

Non-parametric regression estimation from data contaminated by a mixture of Berkson and classical errors

Raymond J. Carroll,

Texas A&M University, College Station, USA

Aurore Delaigle

University of Bristol, UK, and University of Melbourne, Australia

and Peter Hall

University of Melbourne, Australia, and University of California at Davis, USA

[Received July 2006. Revised May 2007]

Summary. 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, which are 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 that is related to the observation of the latter causes an error of classical type. We propose a non-parametric estimator of a regression function from data that are 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

1. Introduction

We consider non-parametric 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) , $1 \leq j \leq n$, which are generated by a so-called Berkson model

$$\begin{aligned} Y_j &= g(X_j) + \eta_j, \\ X_j &= L_j + U_{B,j}, \end{aligned} \tag{1.1}$$

Address for correspondence: Aurore Delaigle, Department of Mathematics, University of Bristol, University Walk, Bristol, BS8 1TW, UK.
E-mail: Aurore.Delaigle@bristol.ac.uk

where $U_{B,j}$, L_j and η_j are mutually independent, $E(\eta_j|X_j) = 0$ and $\text{var}(\eta_j) < \infty$. 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. Model (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 *et al.* (2006) for a non-parametric treatment.

In most situations, the surrogate L cannot be observed without measurement error, which is 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 , $1 \leq j \leq n$, that are 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. Non-parametric methods for inference in settings such as this include kernel approaches (e.g. Fan and Masry (1992), Taupin (2001) and Linton and Whang (2002)) and techniques that are based on simulation and extrapolation arguments (e.g. Cook and Stefanski (1994), Stefanski and Cook (1995), Carroll *et al.* (1999, 2006), Kim and Gleser (2000) and 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 $1 \leq j \leq n$, which are generated by

$$\left. \begin{aligned} Y_j &= g(X_j) + \eta_j, \\ X_j &= L_j + U_{B,j}, \\ W_j &= L_j + U_{C,j}, \end{aligned} \right\} \tag{1.3}$$

where $U_{C,j} \sim f_C$, $U_{B,j} \sim f_B$, $L_j \sim f_L$ and η_j are mutually independent, $E(\eta_j|X_j) = 0$, $\text{var}(\eta) < \infty$, 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 semi-parametric, Bayesian, setting of radiation exposure from nuclear testing; see also Li *et al.* (2007). In this paper we consider non-parametric estimation of the regression function g , for data generated by the model (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.

Non-parametric 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 that are 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 with density es-

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

Section 5 suggests an alternative non-parametric orthogonal series estimator, which is designed for cases where the function g and the densities f_L and f_B are compactly supported. Technical details are collected in Appendix A.

2. Kernel method

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

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

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 $1 \leq j \leq n$, to estimate consistently 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^{Ft} for the corresponding characteristic function. Let K be a kernel function, which is chosen so that its Fourier transform K^{Ft} satisfies $K^{Ft}(0) = 1$ and vanishes outside a compact interval (such kernels are fairly standard in deconvolution problems; see for example Fan and Truong (1993)). Given $h > 0$, put

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

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

$$\begin{aligned} \hat{f}_L(x) &= \frac{1}{nh_1} \sum_{j=1}^n K_C\left(\frac{x - W_j}{h_1}\right), \\ \hat{b}(x) &= \frac{1}{nh_2} \sum_{j=1}^n Y_j K_C\left(\frac{x - W_j}{h_2}\right), \end{aligned} \tag{2.3}$$

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

$$\hat{g}(x) = \frac{1}{h_3} \int \hat{a}(u) K_{-B}\left(\frac{x-u}{h_3}\right) du \tag{2.4}$$

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 \hat{g} is simply $\hat{a} = \hat{b} / \hat{f}_L$. 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, which is introduced to avoid problems with the denominator of \hat{a} at points x where $\hat{f}_L(x)$ 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, \hat{f}_L and \hat{b} are standard kernel estimators and, in particular,

$$\begin{aligned} \hat{f}_L(x) &= \frac{1}{nh_1} \sum_{j=1}^n \mathcal{K}\left(\frac{x - W_j}{h_1}\right), \\ \hat{b}(x) &= \frac{1}{nh_2} \sum_{j=1}^n Y_j \mathcal{K}\left(\frac{x - W_j}{h_2}\right), \end{aligned} \tag{2.5}$$

where \mathcal{K} can be taken to be a conventional kernel. Using these alternative definitions of \hat{f}_L and \hat{b} we may continue to define \hat{g} by equation (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 ρ . 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 \hat{g} is a computationally intensive operation. We split the problem into two parts, selecting (h_1, ρ) and h_3 separately, as follows.

Define

$$\begin{aligned} S_{k1}(l) &= K_C\left(\frac{l - W_k}{h_1}\right) / \left\{ \sum_{k'=1}^n K_C\left(\frac{l - W_{k'}}{h_1}\right) + \rho \right\}, \\ S_{k2}(x) &= h_3^{-1} \int S_{k1}(l) K_{-B}\left(\frac{x - l}{h_3}\right) dl. \end{aligned}$$

Ideally we would use a cross-validation approach, selecting (h_1, ρ) as

$$(\hat{h}_1, \hat{\rho}) = \arg \min_{(h_1, \rho)} \left[\sum_{j=1}^n \left\{ \frac{Y_j - \hat{a}(L_j)}{1 - S_{j1}(L_j)} \right\}^2 \right], \tag{3.1}$$

and then estimating h_3 by

$$\hat{h}_3 = \operatorname{argmin}_{h_3} \left[\sum_{j=1}^n \left\{ \frac{Y_j - \hat{g}(X_j)}{1 - n^{-1} \sum_{k=1}^n S_{k2}(X_k)} \right\}^2 \right]. \tag{3.2}$$

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

The first approach, which is motivated by the case where the error variances are small, is simply to ignore all error that is present in the data, i.e. to replace all L_j s and X_j s by W_j s, and to replace f_B^{Ft} and f_C^{Ft} (in the definitions of K_C and K_{-B}) by 1. The second possibility is to

replace $\exp(-itL_j/h_1)$ and $\exp(-itX_j/h_3)$ respectively in $K_C\{(L_j - W_k)/h_1\}$ and $K_{-B}\{(X_j - l)/h_3\}$ by $\exp(-itW_j/h_1) K^{Ft}(t)/f_C^{Ft}(t/h_1)$ and $\exp(-itW_j/h_3) f_B^{Ft}(-t/h_3)/f_C^{Ft}(-t/h_3)$, 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 ν 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 $S_{kr}(V_j)$ that is obtained by replacing $K_Z\{(V_j - \cdot)/h\}$ by $\hat{K}_{Z,\nu}\{(W_j - \cdot)/h\}$, and $\hat{f}(V_j)$ by

$$\hat{f}_\nu(W_j) = \sum_{k=1}^n Y_k S_{kr;\nu}(W_j),$$

where $\hat{K}_{Z,1}\{(W_j - \cdot)/h\} = K\{(W_j - \cdot)/h\}$, $\hat{K}_{-B,2}\{(W_j - l)/h_3\} = K_C\{(W_j - l)/h_3\}$ and

