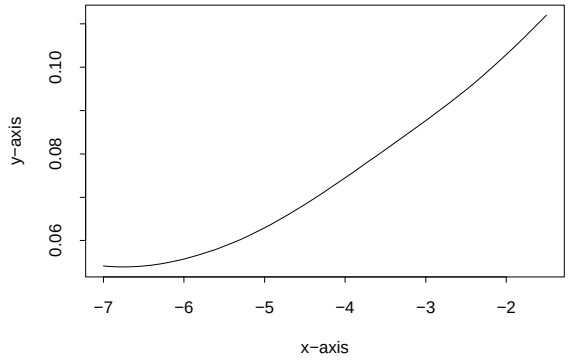


Figure 6: Estimation of the regression curve for the thyroid data in the log-scale. The x-axis is the logarithm of true dose, while the y-axis is the estimated risk of thyroiditis.

We applied our estimation procedure on these data, with smoothing parameters selected via the method described in Section 3.1, with the kernel $K$ as in Section 3.2. The estimator of the regression curve $P(Y = 1|X = x)$ is shown at Figure 6, for values of $x$ in the range $[-7, -1.5]$, which corresponds to the values between the 10th and 90th percentiles of the sample of observed log-doses. The graph shows a continuous, roughly quadratic increase in risk as the true log-doses of radiation increase. Our results are roughly in accord with the parametric analysis of Lyon, et al. (2006), although their use of an excess relative risk model is less flexible than ours.

# 4 THEORETICAL PROPERTIES

## 4.1 Discussion of case of unbounded $g$

In the simple error-free case (i.e. the case where $U_B \equiv U_C \equiv 0$), nonparametric estimation of an unbounded regression curve $g$ defined on the whole real line is a hard problem: in finite samples, the observations are confined to a finite range and in general, only the observations in the neighborhood of the point $x$ where we want to estimate $g$ bring valuable information about the value of $g(x)$. Hence, unless $g(x) \to 0$ as $x \to \infty$, it is usually not possible to construct a good estimator of $g$ outside the range of the observed data.

One of the interesting aspects of errors-in-variables problems is that nonparametric

inference cannot be undertaken in a strictly local sense. In particular, to estimate $g$ at $x$ it is not adequate to rely on noisy observations of $g$ at points close to $x$; observations of $g$ across its support are used to estimate $g$ at a point in the middle of the support. However, especially when the errors are of Berkson type, use of data on an unbounded, infinitely supported function $g$ can involve significant challenges: in finite samples, it is impossible to observe values of $L$ over more than just a finite range, and hence to obtain information across the whole support of $g$.

Consider, for example, the case where $g$ is unbounded and the distribution of the errors $U_B$ has unbounded support. Specifically, assume that the proxy variable $L$ is compactly supported; that the Berkson error $U_B$ has a Laplace distribution, $P(|U_B| > x) = e^{-x}$ for $x > 0$; and that $g(x) = e^{x^2}$. Because the tails of the distribution of $U_B$ decrease more slowly than the tails of $g$ increase, then with high probability, any sample of data on $Y = g(L + U_B) + \eta$ contains many very large values. In particular, for each $D_1 > 0$ and $D_2 \in (0, 1)$, the probability that $g(L_j + U_{B,j}) + \eta_j > n^{D_1}$ for at least $n^{D_2}$ values of $j$ in the range $1 \le j \le n$, converges to 1 as $n \to \infty$. In such instances, the observations on $Y$ are too 'volatile' and the estimator can turn out to be extremely unstable.

Circumstances as extreme as this are awkward to accommodate. One way of avoiding this type of difficulty is to restrict attention to the case of bounded $g$, but that prevents us from treating relatively standard cases such as, for example, polynomial $g$. Although the problem can be very hard, we show below that our estimator can in fact be used for such unbounded $g$; moreover, and perhaps surprisingly, the only way in which our estimator is affected by the fact that, in finite samples, we can only observe data on $L$ over a finite range, is that we can only guarantee consistent estimation of $g$ over a finite, but growing with $n$, interval. We shall prove consistency of $\widehat{g}$ by exploiting its similarities with $\widehat{g}_n$, the estimator of the function $g_n$, which we define as the restriction of $g$ over a finite, but growing, interval. In situations less extreme than the one mentioned in the previous paragraph, $\widehat{g}$ and $\widehat{g}_n$ are sufficiently close for asymptotic properties of $\widehat{g}$ to be derivable from those of $\widehat{g}_n$. More precisely, we assume that:

$$\text{the distribution of } U_B \text{ has all moments finite, and } |g(x)| \le D_3\, x^{D_4} \text{ for all } x, \text{ where } D_3, D_4 > 0 \text{ are constants,} \tag{4.1}$$

and we define $g_n = g \cdot 1_{[-n^{D_5}, n^{D_5}]}(x)$, where, for a set $A$, $1_A(x) = 1$ if $x \in A$ and 0 otherwise. If $\mathcal{R}$ is any compact set, then it can be proved from (4.1) that,

$$\begin{aligned}&\text{for any } D_5 > 0, \text{ no matter how small, } P\{\widehat{g}_n(x) = \widehat{g}(x) \text{ for all } x \in \mathcal{R}\} = \\ &1 - O(n^{-D_6}) \text{ for any } D_6 > 0, \text{ no matter how large.}\end{aligned} \tag{4.2}$$

Provided (4.1) holds, for any given $D_7 > 0$, no matter how small, we can, by choosing $D_5 > 0$ sufficiently small, ensure that,

$$\sup |g_n| = O\left(n^{D_7}\right) \quad \text{and} \quad \int |g_n| = O\left(n^{D_7}\right) . \tag{4.3}$$

We will see at the end of Subsection 4.3 that, together, (4.1)–(4.3) permit us to deal with the case of unbounded, infinitely supported $g$ by working with its truncated version $g_n$. This approach motivates the assumption, influencing (4.4) below, that $g$ may depend on $n$.

The assumption in (4.1) that all moments are finite is satisfied by the most common error distributions. The condition $|g(x)| = O(x^{D_4})$ asks only that $g$ increase no more than polynomially fast, which is a mild constraint.

## 4.2 Notation and Assumptions

Motivated by the arguments in Subsection 4.1, we shall permit $g = g_n$ to depend on $n$, subject to satisfying:

$$\max\left(\sup |g|, \int |g|\right) \le \lambda = \lambda(n)\,, \tag{4.4}$$

where $\lambda \ge 1$. It follows that $a$ and $b$, at (2.1) and (2.3), can also depend on $n$, although to avoid sub-subscripts we do not express this in notation. We adopt a conventional, fixed-function interpretation of $f_B$, $f_C$, $f_L$ and the distribution of $\eta$.