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

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

(a) Choose

$$(\hat{h}_1, \hat{\rho}) = \arg \min_{(h_1, \rho)} \left[\sum_{j=1}^n \left\{ w_1 \frac{Y_j - \hat{a}_1(W_j)}{1 - S_{j1;1}(W_j)} + w_2 \frac{Y_j - \hat{a}_2(W_j)}{1 - S_{j1;2}(W_j)} \right\}^2 \right];$$

then,

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

$$\bar{h}_3 = \arg \min_{h_3} \left[\sum_{j=1}^n \left\{ w_1 \frac{Y_j - \hat{g}_1(W_j)}{1 - n^{-1} \sum_{k=1}^n S_{k2;1}(W_k)} + w_2 \frac{Y_j - \hat{g}_2(W_j)}{1 - n^{-1} \sum_{k=1}^n S_{k2;2}(W_k)} \right\}^2 \right];$$

and, finally,

(c) select $(\hat{h}_1, \hat{\rho}, \hat{h}_3) = (\hat{h}_1, \hat{\rho}, (\sigma_B/\sigma_C)^{2/3} \bar{h}_3)$.

The weight functions w_1 and w_2 were chosen empirically and are such that, when the error variance tends to 0, we select the smoothing parameters via the first procedure only, which, for errors tending to 0, is the same as the cross-validation procedure that would be used in the error-free case. The correction that is 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 $(W_1, Y_1), \dots, (W_n, Y_n)$ according to model (1.3), where the regression function g was one of the following curves:

- (a) $g(x) = (50x^2 + 10x + 25)^{-1}$ (sharp unimodal),
- (b) $g(x) = \phi_{0,1.5}(4x) + \phi_{1,2}(4x) + \phi_{2,5}(4x)$ (asymmetric) and
- (c) $g(x) = 5 \sin(2x) \exp(-16x^2/50)$ (sinusoidal),

where $\phi_{\mu,\sigma}$ is the density of an $N(\mu, \sigma^2)$ variable.

For each example, the variable L was either a centred normal or a multiple of T_8 , a Student- t -distributed variable with eight degrees of freedom, and the error variables U_B and U_C were

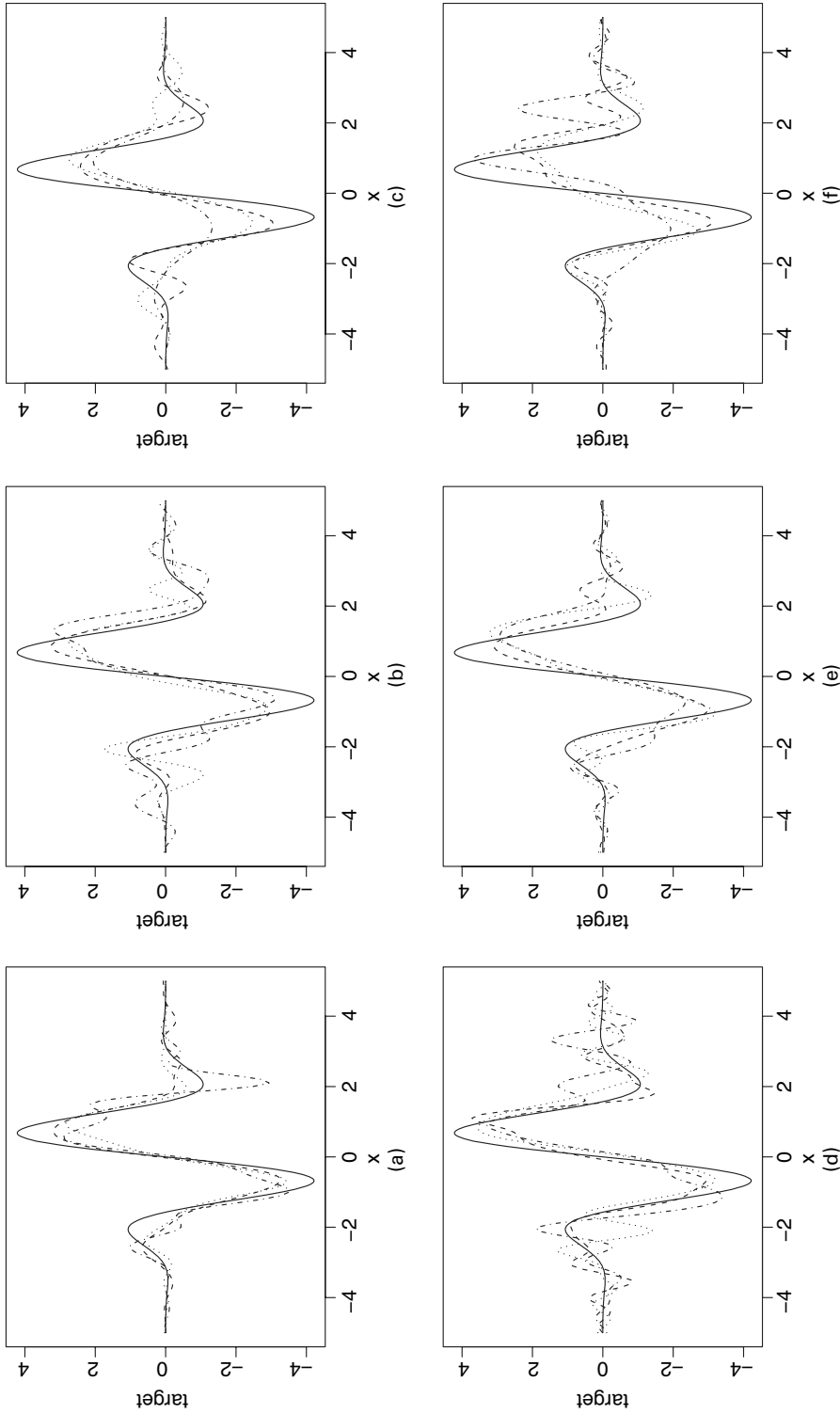


Fig. 1. Estimation of function (c) for samples of size $n = 250$, when $L \sim T_8 \sqrt{0.75}$, U_B and U_C are (a)–(c) Laplace or (d)–(f) normal, when $(\sigma_B^2/\sigma_L^2, \sigma_C^2/\sigma_L^2)$ is (a), (d) (0.1, 0.1), (b), (e) (0.1, 0.2) and (c), (f) (0.2, 0.2) (—, target curve; - - - - - , d1; ·····, d5; ·····, d9)

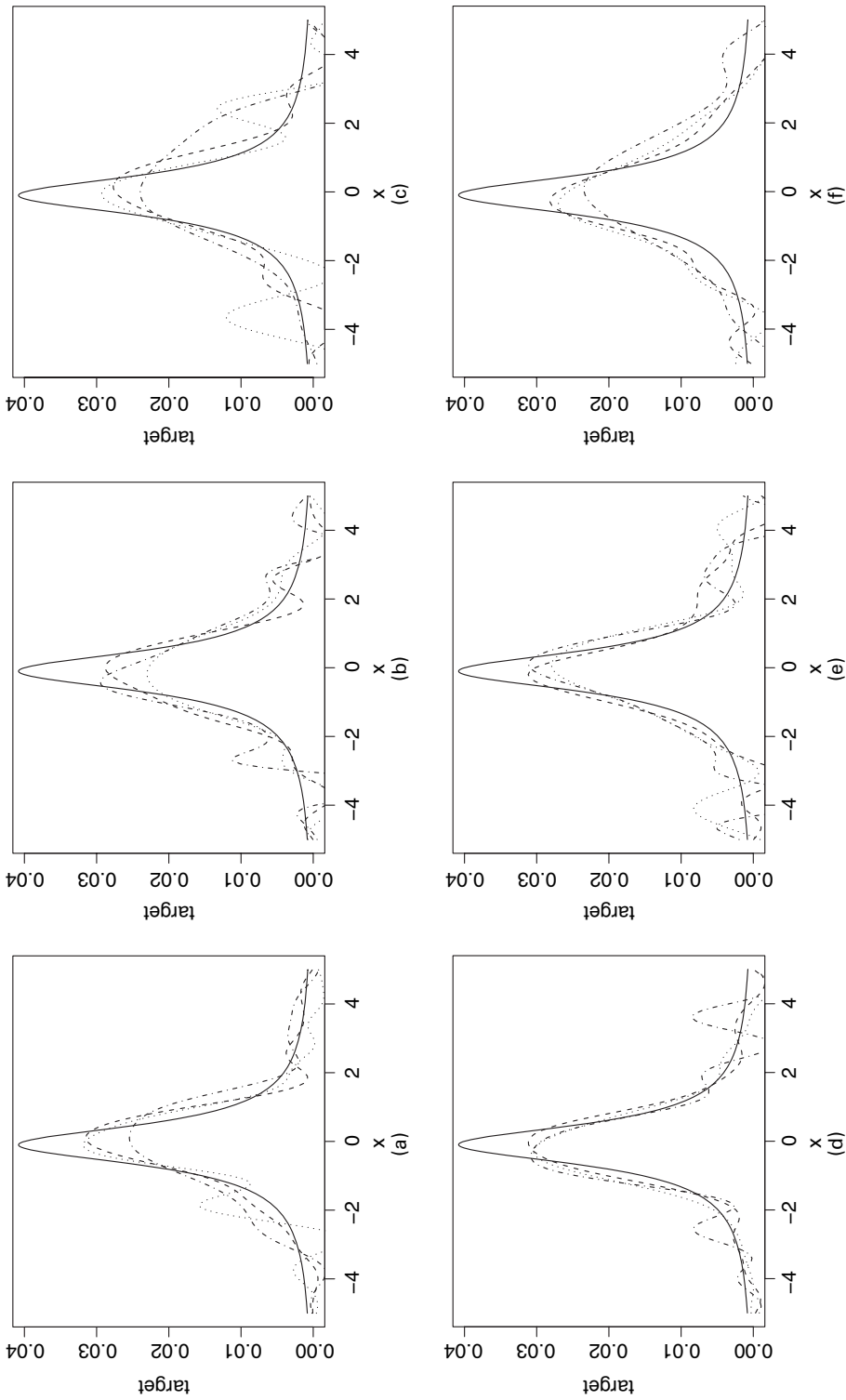


Fig. 2. Estimation of function (a)–(c) $n = 100$ or (d)–(f) $n = 250$, when $L \sim \mathcal{N}(0,2)$, U_B and U_C are Laplace with $(\sigma_B^2/\sigma_L^2, \sigma_C^2/\sigma_L^2)$ equal to (a), (d) $(0.1,0.1)$, (b), (e) $(0.1,0.2)$ and (c), (f) $(0.2,0.2)$ (—, target curve; - - - - - , d1; ······, d5; ······, d9)

normal or Laplace variables, centred at zero and having a variance equal to 10% and 20% respectively of the variance of L . The variable η was $N(0, \sigma_\eta^2)$, where $\sigma_\eta^2 = 0.1 \text{ var}|g|$. Here, $\text{var}|g|$ was defined by $\text{var}|g| = \int_{q_{0.01}}^{q_{0.99}} (|g| - E|g|)^2 / (q_{0.99} - q_{0.01})$, where $E(|g|) = \int_{q_{0.01}}^{q_{0.99}} |g| / (q_{0.99} - q_{0.01})$ and q_α was the α th quantile of $|g|$ rescaled to integrate to 1.

In each case, we considered samples of size $n = 100$ or $n = 250$, we generated 200 replicated samples from the random vector (W, Y) and we constructed the corresponding estimator \hat{g} , by using the data-driven method of Section 3.1 and the kernel K with Fourier transform $K^{Ft}(t) = (1 - t^2)^3 \mathbf{1}_{[-1, 1]}(t)$, which is commonly used in deconvolution problems. We report the integrated squared error $\text{ISE}(x) = \int \{\hat{g}(x) - g(x)\}^2 dx$. In all figures, the estimates that are 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 that are not presented here.

In deconvolution problems it is quite common to consider two classes of errors, called ordinary smooth and supersmooth errors. Roughly, an error of the first and second type respectively has a characteristic function behaving like a negative polynomial and exponential in the tails. Rates of convergence in errors-in-variables problems are 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 that are obtained when estimating curve (c), in the case where $L \sim T_8 \sqrt{0.75}$, 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 Fig. 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, which are not reported here, show that the Laplace error case systematically outperforms the normal error case, which occasionally performs very poorly, especially when σ_B^2 is large.