Biases for the estimators $\widehat{f_L}$ and $\widehat{b}$, defined at (2.3), are respectively given by

$$\begin{aligned}\text{bias}_f(x) &= E\{\widehat{f_L}(x) - f_L(x)\} = \frac{1}{2\pi} \int f_L^{\text{Ft}}(t) \left\{K^{\text{Ft}}(h_1 t) - 1\right\} e^{-itx}\, dt\,, \\ \text{bias}_b(x) &= E\{\widehat{b}(x) - b(x)\} = \frac{1}{2\pi} \int b^{\text{Ft}}(t) \left\{K^{\text{Ft}}(h_2 t) - 1\right\} e^{-itx}\, dt\,.\end{aligned}$$

Define too

$$\text{bias}_g(x) = \frac{1}{2\pi} \int g^{\text{Ft}}(t) \left\{ K^{\text{Ft}}(h_3 t) - 1 \right\} e^{-itx} \, dt \,,$$

it being assumed in each case that the integral is convergent in the Riemann sense. To interpret $\text{bias}_g$, consider the case where $g$ is a probability density, and we observe noisy data generated as $\zeta = \eta_g + \eta$, where $\eta_g$ has density $g$, and $\eta$ is independent of $\eta_g$ and has a known distribution with a characteristic function that does not vanish on the real line. Then, $\text{bias}_g$ represents the bias of the standard deconvolution kernel estimator of $g$ with bandwidth $h_3$.

Taking, for simplicity, $h_1 = h_2$, let $\text{supbi}(h_1)$ denote the maximum of the suprema of the biases $\text{bias}_b$ and $\text{bias}_f$, and define also $\delta$, closely related to root mean squared error:

$$\text{supbi}(h_1) = \max_{c=b,f} \sup_{-\infty<x<\infty} |\text{bias}_c(x)| \,, \quad \delta = \lambda^{-1} \, \text{supbi}(h_1) + (n h_1^{2\alpha+1})^{-1/2} \,,$$

where $\lambda$ is as at (4.4) and $\alpha > 1$ will be determined by (4.6); let $\mathcal{R}$ denote a finite union of compact intervals on which $f_L$ is bounded away from zero; and assume that

> (a) $f_L$ is uniformly bounded; (b) there exists an open set $\mathcal{S}$ containing $\mathcal{R}$, such that $f_L$ is bounded away from zero on $\mathcal{S}$; and (c) for a constant $\xi \in (0, \infty]$, $f_L(x) \geq C_1 \, (1 + |x|)^{-\xi}$ for all $|x|$. (4.5)

We permit $\xi = \infty$ in (4.5), in which case (4.5)(c) is degenerate and only (4.5)(a) and (4.5)(b) are effective.

Next we state assumptions about the known densities $f_C$ and $f_B$, the kernel $K$ and the regression mean $g$; see (4.6)(a)–(d), respectively. With $C_2 > 0$ denoting a constant and $\lfloor \beta \rfloor$ the integer part of $\beta$, our assumptions are:

> for constants $\alpha, \beta, \gamma$ such that $\alpha, \beta > 1$ and $\beta + \gamma \geq 1$ is an integer, (a) $|(d/dt)^j \, f_C^{\text{Ft}}(st)^{-1}| \leq C_2 \, s^j \, (1 + st)^{\alpha - j}$, for $j = 0$ and 1, all $|t| \leq 1$ and all $s \geq 1$; (b) $|(d/dt)^j \, f_{-B}^{\text{Ft}}(st)^{-1}| \leq C_2 \, s^j \, (1 + st)^{\beta - j}$ when $0 \leq j \leq \min(\beta, \beta + \gamma + 1)$, and $|(d/dt)^j \, f_{-B}^{\text{Ft}}(st)^{-1}| \leq C_2 \, s^{\lfloor \beta \rfloor}$ when $\min(\beta, \beta + \gamma + 1) < j \leq \max(\beta, \beta + \gamma + 1)$, for all $|t| \leq 1$ and all $s \geq 1$; (c) $|(d/dt)^j \, K^{\text{Ft}}(t)| \leq C_2$ for $0 \leq j \leq \beta + \gamma + 1$ and for all $t$, and $K^{\text{Ft}}(0) = 1$ and $K^{\text{Ft}}(t)$ vanishes outside $[-1, 1]$; and (d) $g$ satisfies (4.4). (4.6)

The conditions imposed on $f_C$ and $f_B$ are a variation of the assumptions $|f_C^{\rm Ft}(t)| \geq$ const. $(1+|t|)^{-\alpha}$ and $|f_{-B}^{\rm Ft}(t)| \geq$ const. $(1+|t|)^{-\beta}$, respectively, which are typically encountered in errors-in-variables problems. They apply if, for example, the error distributions are of Laplace type, and in particular if $f_C = \phi(\,\cdot\,|\,\alpha)$ and $f_B = \phi(\,\cdot\,|\,\beta)$, where $\phi(\,\cdot\,|\,\omega)$ is the density of the distribution function with characteristic function $(1+t^2)^{-\omega}$ for all $t$. Then, (a) holds with $\alpha = 2\omega > 1$, and (b) holds with $\beta = 2\omega$ and $\gamma$ depending on $\omega$. More particularly, the case where $f_{-B}^{\rm Ft}$ is the inverse of a polynomial, which occurs if, for example, $f_{-B}^{\rm Ft}(t) = (1+t^2)^{-\omega}$ and $\omega$ is an integer, is of special interest. More generally, suppose that

$$f_{-B}^{\rm Ft}(t)^{-1} = 1 + \sum_{j=1}^{p} c_j\, t^j\,, \tag{4.7}$$

where $2 \leq p < \infty$ and the $c_j$'s are constants. Then, it is readily checked that (4.6)(b) holds. Smoother types of errors, for example those with Fourier transform bounded below by a negative exponential, can be considered as well. Results similar to those given in Section 4.3 hold for such errors too, but with considerably slower convergence rates. See the discussion at the end of Section 4.3.