Fig. 2 illustrates the way in which the estimator improves as the sample size increases. We compare the results that are obtained when estimating curve (a) for samples of size $n = 100$ or $n = 250$. Here, $L \sim N(0, 2)$, 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 the estimators, in all cases, as n increases from 100 to 250.

Fig. 3 illustrates the performance of the estimator in a case where the classical error is smoother than the Berkson error—a situation which is encountered very often in real data applications. We compare the results that are obtained when estimating curve (b) 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, $L \sim N(0, 1)$ and the 200 generated samples are of size $n = 250$. Here, and also in all other cases that 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 Fig. 3 we also illustrate the effect of the errors that are present in the data; we show the local linear estimators that are obtained when ignoring the error, i.e. when using the procedure with plug-in bandwidth described by Fan and Gijbels (1996). The graphs show that ignoring the error leads to severely biased estimators.

Of course, as for any non-parametric method in the usual ‘error-free’ regression problem, 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 σ_B^2/σ_L^2 and σ_C^2/σ_L^2 , the performance of the estimator depends on the variance of U_B, U_C and L . For example, Fig. 4 illustrates the results of estimating regression function (c) in the case where $n = 250, L \sim N(0, \sigma_L^2)$ and $Z \sim \text{Laplace}(\sigma_Z \sqrt{0.5})$, with $Z = U_B$ or $Z = U_C$, 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

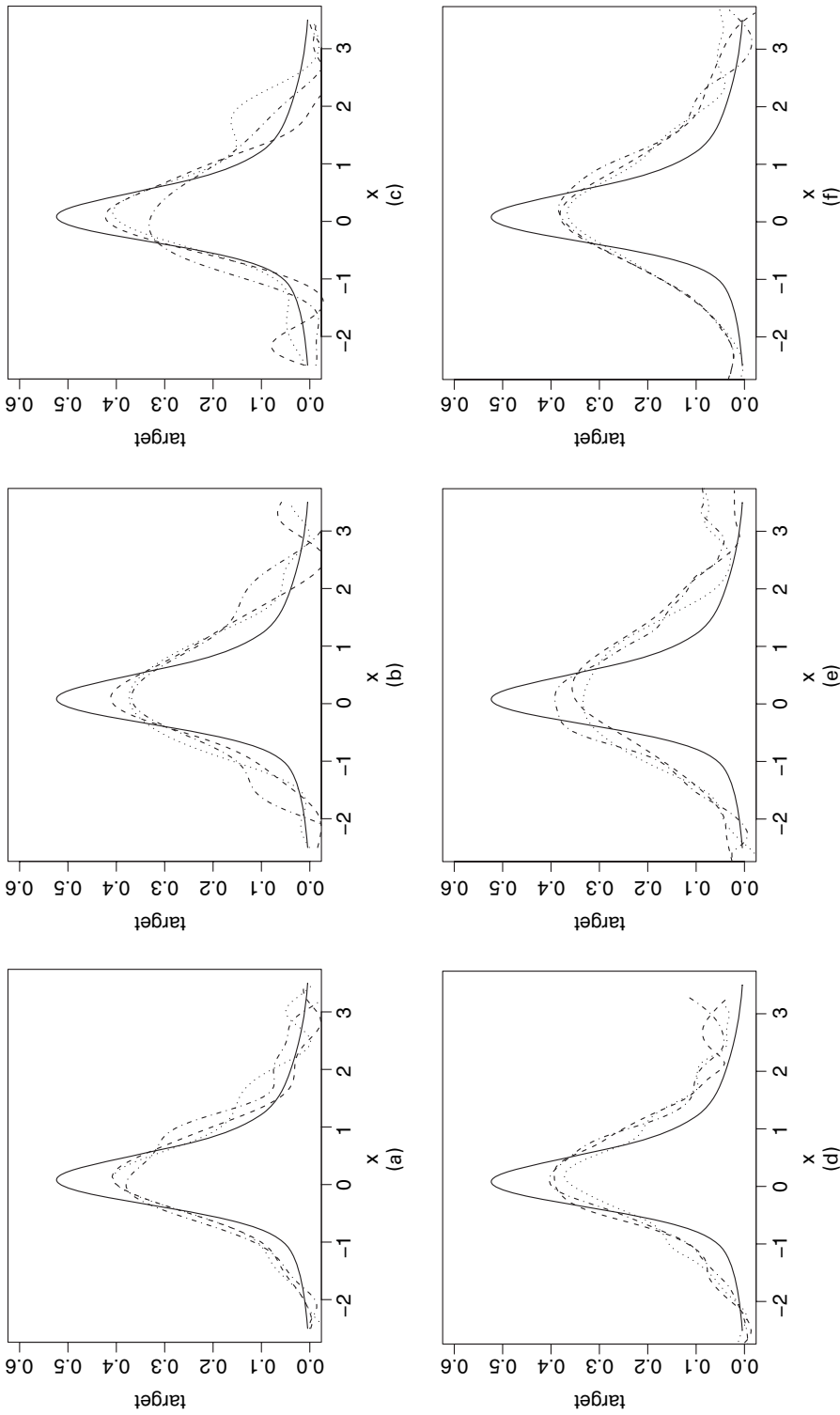


Fig. 3. Estimation of function (b) for samples of size $n = 250$, when $L \sim N(0, 1)$, U_B is Laplace and U_C is normal, with $(\sigma_B^2/\sigma_L^2, \sigma_C^2/\sigma_L^2)$ equal to (a), (d) (0.1, 0.1), (b), (e) (0.1, 0.2) and (c), (f) (0.2, 0.1); (a)–(c) estimator (2.4); (d)–(f) the local linear estimator that ignores the error in the data (—, target curve; - - - - - , d1; ······, d5; - · - · - ·, d9)