The restriction imposed on $K$ reflects the fact that the compact support of $K^{\rm Ft}$ can be taken, without loss of generality, to be contained in $[-1,1]$, and asks as well that $K^{\rm Ft}$ be sufficiently smooth. In practice it is common to define $K$ by $K^{\rm Ft}(t) = (1-t^{r_1})^{r_2}$ for $t \in [-1,1]$, and $K^{\rm Ft} = 0$ otherwise, where $r_1$ is an even integer and $r_2$ is a positive integer. In such cases, (4.6)(c) holds provided

$$r_2 > \beta + \gamma + 1 \tag{4.8}$$

and the value of $\gamma$ is limited only by the size of $r_2$; $\gamma$ does not depend on selection of the constants $p$ and $c_1, \ldots, c_p$ in (4.7). In particular, by choosing $r_2$ sufficiently large we can take $\gamma$ arbitrarily large in (4.6). These considerations generally permit us to take $\xi = \infty$ in (4.5)(c); see the discussion immediately below Theorem 4.1. In such cases, (4.5) imposes especially mild conditions on $f_L$.

More generally, (4.5) and (4.6), and the statement of our main results in Section 4.3, are tailored to permit relatively weak conditions on $f_L$ and $g$. For example, no smoothness

assumptions are imposed at this point. Indeed, we shall raise the smoothness issue only through the bias terms $\text{bias}_b$, $\text{bias}_f$ and $\text{bias}_g$.

## 4.3 Properties in the Mixed Error Case

Here we assume the model (1.3), when neither of the errors $U_C$ and $U_B$ has a degenerate distribution. The estimator $\widehat{g}$ is given by (2.4), with $\widehat{f}_L$ and $\widehat{b}$ defined at (2.3). For simplicity we omit the case $\beta + \gamma = \xi$ in (4.9) below; it is the same as for $\beta + \gamma < \xi$, except that a factor $\log n$ is included. Upper bounds to convergence rates are given below. We do not have minimax lower bounds that reflect the upper bounds.

**Theorem 4.1.** *If* (4.5) *and* (4.6) *hold, and if* $h_1 = h_2$ *and* $0 < h_1, h_3 \le B_1$ *where* $B_1 > 0$*, then, for a constant* $B_2 > 0$ *not depending on* $n$*,* $h_1$*,* $h_2$ *or* $h_3$*,*

$$\sup_{x \in \mathcal{R}} E\Big|\widehat{g}(x) - g(x) - \mathit{bias}_g(x)\Big| \le B_2\, \lambda\, \big(\rho + \delta + \rho^{-1}\, \delta^2\big)\; h_3^{-\beta} \times \begin{cases} \big(1 + h_3^{\beta+\gamma}\, \rho^{(\beta+\gamma)/\xi - 1}\big) & \text{if } \beta + \gamma < \xi \\ \big(1 + h_3^{\beta+\gamma}\big) & \text{if } \beta + \gamma > \xi\,, \end{cases} \tag{4.9}$$

*where* $\lambda$ *is as at* (4.4).

To interpret this theorem, let us first consider the case where $g$ is a bounded, integrable function. Then the contribution of $g$ to bias is represented in (4.9) by

$$\text{bias}_g(x) = \int K(u)\, \{g(x - h_3 u) - g(x)\}\, du = O\big(h_3^k\big)\,, \tag{4.10}$$

where the first identity holds under the conditions of Theorem 4.1, and the second identity holds provided $g$ has $k$ bounded derivatives,

$$\int (1 + |u|)^k\, |K(u)|\, du < \infty\,, \tag{4.11}$$

and $\kappa_j = \int u^j\, K(u)\, du = 0$ for $j = 1, \ldots, k-1$. Kernels satisfying these conditions, as well as those in (4.6)(c), are commonly used in practice. For example, the kernels employed in Section 3 are of this type for $k = 2$.

Bias formulae such as (4.10) are of course conventional. It is the remaining contribution to convergence rate, bounded by the right-hand side of (4.9), that is most affected by the

errors-in-variables aspect of the problem and is therefore of greatest interest. Take the ridge parameter $\rho$ to equal a constant multiple of $\delta$ and assume that we can choose $\gamma$, in (4.6), so large that $h_3^\gamma = O(\rho)$ (this assumption is not an issue if $f_{-B}$ satisfies (4.7). Related interpretations of (4.9) are also possible where the simplifications obtainable when $f_{-B}$ is given by (4.7) do not apply, but those instances are not so transparent, since then both $\gamma$ and $\xi$ can impact on the overall convergence rate). Then (4.9) further simplifies to:

$$\sup_{x\in\mathcal{R}} E\Big|\widehat{g}(x) - g(x) - \mathrm{bias}_g(x)\Big| = O\big(\lambda\,\delta\,h_3^{-\beta}\big)\,, \tag{4.12}$$

where, since $g$ is bounded and integrable, $\lambda$ can be taken constant and so can be omitted from (4.12). Results (4.10) and (4.12) imply the following rate of convergence of $\widehat{g}$ to $g$:

$$\sup_{x\in\mathcal{R}} E|\widehat{g}(x) - g(x)| = O\big(h_3^k + \delta\,h_3^{-\beta}\big)\,. \tag{4.13}$$

If $f_L$ has $k$ bounded derivatives, then $\delta \sim$ const. $\times\, n^{-k/(2\alpha+2k+1)}$ provided we take $h_1 \sim$ const. $n^{-1/(2\alpha+2k+1)}$. This order of $\delta$ is also the minimax-optimal, root squared-error convergence rate for estimators of $g$, in the case where $U_B$ is identically zero. See, for example, Fan and Truong (1993). Thus, the factor $h_3^{-\beta}$, on the right-hand side of (4.13), can be interpreted as the amount by which the conventional convergence rate, $\delta$, is degraded by introducing the additional error $U_B$.

Of course, the factor $h_3^{-\beta}$ diverges as the bandwidth $h_3$ becomes smaller. On the other hand, the bias term $h_3^k$ reduces to zero as $h_3$ decreases, so there will be an optimal order of magnitude of $h_3$ for which the contributions $h_3^k$ and $\delta\,h_3^{-\beta}$ are in balance, leading to:

$$\sup_{x\in\mathcal{R}} E|\widehat{g}(x) - g(x)| = O\left(n^{-k^2/\{(2\alpha+2k+1)(\beta+k)\}}\right)\,. \tag{4.14}$$

Result (4.9) reveals the potential deleterious effects of taking the ridge, $\rho$, too large (that is, of larger order than $\delta$) or too small. In particular, the order of magnitude of the right-hand side of (4.9) is made larger by choosing $\rho$ to be of either strictly larger order, or strictly smaller order, than $\delta$.