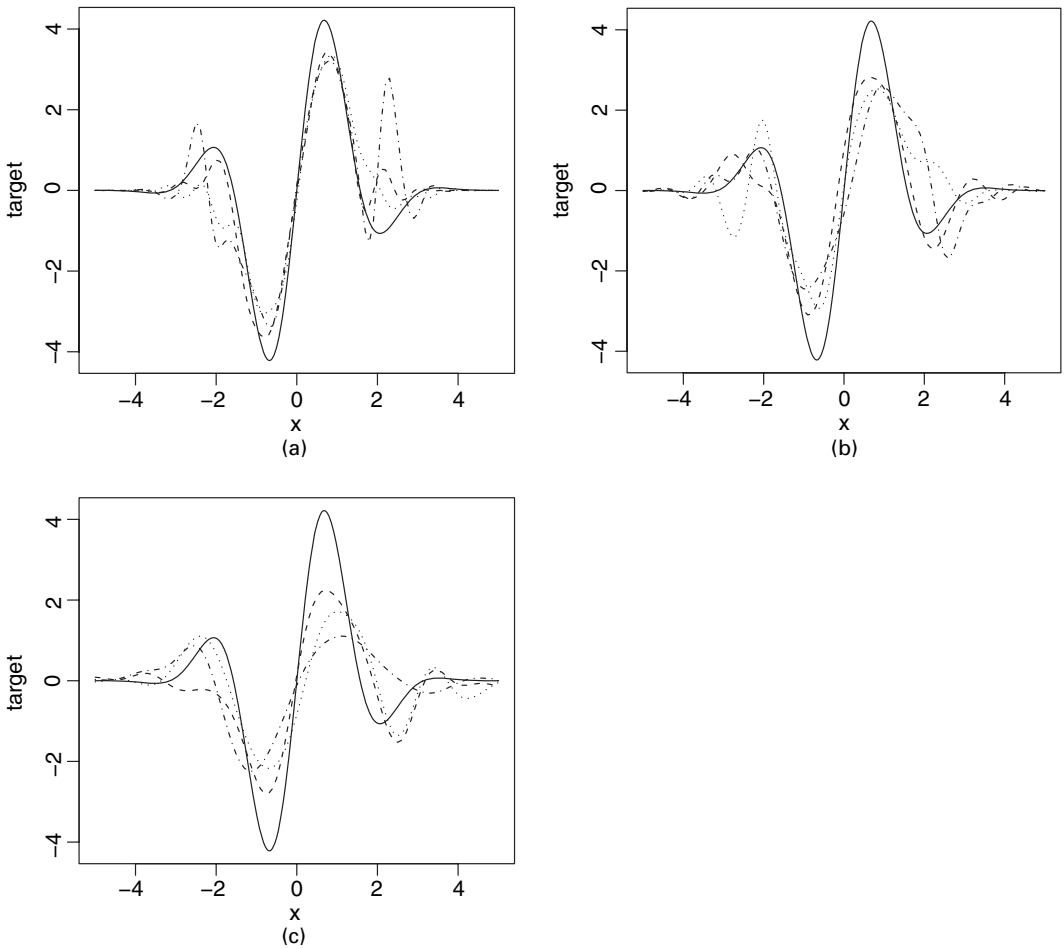


Fig. 4. Estimation of function (c) when $n = 250$, $L \sim N(0, \sigma_L^2)$ and $Z \sim \text{Laplace}(\sigma_Z\sqrt{0.5})$, with $Z = U_B$ or $Z = U_C$, for $(\sigma_L^2, \sigma_B^2, \sigma_C^2)$ equal to (a) (0.5,0.05,0.1), (b) (1,0.1,0.2) or (c) (2,0.2,0.4) (—, target curve; - - - -, d1; ·····, d5; - · - ·, d9)

increase the observations become more widespread, and, for a given sample size, it becomes more difficult 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 $g(x) = x^2$, $L \sim N(0, 1)$ and U_B and U_C are Laplace, with $\sigma_B^2 = \sigma_C^2 = 0.1$. Here, since $|g|$ integrates to ∞ , 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 difficult to deal with such unbounded functions, our estimator can estimate g on a compact interval, of length growing with the sample size. In Fig. 5, we illustrate these results by showing the decile curves that are obtained for samples of size $n = 100, 250, 500$. We see clearly that, as the sample size increases, the estimator can estimate g correctly on growing intervals. 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.

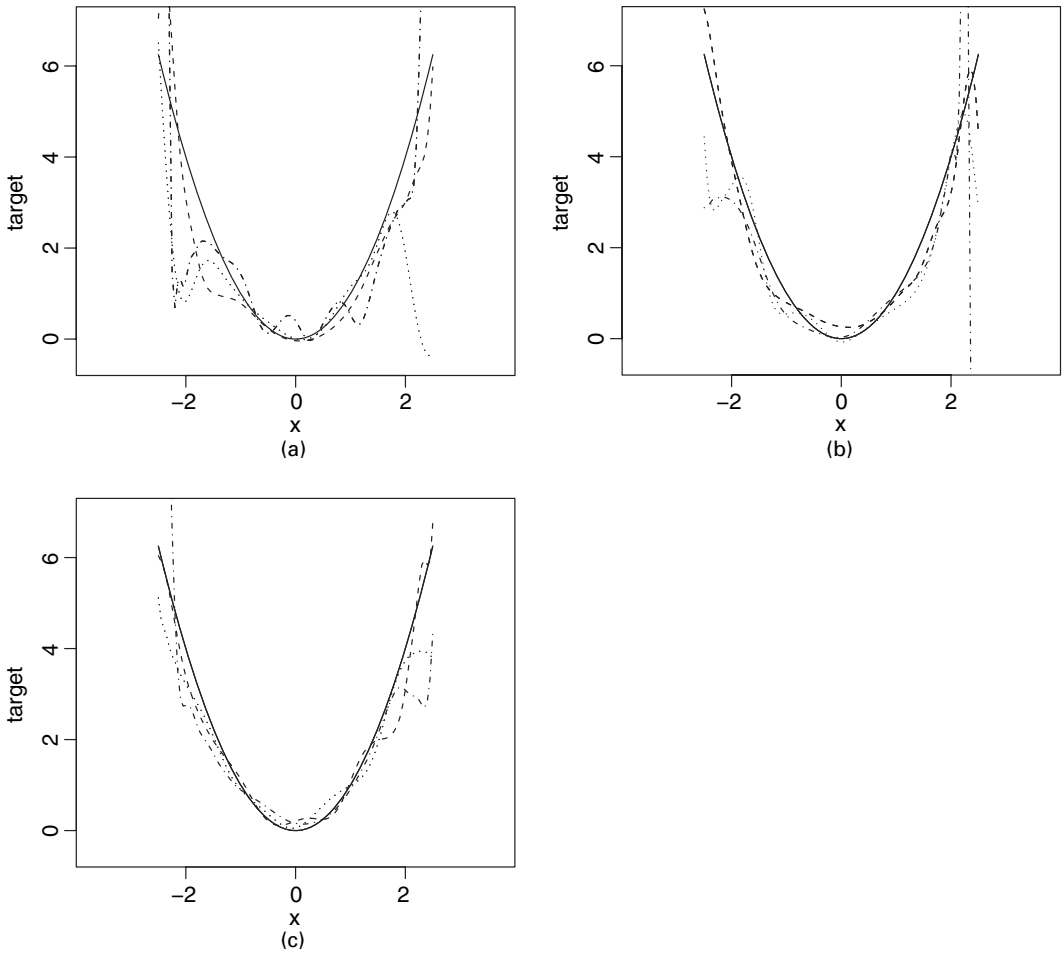


Fig. 5. Estimation of the curve $g(x) = x^2$ when $L \sim N(0,1)$ and $Z \sim \text{Laplace}(\sqrt{0.05})$, with $Z = U_B$ or $Z = U_C$ and the sample size is (a) $n = 100$, (b) $n = 250$ or (c) $n = 500$ (—, target curve; - - - - -, d1; ·····, d5; ······, d9)

3.3. Data example

We applied the kernel method to data from the Nevada test site 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 2000 individuals who had been 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.*, 2006), as is a reanalysis of the thyroid disease data (Lyon *et al.*, 2006). We analyse a subset of the revised dosimetry data, namely the 1278 women in the study, 103 of whom developed thyroiditis.

In this example, X and W respectively are the logarithm of the true and observed radiation exposure and $Y = 0$ or $Y = 1$ 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 illustra-

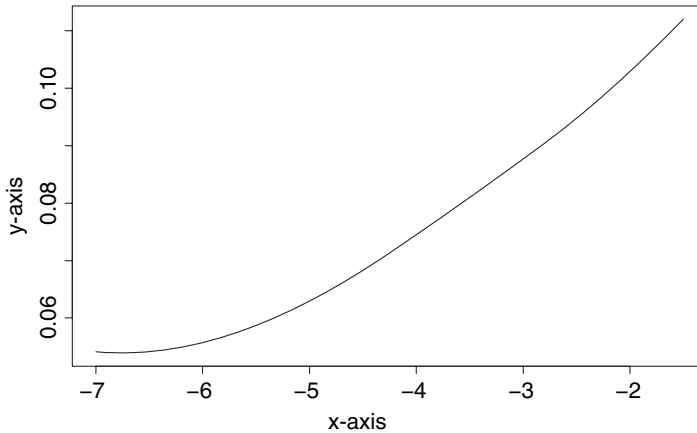


Fig. 6. Estimation of the regression curve for the thyroid data in the log-scale: the x -axis is the logarithm of the true dose, whereas the y -axis is the estimated risk of thyroiditis

tion 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.

We applied our estimation procedure on these data, with smoothing parameters selected via the method that was 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 in Fig. 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 $U_B \equiv U_C \equiv 0$), non-parametric estimation of an unbounded regression curve g that is defined on the whole real line is a difficult problem: in finite samples, the observations are confined to a finite range and, in general, only the observations in the neighbourhood of the point x where we want to estimate g bring valuable information about the value of $g(x)$. Hence, unless $g(x) \rightarrow 0$ as $x \rightarrow \infty$, 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 non-parametric 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 that are 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, $P(|U_B| > x) = \exp(-x)$ for $x > 0$

and that $g(x) = \exp(x^2)$. 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 $1 \leq j \leq n$, converges to 1 as $n \rightarrow \infty$. 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 polynomial g . Although the problem can be very difficult 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 \hat{g} by exploiting its similarities with \hat{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 that are less extreme than the one mentioned in the previous paragraph, \hat{g} and \hat{g}_n are sufficiently close for asymptotic properties of \hat{g} to be derivable from those of \hat{g}_n . More precisely, we assume that

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

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

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

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

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

We shall see at the end of Section 4.3 that, together, results (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 condition (4.4) below, that g may depend on n .

The assumption in expression (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 Section 4.1, we shall permit $g = g_n$ to depend on n , subject to satisfying

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

where $\lambda \geq 1$. It follows that a and b , at expressions (2.1) and (2.3), can also depend on n , although to avoid 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 η .

Biases for the estimators \hat{f}_L and \hat{b} , which are defined at expression (2.3), are respectively given by

$$\text{bias}_f(x) = E\{\hat{f}_L(x) - f_L(x)\} = \frac{1}{2\pi} \int f_L^{\text{Ft}}(t)\{K^{\text{Ft}}(h_1t) - 1\} \exp(-itx) dt,$$

$$\text{bias}_b(x) = E\{\hat{b}(x) - b(x)\} = \frac{1}{2\pi} \int b^{\text{Ft}}(t)\{K^{\text{Ft}}(h_2t) - 1\} \exp(-itx) dt.$$

Define also

$$\text{bias}_g(x) = \frac{1}{2\pi} \int g^{\text{Ft}}(t)\{K^{\text{Ft}}(h_3t) - 1\} \exp(-itx) dt,$$

it being assumed in each case that the integral is convergent in the Riemann sense. To interpret bias_g , consider the case where g is a probability density, and we observe noisy data that are generated as $\zeta = \eta_g + \eta$, where η_g has density g , and η is independent of η_g and has a known distribution with a characteristic function that does not vanish on the real line. Then, 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 bias_b and bias_f , and define also δ , which is closely related to root-mean-squared error:

$$\text{supbi}(h_1) = \max_{c=b,f} \sup_{-\infty < x < \infty} |\text{bias}_c(x)|,$$

$$\delta = \lambda^{-1} \text{supbi}(h_1) + (nh_1^{2\alpha+1})^{-1/2},$$

where λ is as at condition (4.4) and $\alpha > 1$ will be determined by assumption (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 is an open set S containing \mathcal{R} , such that f_L is bounded away from zero on 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 assumption (4.5), in which case part (c) is degenerate and only parts (a) and (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 assumptions (4.6), parts (a)–(d) respectively. With $C_2 > 0$ denoting a constant and $\lfloor \beta \rfloor$ the integer part of β , our assumptions are

for constants α, β and γ 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 $j=1$, 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 condition (4.4). (4.6)

The conditions that are imposed on f_C and f_B are a variation of the assumptions $|f_C^{\text{Ft}}(t)| \geq \text{constant} \times (1 + |t|)^{-\alpha}$ and $|f_B^{\text{Ft}}(t)| \geq \text{constant} \times (1 + |t|)^{-\beta}$ 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, part (a) holds with $\alpha = 2\omega > 1$, and part (b) holds with $\beta = 2\omega$ and γ depending on ω . More particularly, the case where f_{-B}^{Ft} is the inverse of a polynomial, which occurs if, for example, $f_{-B}^{Ft}(t) = (1+t^2)^{-\omega}$ and ω is an integer, is of special interest. More generally, suppose that

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

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

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

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

and the value of γ is limited only by the size of r_2 ; γ does not depend on selection of the constants p and c_1, \dots, c_p in assumption (4.7). In particular, by choosing r_2 sufficiently large we can take γ arbitrarily large in condition (4.6). These considerations generally permit us to take $\xi = \infty$ in part (c) of condition (4.5); see the discussion immediately below theorem 1. In such cases, condition (4.5) imposes especially mild conditions on f_L .