It is straightforward to combine Theorem 4.1, and the results in Section 4.1, to handle the case of fixed but unbounded $g$. Specifically, taking $g_n = g \cdot 1_{[-n^D, n^D]}(x)$, with $D > 0$

arbitrarily small, and assuming that (4.1) holds, (4.2) permits us to take $\lambda = O(n^r)$ for any $r > 0$ in (4.9), provided we replace $\text{bias}_g(x)$ there by,

$$\text{bias}_{g_n}(x) = \int K(u) \{g_n(x - h_3 u) - g_n(x)\} \, du = O\big(n^s h_3^k\big) \tag{4.15}$$

for all $s > 0$. The second identity in (4.15) holds provided $g_n^{(k)}$ exists and $|g_n^{(k)}|$ grows no more than polynomially fast, $\kappa_j = 0$ for $j = 1, \ldots, k-1$, and, in a mild strengthening of (4.11), $\int |u|^{k+c} \, |K(u)| \, du < \infty$ for some $c > 0$. Therefore, and using also (4.12), the following version of (4.14) follows from (4.2) and (4.9): For all $r > 0$,

$$\sup_{x \in \mathcal{R}} |\widehat{g}(x) - g(x)| = O_p\left(n^{-(k^2-r)/\{(2\alpha+2k+1)(\beta+k)\}}\right) . \tag{4.16}$$

Thus it can be seen that unboundeness of $g$ barely changes the convergence rate.

Note too from (4.14) and (4.16) that our bounds on the rate of convergence increase as the smoothness of either error distribution increases; that is, as $\alpha$ or $\beta$ increases. These results correctly suggest that if either of the errors were "supersmooth," for example Gaussian, the convergence rate would be slower than the inverse of any polynomial in $n$. In fact, no estimator can converge at a polynomial rate in the supersmooth cases.

# 5 ORTHOGONAL SERIES METHOD

An alternative estimator of $g$ can be considered in the case where $g$, $f_B$ and $f_L$ are compactly supported. Here and below, we assume that $f_L$, $g$ and $f_B$ have been rescaled so that all three support intervals are contained within $\mathcal{I} = [-\pi, \pi]$. In this case, it follows from work of Delaigle, Hall and Qiu (2006) that no nonparametric estimator can identify $g$ outside the interval $[a_L + a_B, b_L - a_B]$, where $[-a_B, a_B]$ and $[a_L, b_L]$ denote the supports of, respectively, $f_B$ and $f_L$. Here, for simplicity, we have assumed that $f_B$ is symmetric. The estimator we describe below is able to identify $g$ on the interval $[a_L + a_B, b_L - a_B]$, whatever the support, compact or not, of the classical error density $f_C$. The trigonometric-series expansion of a function $k$ with support contained in $\mathcal{I}$ may be written as,

$$k(x) = k_0 + \sum_{j=1}^{\infty} \{k_{1j} \cos(jx) + k_{2j} \sin(jx)\} ,$$

with $k_0 = (2\pi)^{-1} \int_{\mathcal{I}} k$ and, for $\ell = 1, 2$, $k_{\ell,j} = \pi^{-1} \int_{\mathcal{I}} k(x) \, \mathrm{cs}_{\ell,j}(x) \, dx$, where $\mathrm{cs}_{\ell,j}(x) = \cos(jx)$ or $\sin(jx)$ according as $\ell = 1$ or 2, respectively. Using the sine-cosine decomposition of $g$ and $a$, we have, from Delaigle, Hall and Qiu (2006)

$$\begin{pmatrix} c_{1j} \\ c_{2j} \end{pmatrix} = \frac{1}{\delta_{1j}^2 + \delta_{2j}^2} \begin{pmatrix} \delta_{1j} & \delta_{2j} \\ -\delta_{2j} & \delta_{1j} \end{pmatrix} \begin{pmatrix} r_{1j} \\ r_{2j} \end{pmatrix}, \tag{5.1}$$

where, for $\ell = 1, 2$, $(c_{\ell,j}, \delta_{\ell,j}, r_{\ell,j})$ denotes $(g_{\ell,j}, \alpha_{\ell,j}, a_{\ell,j})$, with $\alpha_{\ell,j} = E\{\mathrm{cs}_{\ell,j}(U_B)\}$, and an estimator of the Fourier coefficients of $g$ can be deduced from (5.1), by replacing the Fourier coefficients $a_0$ and $a_{\ell,j}$ by $\widehat{a}_0 = (2\pi)^{-1} \int_{\mathcal{I}} \widehat{a}$ and $\widehat{a}_{\ell,j} = \pi^{-1} \int_{\mathcal{I}} \widehat{a} \, \mathrm{cs}_{\ell,j}$, where we choose $\widehat{a}$ to be a sine-cosine series estimator of $a$, defined by borrowing ideas of Hall and Qiu (2005) in the context of pure classical errors.

More precisely, define $b = a \, f_L$, $p_{\ell,j} = \pi^{-1} E\{\mathrm{cs}_{\ell,j}(W)\}$ and $q_{\ell,j} \equiv \pi^{-1} E\{Y \, \mathrm{cs}_{\ell,j}(W)\}$, for $\ell = 1, 2$. Then it can be shown that (5.1) holds with $(c_{\ell,j}, \delta_{\ell,j}, r_{\ell,j})$ equal to either $(f_{L\ell,j}, \beta_{\ell,j}, p_{\ell,j})$ or $(b_{\ell,j}, \beta_{\ell,j}, q_{\ell,j})$, where $\beta_{\ell,j} = E\{\mathrm{cs}_{\ell,j}(U_C)\}$, $\ell = 1, 2$. Substituting the estimators $\widehat{p}_{\ell,j} = (\pi n)^{-1} \sum_i \mathrm{cs}_{\ell,j}(W_i)$ and $\widehat{q}_{\ell,j} = (\pi n)^{-1} \sum_i Y_i \, \mathrm{cs}_{\ell,j}(W_i)$ for $p_{\ell,j}$ and $q_{\ell,j}$, we see that $a$ can be estimated by

$$\widehat{a}(x) \;=\; \frac{\widehat{b}_0 + \sum_{j \geq 1} \{\widehat{b}_{1j} \, \cos(jx) + \widehat{b}_{2j} \, \sin(jx)\}}{(2\pi)^{-1} + \sum_{j \geq 1} \{\widehat{f}_{L_{1j}} \, \cos(jx) + \widehat{f}_{L_{2j}} \, \sin(jx)\}} \, .$$

In practice, we need to truncate the series for $\widehat{a}$ and only keep the terms corresponding to $j \leq M_1$, where, for example, $M_1$ can be chosen by a thresholding rule as in Hall and Qiu (2005). The series for $\widehat{g}$ needs also be truncated, to keep only the terms $j \leq M_2$, where $M_2$ can be selected by a cross-validation procedure of the type introduced in Delaigle, Hall and Qiu (2006).