More generally, conditions (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 bias_b , bias_f and 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 \hat{g} is given by equation (2.4), with \hat{f}_L and \hat{b} defined at expression (2.3). For simplicity we omit the case $\beta + \gamma = \xi$ in expression (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 1. If conditions (4.5) and (4.6) hold, and if $h_1 = h_2$ and $0 < h_1, h_3 \leq B_1$ where $B_1 > 0$, then, for a constant $B_2 > 0$ not depending on n, h_1, h_2 or h_3 ,

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

where λ is as at condition (4.4).

To interpret this theorem, let us first consider the case where g is a bounded integrable func-

tion. Then the contribution of g to bias is represented in expression (4.9) by

$$\text{bias}_g(x) = \int K(u)\{g(x - h_3u) - g(x)\} du = O(h_3^k), \tag{4.10}$$

where the first identity holds under the conditions of theorem 1, and the second identity holds provided that 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 $j = 1, \dots, k - 1$. Kernels satisfying these conditions, as well as those in part (c) of condition (4.6), are commonly used in practice. For example, the kernels that are employed in Section 3 are of this type for $k = 2$.

Bias formulae such as equation (4.10) are of course conventional. It is the remaining contribution to the convergence rate, bounded by the right-hand side of expression (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 ρ to equal a constant multiple of δ and assume that we can choose γ , in condition (4.6), so large that $h_3^2 = O(\rho)$ (this assumption is not an issue if f_{-B} satisfies assumption (4.7)). Related interpretations of expression (4.9) are also possible where the simplifications that are obtainable when f_{-B} is given by expression (4.7) do not apply, but those instances are not so transparent, since then both γ and ξ can impact on the overall convergence rate). Then expression (4.9) further simplifies to

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

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

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

If f_L has k bounded derivatives, then $\delta \sim \text{constant} \times n^{-k/(2\alpha+2k+1)}$ provided that we take $h_1 \sim \text{constant} \times n^{-1/(2\alpha+2k+1)}$. This order of δ is also the minimax optimal, root squared error convergence rate for estimators of g , in the case where U_B is identically 0. See, for example, Fan and Truong (1993). Thus, the factor $h_3^{-\beta}$, on the right-hand side of equation (4.13), can be interpreted as the amount by which the conventional convergence rate δ is degraded by introducing the additional error U_B .

Of course, the factor $h_3^{-\beta}$ diverges as the bandwidth h_3 becomes smaller. In contrast, the bias term h_3^k reduces to 0 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

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

Result (4.9) reveals the potential deleterious effects of taking the ridge ρ too large (i.e. of larger order than δ) or too small. In particular, the order of magnitude of the right-hand side of expression (4.9) is made larger by choosing ρ to be of either strictly larger order, or strictly smaller order, than δ .

It is straightforward to combine theorem 1 and the results in Section 4.1, to handle the case of fixed but unbounded g . Specifically, taking $g_n = g \mathbf{1}_{[-n^D, n^D]}(x)$, with $D > 0$ arbitrarily small, and assuming that assumption (4.1) holds, result (4.2) permits us to take $\lambda = O(n^r)$ for any $r > 0$

in expression (4.9), provided that we replace $\text{bias}_g(x)$ there by

$$\text{bias}_{g_n}(x) = \int K(u)\{g_n(x - h_3u) - g_n(x)\} du = O(n^s h_3^k) \tag{4.15}$$

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

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

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

Note also from equations (4.14) and (4.16) that our bounds on the rate of convergence increase as the smoothness of either error distribution increases, i.e. as α or β increases. These results correctly suggest that if either of the errors were ‘supersmooth’, e.g. 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 *et al.* (2006) that no non-parametric 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 that we describe below can 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 $l = 1, 2$, $k_{l,j} = \pi^{-1} \int_{\mathcal{I}} k(x) \text{cs}_{l,j}(x) dx$, where $\text{cs}_{l,j}(x)$ is $\cos(jx)$ or $\sin(jx)$ according to whether $l = 1$ or $l = 2$ respectively. Using the sine–cosine decomposition of g and a , we have, from Delaigle *et al.* (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 $l = 1, 2$, $(c_{l,j}, \delta_{l,j}, r_{l,j})$ denotes $(g_{l,j}, \alpha_{l,j}, a_{l,j})$, with $\alpha_{l,j} = E\{\text{cs}_{l,j}(U_B)\}$, and an estimator of the Fourier coefficients of g can be deduced from equation (5.1), by replacing the Fourier coefficients a_0 and $a_{l,j}$ by $\hat{a}_0 = (2\pi)^{-1} \int_{\mathcal{I}} \hat{a}$ and $\hat{a}_{l,j} = \pi^{-1} \int_{\mathcal{I}} \hat{a} \text{cs}_{l,j}$, where we choose \hat{a} to be a sine–cosine series estimator of a , which is defined by borrowing ideas of Hall and Qiu (2005) in the context of pure classical errors.

More precisely, define $b = a f_L$, $p_{l,j} = \pi^{-1} E\{\text{cs}_{l,j}(W)\}$ and $q_{l,j} \equiv \pi^{-1} E\{Y \text{cs}_{l,j}(W)\}$, for $l = 1, 2$. Then it can be shown that result (5.1) holds with $(c_{l,j}, \delta_{l,j}, r_{l,j})$ equal to either $(f_{L,l,j}, \beta_{l,j}, p_{l,j})$ or $(b_{l,j}, \beta_{l,j}, q_{l,j})$, where $\beta_{l,j} = E\{\text{cs}_{l,j}(U_C)\}$, $l = 1, 2$. Substituting the estimators

$$\hat{p}_{l,j} = (\pi n)^{-1} \sum_i \text{cs}_{l,j}(W_i)$$

and

$$\hat{q}_{l,j} = (\pi n)^{-1} \sum_i Y_i \text{cs}_{l,j}(W_i)$$

for $p_{l,j}$ and $q_{l,j}$, we see that a can be estimated by

$$\hat{a}(x) = \frac{\hat{b}_0 + \sum_{j \geq 1} \{\hat{b}_{1j} \cos(jx) + \hat{b}_{2j} \sin(jx)\}}{(2\pi)^{-1} + \sum_{j \geq 1} \{\hat{f}_{L1j} \cos(jx) + \hat{f}_{L2j} \sin(jx)\}}.$$

In practice, we need to truncate the series for \hat{a} and to keep only 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 \hat{g} needs also to 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 that was introduced in Delaigle *et al.* (2006).