# 6 OUTLINE PROOF OF THEOREM 4.1

Define $\bar{f}_L = \max(\widehat{f}_L, 0)$, $\Delta_b = \widehat{b} - b$, $\Delta_f = \widehat{f}_L - f_L$, $\bar{\Delta}_f = \bar{f}_L - f_L$, $k_x(u) = h_3^{-1} \, K_{-B}\{(u - x)/h_3\}$ and $Q_1 = 2 \, (|a| \, \Delta_f^2 + |\Delta_b \, \Delta_f|)/\{\rho \, (f_L + \rho)\}$. It can be shown that,

$$\widehat{g}(x) - g(x) = \mathrm{bias}_g(x) - \rho \int \frac{a \, k_x}{f_L + \rho} + A_b(x) - A_f(x) + Q_2(x) \, , \tag{6.1}$$

where

$$A_b(x) = \int \frac{\Delta_b\, k_x}{f_L + \rho}\,, \quad A_f(x) = \int \frac{b\,\Delta_f\, k_x}{(f_L+\rho)^2}\,, \quad |Q_2(x)| \le \int Q_1\,|k_x|\,. \tag{6.2}$$

Using (b) and (c) of (4.6) it can be shown that $|K_{-B}(x\,|\,h)| \le C_1\, h^{-\beta}\,(1+|x|)^{-(\beta+\gamma+1)}$ for all real $x$ and all $h > 0$, where $C_1, C_2, \ldots$ will denote positive constants. This leads to the result, $h_3^\beta \int |k_x|\,(f_L+\rho)^{-1} \le C_2\,(I_1 + I_2)$, where, with $C > 0$ chosen so small that $x + h_3 u \in \mathcal{S}$ whenever $x \in \mathcal{R}$ and $|h_3 u| \le C$, and $\mathcal{R}$ and $\mathcal{S}$ as in (4.5), we define $I_1 = \int_{|h_3 u| \le C} (1+|u|)^{-(\beta+\gamma+1)}\,du \le C_3$,

$$I_2 = \int_{|h_3 u| > C} \big\{(h_3|u|)^{-\xi} + \rho\big\}^{-1}\,(1+|u|)^{-(\beta+\gamma+1)}\,du \le C_4 \begin{cases} h_3^{\beta+\gamma}\,\rho^{(\beta+\gamma)/\xi - 1} & \text{if } \beta+\gamma < \xi \\ h_3^{\beta+\gamma} & \text{if } \beta+\gamma > \xi\,. \end{cases}$$

Combining the results in this paragraph we deduce that

$$h_3^\beta \int \frac{|k_x|}{f_L+\rho} \le C_5 \begin{cases} \big(1 + h_3^{\beta+\gamma}\,\rho^{(\beta+\gamma)/\xi-1}\big) & \text{if } \beta+\gamma < \xi \\ \big(1 + h_3^{\beta+\gamma}\big) & \text{if } \beta+\gamma > \xi\,, \end{cases} \tag{6.3}$$

uniformly in $x \in \mathcal{R}$.

Using (4.6)(a) and the definition of supbi($h_1$) it can be shown that, for $c = b$ and $c = f$, we have, uniformly in $x$,

$$E\big\{\Delta_b(u)^2\big\} + |a|^2 E\big\{\Delta_f(u)^2\big\} \le C_6\,\lambda^2\,\delta^2\,. \tag{6.4}$$

Using (6.3), (6.4) and the result $|b| \le C_7\,\lambda\, f_L$, it can be proved that, for $c = b$ and $c = f$,

$$\big\{E A_c(x)^2\big\}^{1/2} \le C_8\,\lambda\,\delta\,h_3^{-\beta} \begin{cases} \big(1 + h_3^{\beta+\gamma}\,\rho^{(\beta+\gamma)/\xi-1}\big) & \text{if } \beta+\gamma < \xi \\ \big(1 + h_3^{\beta+\gamma}\big) & \text{if } \beta+\gamma > \xi\,, \end{cases} \tag{6.5}$$

$$\rho\left|\int \frac{a\,k_x}{f_L+\rho}\right| \le C_9\,\lambda\,\rho\,h_3^{-\beta} \begin{cases} \big(1 + h_3^{\beta+\gamma}\,\rho^{(\beta+\gamma)/\xi-1}\big) & \text{if } \beta+\gamma < \xi \\ \big(1 + h_3^{\beta+\gamma}\big) & \text{if } \beta+\gamma > \xi\,, \end{cases} \tag{6.6}$$

where both formulae hold uniformly in $x \in \mathcal{R}$. Together, (6.1), (6.5) and (6.6) give:

$$\widehat{g}(x) - g(x) = \mathrm{bias}_g(x) + Q_2(x) + Q_3(x)\,, \tag{6.7}$$

where $Q_2$ is as before, and so satisfies the last inequality at (6.2), and

$$\left\{EQ_3(x)^2\right\}^{1/2} \le C_{10}\,\lambda\,(\rho+\delta)\,h_3^{-\beta}\begin{cases}\left(1+h_3^{\beta+\gamma}\,\rho^{(\beta+\gamma)/\xi-1}\right) & \text{if } \beta+\gamma<\xi\\ \left(1+h_3^{\beta+\gamma}\right) & \text{if } \beta+\gamma>\xi\,,\end{cases} \tag{6.8}$$

uniformly in $x \in \mathcal{R}$. Properties (6.3) and (6.4), and the last inequality at (6.2), entail:

$$E|Q_2(x)| \le C_{11}\,\lambda\,\delta^2\,\rho^{-1}\,h_3^{-\beta}\begin{cases}\left(1+h_3^{\beta+\gamma}\,\rho^{(\beta+\gamma)/\xi-1}\right) & \text{if } \beta+\gamma<\xi\\ \left(1+h_3^{\beta+\gamma}\right) & \text{if } \beta+\gamma>\xi\,.\end{cases} \tag{6.9}$$

Results (6.7)–(6.9) imply that $\widehat{g}(x)-g(x) = \text{bias}_g(x)+Q_4(x)$, where, uniformly in $x \in \mathcal{R}$,

$$E|Q_4(x)| \le C_{12}\,\lambda\left\{\rho+\delta+\delta^2\,\rho^{-1}\right\}h_3^{-\beta}\begin{cases}\left(1+h_3^{\beta+\gamma}\,\rho^{(\beta+\gamma)/\xi-1}\right) & \text{if } \beta+\gamma<\xi\\ \left(1+h_3^{\beta+\gamma}\right) & \text{if } \beta+\gamma>\xi\,.\end{cases}$$

Theorem 4.1 follows directly from these properties.

# Acknowledgments

Carroll's research was supported by a grant from the National Cancer Institute, and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences. Delaigle's research was supported by a Hellman Fellowship and a Maurice Belz Fellowship. We thank Dr. J. Lynn Lyon for giving us access to the Nevada Test Site data and Dr. F. Owen Hoffman for many helpful discussions on the mixture of Berkson and classical uncertainties.

# REFERENCES

Berkson, J. (1950). Are there two regression problems? *J. Amer. Statist. Assoc.* **45**, 164–180.

Buonaccorsi, J. P. and Lin, C.-D. (2002). Berkson measurement error in designed repeated measures studies with random coefficients. *J. Statist. Plann. Inf.* **104**, 53-72.

Carroll, R. J., Maca, J. D. and Ruppert, D. (1999). Nonparametric regression in the presence of measurement error. *Biometrika* **86**, 541–554.

Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models*, second edition. Chapman and Hall CRC Press, Boca Raton.

Cook, J. R. and Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *J. Amer. Statist. Assoc.* **89**, 1314–1328.

Delaigle, A. (2007). Nonparametric density estimation from data contaminated by Berkson errors, classical errors, or a mixture of both. *Can. J. Statist.*, **35**, 1–16.

Delaigle, A., Hall, P. and Qiu, P. (2006). Nonparametric methods for solving the Berkson errors-in-variables problem. *J. Roy. Statist. Soc., B* **68**, 201–220.

Devanarayan, V. and Stefanski, L. A. (2002). Empirical simulation extrapolation for measurement error models with replicate measurements. *Statist. Probab. Lett.* **59**, 219-225.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*, Chapman and Hall: London.

Fan, J. and Masry, E. (1992). Multivariate regression estimation with errors-in-variables: asymptotic normality for mixing processes. *J. Multivariate Anal.* **43**, 237–271.

Fan, J. and Truong, Y. K. (1993). Nonparametric regression with errors in variables. *Ann. Statist.* **21**, 1900–1925.

Hall, P. and Qiu, P. (2005). Discrete-transform approach to deconvolution problems. *Biometrika* **92**, 135–148.

Huwang, L. and Huang, H. Y. S. (2000). On errors-in-variables in polynomial regression – Berkson case. *Statist. Sinica* **10**, 923-936.

Kerber, R. L., Till, J. E., Simon, S. L., Lyon, J. L. Thomas, D. C., Preston-Martin, S., Rollison, M. L., Lloyd, R. D. and Stevens, W. (1993). A cohort study of thyroid disease in relation to fallout from nuclear weapons testing. *J. Amer. Medical Assoc.* **270**, 2076-2083.

Kim, J. and Gleser, L. J. (2000). SIMEX approaches to measurement error in ROC studies. *Comm. Statist. Theory Meth.* **29**, 2473–2491.

Li, Y., Guolo, A., Hoffman, F. O. and Carroll, R. J. (2007). Shared uncertainty in measurement error problems, with application to Nevada Test Site Fallout data. *Biometrics*, to appear.

Linton, O. and Whang, Y. J. (2002). Nonparametric estimation with aggregated data. *Econometric Theory* **18**, 420–468.

Lyon, J. L., Alder, S. C., Stone, M. B., Scholl, A., Reading, J. C. Holubkov, R., Sheng, X., White, G. L., Hegmann, K. T., Anspaugh, L., Hoffman, F. O., Simon, S. L., Thomas, B., Carroll, R. J. and Meikle, A. W. (2006). Thyroid disease associated with exposure to the Nevada Test Site radiation: a reevaluation based on corrected dosimetry and examination data. *Epidemiology* **17**, 604-614.

Mallick, B., Hoffman, F. O. and Carroll, R. J. (2002). Semiparametric regression modeling

with mixtures of Berkson and classical error, with application to fallout from the Nevada test site. *Biometrics* **58**, 13-20.

Reeves, G. K., Cox, D. R., Darby, S. C. and Whitley, E. (1998). Some aspects of measurement error in explanatory variables for continuous and binary regression models. *Stat. Medicine* **17**, 2157-2177.

Schafer, D. W. and Gilbert, E. S. (2006). Some statistical implications of dose uncertainty in radiation dose-response analyses. *Radiation Research* **166**, 303-312.

Simon, S. L., Till, J. E., Lloyd, R. D., Kerber, R. L., Thomas, D. C., Preston-martin, S., Lyon, J. L. and Stevens, W. (1995). The Utah Leukemia case-control study: dosimetry methodology and results. *Health Physics* **68**, 460-471.

Simon, S. L., Anspaugh, L. R., Hoffman, F. O., et al. (2006). Update of dosimetry for the Utah Thyroid Cohort Study. *Radiation Research* **165**, 208-22.

Stefanski, L. A. and Cook, J. R. (1995). Simulation-extrapolation: The measurement error jackknife. *J. Amer. Statist. Assoc.* **90**, 1247–1256.

Stevens, W., Till, J. E., Thomas, D. C., et al. (1992). Assessment of leukemia and thyroid disease in relation to fallout in Utah: report of a cohort study of thyroid disease and radioactive fallout from the Nevada test site. University of Utah.

Stram, D. O., Huberman, M. and Wu, A. H. (2002). Is residual confounding a reasonable explanation for the apparent protective effects of beta-carotene found in epidemiological studies of lung cancer in smokers? *Amer. J. Epidemiol.* **155**, 622–628.

Taupin, M. L. (2001). Semi-parametric estimation in the nonlinear structural errors-in-variables model. *Ann. Statist.* **29**, 66–93.

Wang, L. (2003). Estimation of nonlinear Berkson-type measurement error models. *Statist. Sinica* **13**, 1201–1210.

# NOT-FOR-PUBLICATION APPENDIX: DETAILS FOR SECTION 6

Put $\bar{f}_L = \max(\widehat{f}_L, 0)$, $\Delta_b = \widehat{b} - b$, $\Delta_f = \widehat{f}_L - f_L$ and $\bar{\Delta}_f = \bar{f}_L - f_L$, and observe that

$$\frac{1}{\bar{f}_L + \rho} = \frac{1}{f_L + \rho} - \frac{\bar{\Delta}_f}{(f_L + \rho)(\bar{f}_L + \rho)}\,.$$

Therefore,

$$\frac{b + \Delta_b}{\bar{f}_L + \rho} = \frac{b + \Delta_b}{f_L + \rho} - \frac{b\,\bar{\Delta}_f}{(f_L + \rho)^2} + \frac{b\,\bar{\Delta}_f^2}{(f_L + \rho)^2(\bar{f}_L + \rho)} - \frac{\Delta_b\,\bar{\Delta}_f}{(f_L + \rho)(\bar{f}_L + \rho)}\,.$$

Also,

$$\frac{b}{f_L} - \frac{b}{f_L + \rho} = \frac{\rho\, b}{f_L\,(f_L + \rho)} = \frac{\rho\, a}{f_L + \rho}\,,$$

and so

$$\frac{b + \Delta_b}{\bar{f}_L + \rho} = \frac{b}{f_L} - \frac{\rho\, a}{f_L + \rho} + \frac{\Delta_b}{f_L + \rho} - \frac{b\,\bar{\Delta}_f}{(f_L + \rho)^2} + Q_1\,,$$

where $Q_1$, denoting quadratic terms, satisfies

$$|Q_1| \le \frac{|a|\,\bar{\Delta}_f^2 + |\Delta_b\,\bar{\Delta}_f|}{(f_L + \rho)(\bar{f}_L + \rho)} \le \frac{|a|\,\Delta_f^2 + |\Delta_b\,\Delta_f|}{\rho\,(f_L + \rho)}\,. \tag{A.1}$$

Here we have used the fact that $|\bar{\Delta}_f| \le |\Delta_f|$.

Now, $|\bar{\Delta}_f - \Delta_f| = |\widehat{f}_L|\, I(\widehat{f}_L < 0) \le |\Delta_f|\, I(\widehat{f}_L < 0)$, from which result and (A.1) it follows that

$$\frac{b + \Delta_b}{\bar{f}_L + \rho} = \frac{b}{f_L} - \frac{\rho\, a}{f_L + \rho} + \frac{\Delta_b}{f_L + \rho} - \frac{b\,\Delta_f}{(f_L + \rho)^2} + Q_2\,, \tag{A.2}$$

where $Q_2$ satisfies:

$$|Q_2| \le \frac{|a|\,\Delta_f^2 + |\Delta_b\,\Delta_f|}{\rho\,(f_L + \rho)} + \frac{|b\,\Delta_f|\, I(\widehat{f}_L < 0)}{(f_L + \rho)^2}\,. \tag{A.3}$$

However, if $\widehat{f}_L < 0$ then $|\Delta_f| > f_L$, whence it follows that

$$\frac{|b\,\Delta_f|\, I(\widehat{f}_L < 0)}{(f_L + \rho)^2} \le \frac{|b|\,\Delta_f^2}{f_L\,(f_L + \rho)^2} = \frac{|a|\,\Delta_f^2}{(f_L + \rho)^2}\,.$$

Hence, (A.3) implies that

$$|Q_2| \le Q_3 \equiv 2\,\frac{|a|\,\Delta_f^2 + |\Delta_b\,\Delta_f|}{\rho\,(f_L + \rho)}\,. \tag{A.4}$$

Define $k_x(u) = h_3^{-1}\, K_{-B}\{(u-x)/h_3\}$. Then, (A.2) and (A.4) imply that

$$\widehat{g}(x) = \int a\, k_x - \rho \int \frac{a\, k_x}{f_L + \rho} + A_b(x) - A_f(x) + Q_4(x)\,, \tag{A.5}$$

where

$$A_b(x) = \int \frac{\Delta_b\, k_x}{f_L+\rho}\,, \quad A_f(x) = \int \frac{b\,\Delta_f\, k_x}{(f_L+\rho)^2}\,, \quad |Q_4(x)| \le \int Q_3\, |k_x|\,. \tag{A.6}$$

Now,

$$\int a\, k_x = \frac{1}{2\pi} \int g^{\mathrm{Ft}}(t)\, K^{\mathrm{Ft}}(h_3 t)\, e^{-itx}\, dt = g(x) + \mathrm{bias}_g(x)\,,$$

and so (A.5) entails:

$$\widehat{g}(x) - g(x) = \mathrm{bias}_g(x) - \rho \int \frac{a\, k_x}{f_L+\rho} + A_b(x) - A_f(x) + Q_4(x)\,. \tag{A.7}$$

Observe that,

$$\begin{aligned} h^\beta\, K_{-B}(x\,|\,h) &= \frac{h^\beta}{2\pi} \int K^{\mathrm{Ft}}(t)\, f_B^{\mathrm{Ft}}(t/h)^{-1}\, e^{-itx}\, dt \\ &= \frac{1}{2\pi} \int K^{\mathrm{Ft}}(t)\, \{s^{-\beta}\, f_B^{\mathrm{Ft}}(st)^{-1}\} \left(\frac{d}{dt}\, \frac{-e^{-itx}}{ix}\right) dt\,, \end{aligned} \tag{A.8}$$

where $s = h^{-1}$ and should be interpreted as the $s$ in (4.6)(b). Using parts (b) and (c) of (4.6); employing the first line of (A.8), and taking the absolute value of the integrand, to derive an upper bound for $|K_{-B}(x\,|\,h)|$ that does not depend on $x$; deriving an upper bound that depends on $x$, by integrating by parts $\beta+\gamma+1$ times in the manner suggested by the second line of (A.8); and combining these two bounds; we deduce that

$$|K_{-B}(x\,|\,h)| \le C_1\, h^{-\beta}\, (1+|x|)^{-(\beta+\gamma+1)} \tag{A.9}$$

for all real $x$ and all $h > 0$, where $C_1, C_2, \ldots$ will denote positive constants. Therefore,

$$\begin{aligned} I \equiv h_3^\beta \int \frac{|k_x|}{f_L+\rho} &\le C_1 \int \{f_L(x+h_3u)+\rho\}^{-1}\, (1+|u|)^{-(\beta+\gamma+1)}\, du \\ &\le C_2\, (I_1 + I_2)\,, \end{aligned} \tag{A.10}$$

where, with $C > 0$ fixed and chosen so small that $x + h_3 u \in \mathcal{S}$ whenever $x \in \mathcal{R}$ and $|h_3 u| \le C$, and $\mathcal{R}$ and $\mathcal{S}$ as in (3.2), we define

$$I_1 = \int_{|h_3 u| \le C} (1+|u|)^{-(\beta+\gamma+1)}\, du \le C_3\,,$$

$$
\begin{aligned}
I_2 &= \int_{|h_3 u|>C} \left\{ (h_3|u|)^{-\xi} + \rho \right\}^{-1} (1+|u|)^{-(\beta+\gamma+1)}\, du \\
&\le C_4 \left\{ \int_{Ch_3^{-1}}^{h_3^{-1}\rho^{-1/\xi}} \frac{(h_3 u)^{\xi}}{(1+u)^{\beta+\gamma+1}}\, du + \rho^{-1} \int_{h_3^{-1}\rho^{-1/\xi}}^{\infty} (1+|u|)^{-(\beta+\gamma+1)}\, du \right\} \\
&\le C_5 \begin{cases} h_3^{\beta+\gamma}\, \rho^{(\beta+\gamma)/\xi - 1} & \text{if } \beta+\gamma < \xi \\ h_3^{\beta+\gamma} & \text{if } \beta+\gamma > \xi \end{cases} \quad + \quad C_5\, \rho^{-1} \left( h_3\, \rho^{1/\xi} \right)^{\beta+\gamma} \\
&\le C_6 \begin{cases} h_3^{\beta+\gamma}\, \rho^{(\beta+\gamma)/\xi - 1} & \text{if } \beta+\gamma < \xi \\ h_3^{\beta+\gamma} & \text{if } \beta+\gamma > \xi\,. \end{cases}
\end{aligned}
$$

Here we have used the fact that $f_L(u) \ge |u|^{-\xi}$ for all sufficiently large $|u|$. Combining the results from (A.10) down we deduce that

$$
h_3^{\beta} \int \frac{|k_x|}{f_L + \rho} \le C_7 \begin{cases} \left(1 + h_3^{\beta+\gamma}\, \rho^{(\beta+\gamma)/\xi - 1}\right) & \text{if } \beta+\gamma < \xi \\ \left(1 + h_3^{\beta+\gamma}\right) & \text{if } \beta+\gamma > \xi\,, \end{cases} \tag{A.11}
$$

uniformly in $x \in \mathcal{R}$.

Analogously to (A.9) it can be shown using (4.6)(a) that for all $x$ and all $h > 0$,

$$
|K_C(x \,|\, h)| \le C_8\, h_1^{-\alpha}\, (1+|x|)^{-1}\,. \tag{A.12}
$$

It follows from the definition of supbi($h_1$) that $\max(|E\Delta_b|, |E\Delta_f|) \le \text{supbi}(h_1)$. Using (A.12) to obtain the second inequality below, it can be proved that, uniformly in $u$,

$$
\text{var}\{\Delta_f(u)\} \le \frac{1}{n h_1^2} \int K_C\Big(\frac{u-w}{h_1}\Big)^2 f_W(w)\, dw \le C_9 \left(n h_1^{2\alpha+1}\right)^{-1}. \tag{A.13}
$$

An identical bound applies to var$\{\Delta_b(u)\}$, although with a factor $\lambda^2$; recall the assumption that var$(\eta) < \infty$, imposed as part of model (1.3). Combining the results so far in this paragraph we deduce that, for $c = b$ and $c = f$, we have, uniformly in $u$,

$$
E\big\{\Delta_b(u)^2\big\} + |a|^2 E\big\{\Delta_f(u)^2\big\} \le C_{10}\, \lambda^2\, \delta^2\,. \tag{A.14}
$$

The function $f_B$, being a density, is integrable, and hence, in view of (2.1), the function $a$ is bounded above by a constant multiple of $\lambda$. Hence, by (2.2), $|b| \le C_{11}\, \lambda\, f_L$ for a constant $C_{11} > 0$. This property, (A.11) and (A.14) together imply that, for $c = b$ and $c = f$,

$$
\big\{EA_c(x)^2\big\}^{1/2} \le C_{12}\, \lambda\, \delta\, h_3^{-\beta} \begin{cases} \left(1 + h_3^{\beta+\gamma}\, \rho^{(\beta+\gamma)/\xi - 1}\right) & \text{if } \beta+\gamma < \xi \\ \left(1 + h_3^{\beta+\gamma}\right) & \text{if } \beta+\gamma > \xi\,. \end{cases} \tag{A.15}
$$

Formula (A.11) also implies that

$$\rho \left| \int \frac{a\, k_x}{f_L + \rho} \right| \leq C_{13}\, \lambda\, \rho\, h_3^{-\beta} \begin{cases} \left(1 + h_3^{\beta+\gamma}\, \rho^{(\beta+\gamma)/\xi - 1}\right) & \text{if } \beta + \gamma < \xi \\ \left(1 + h_3^{\beta+\gamma}\right) & \text{if } \beta + \gamma > \xi\,, \end{cases} \tag{A.16}$$

where both (A.15) and (A.16) hold uniformly in $x \in \mathcal{R}$. Together, (A.7), (A.15) and (A.16) give:

$$\widehat{g}(x) - g(x) = \text{bias}_g(x) + Q_4(x) + Q_5(x)\,, \tag{A.17}$$

where $Q_4$ is as before, and so satisfies the last inequality at (A.6), and

$$\left\{ E Q_5(x)^2 \right\}^{1/2} \leq C_{14}\, \lambda\, (\rho + \delta)\, h_3^{-\beta} \begin{cases} \left(1 + h_3^{\beta+\gamma}\, \rho^{(\beta+\gamma)/\xi - 1}\right) & \text{if } \beta + \gamma < \xi \\ \left(1 + h_3^{\beta+\gamma}\right) & \text{if } \beta + \gamma > \xi\,, \end{cases} \tag{A.18}$$

uniformly in $x \in \mathcal{R}$.

Properties (A.4), (A.11) and (A.14), and the last inequality at (A.6), entail:

$$\begin{aligned} E|Q_4(x)| &\leq \int E(Q_3)\, |k_x| \\ &\leq C_{15}\, \lambda\, \delta^2\, \rho^{-1}\, h_3^{-\beta} \begin{cases} \left(1 + h_3^{\beta+\gamma}\, \rho^{(\beta+\gamma)/\xi - 1}\right) & \text{if } \beta + \gamma < \xi \\ \left(1 + h_3^{\beta+\gamma}\right) & \text{if } \beta + \gamma > \xi\,. \end{cases} \end{aligned} \tag{A.19}$$

Combining (A.17)–(A.19) we deduce that

$$\widehat{g}(x) - g(x) = \text{bias}_g(x) + Q_6(x)\,, \tag{A.20}$$

where, uniformly in $x \in \mathcal{R}$,

$$E|Q_6(x)| \leq C_{16}\, \lambda \left\{ \rho + \delta + \delta^2\, \rho^{-1} \right\} h_3^{-\beta} \begin{cases} \left(1 + h_3^{\beta+\gamma}\, \rho^{(\beta+\gamma)/\xi - 1}\right) & \text{if } \beta + \gamma < \xi \\ \left(1 + h_3^{\beta+\gamma}\right) & \text{if } \beta + \gamma > \xi\,. \end{cases} \tag{A.21}$$

Theorem 4.1 follows directly from (A.20) and (A.21).