Acknowledgements

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.

Appendix A: Outline proof of theorem 1

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

$$\hat{g}(x) - g(x) = \text{bias}_g(x) - \rho \int \frac{ak_x}{f_L + \rho} + A_b(x) - A_f(x) + Q_2(x), \tag{A.1}$$

where

$$\left. \begin{aligned} A_b(x) &= \int \Delta_b k_x / (f_L + \rho), \\ A_f(x) &= \int b \Delta_f k_x / (f_L + \rho)^2, \\ |Q_2(x)| &\leq \int Q_1 |k_x|. \end{aligned} \right\} \tag{A.2}$$

Using parts (b) and (c) of condition (4.6) it can be shown that $|K_{-B}(x|h)| \leq C_1 h^{-\beta} (1 + |x|)^{-(\beta+\gamma+1)}$ for all real x and all $h > 0$, where C_1, C_2, \dots will denote positive constants. This leads to the result $h_3^\beta \int |k_x|(f_L + \rho)^{-1} \leq C_2(I_1 + I_2)$, where, with $C > 0$ chosen so small that $x + h_3u \in \mathcal{S}$ whenever $x \in \mathcal{R}$ and $|h_3u| \leq C$, and \mathcal{R} and \mathcal{S} as in condition (4.5), we define $I_1 = \int_{|h_3u| \leq C} (1 + |u|)^{-(\beta+\gamma+1)} du \leq C_3$,

$$I_2 = \int_{|h_3u| > C} \{(h_3|u|)^{-\xi} + \rho\}^{-1} (1 + |u|)^{-(\beta+\gamma+1)} du \leq C_4 \begin{cases} h_3^{\beta+\gamma} \rho^{(\beta+\gamma)/\xi-1} & \text{if } \beta + \gamma < \xi, \\ h_3^{\beta+\gamma} & \text{if } \beta + \gamma > \xi. \end{cases}$$

Combining the results in this paragraph we deduce that

$$h_3^\beta \int \frac{|k_x|}{f_L + \rho} \leq C_5 \begin{cases} (1 + h_3^{\beta+\gamma} \rho^{(\beta+\gamma)/\xi-1}) & \text{if } \beta + \gamma < \xi, \\ (1 + h_3^{\beta+\gamma}) & \text{if } \beta + \gamma > \xi, \end{cases} \tag{A.3}$$

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

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

$$E\{\Delta_b(u)^2\} + |a|^2 E\{\Delta_f(u)^2\} \leq C_6 \lambda^2 \delta^2. \tag{A.4}$$

Using expressions (A.3), (A.4) and the result $|b| \leq C_7 \lambda f_L$, it can be proved that, for $c = b$ and $c = f$,

$$E\{A_c(x)^2\}^{1/2} \leq C_8 \lambda \delta h_3^{-\beta} \begin{cases} (1 + h_3^{\beta+\gamma} \rho^{(\beta+\gamma)/\xi-1}) & \text{if } \beta + \gamma < \xi, \\ (1 + h_3^{\beta+\gamma}) & \text{if } \beta + \gamma > \xi, \end{cases} \tag{A.5}$$

$$\rho \left| \int \frac{ak_x}{f_L + \rho} \right| \leq C_9 \lambda \rho h_3^{-\beta} \begin{cases} (1 + h_3^{\beta+\gamma} \rho^{(\beta+\gamma)/\xi-1}) & \text{if } \beta + \gamma < \xi, \\ (1 + h_3^{\beta+\gamma}) & \text{if } \beta + \gamma > \xi, \end{cases} \tag{A.6}$$

where both formulae hold uniformly in $x \in \mathcal{R}$. Together, expressions (A.1), (A.5) and (A.6) give

$$\hat{g}(x) - g(x) = \text{bias}_g(x) + Q_2(x) + Q_3(x), \tag{A.7}$$

where Q_2 is as before, and so satisfies the last inequality at expression (A.2), and

$$E\{Q_3(x)^2\}^{1/2} \leq C_{10} \lambda (\rho + \delta) h_3^{-\beta} \begin{cases} (1 + h_3^{\beta+\gamma} \rho^{(\beta+\gamma)/\xi-1}) & \text{if } \beta + \gamma < \xi, \\ (1 + h_3^{\beta+\gamma}) & \text{if } \beta + \gamma > \xi, \end{cases} \tag{A.8}$$

uniformly in $x \in \mathcal{R}$. Properties (A.3) and (A.4), and the last inequality at expression (A.2), entail

$$E|Q_2(x)| \leq C_{11} \lambda \delta^2 \rho^{-1} h_3^{-\beta} \begin{cases} (1 + h_3^{\beta+\gamma} \rho^{(\beta+\gamma)/\xi-1}) & \text{if } \beta + \gamma < \xi, \\ (1 + h_3^{\beta+\gamma}) & \text{if } \beta + \gamma > \xi. \end{cases} \tag{A.9}$$

Results (A.7)–(A.9) imply that $\hat{g}(x) - g(x) = \text{bias}_g(x) + Q_4(x)$, where, uniformly in $x \in \mathcal{R}$,

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

Theorem 1 follows directly from these properties.

References

Berkson, J. (1950) Are there two regression problems? *J. Am. Statist. Ass.*, **45**, 164–180.
 Buonaccorsi, J. P. and Lin, C.-D. (2002) Berkson measurement error in designed repeated measures studies with random coefficients. *J. Statist. Planng 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*, 2nd edn. Boca Raton: Chapman and Hall–CRC.
 Cook, J. R. and Stefanski, L. A. (1994) Simulation-extrapolation estimation in parametric measurement error models. *J. Am. Statist. Ass.*, **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. R. 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*. London: Chapman and Hall.
 Fan, J. and Masry, E. (1992) Multivariate regression estimation with errors-in-variables: asymptotic normality for mixing processes. *J. Multiv. 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.
 Huang, L. and Huang, H. Y. S. (2000) On errors-in-variables in polynomial regression—Berkson case. *Statist. Sin.*, **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. Am. Med. Ass.*, **270**, 2076–2083.

- Kim, J. and Gleser, L. J. (2000) SIMEX approaches to measurement error in ROC studies. *Communs 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 be published.
- Linton, O. and Whang, Y. J. (2002) Nonparametric estimation with aggregated data. *Econometr. 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. *Statist. Med.*, **17**, 2157–2177.
- Schafer, D. W. and Gilbert, E. S. (2006) Some statistical implications of dose uncertainty in radiation dose-response analyses. *Radian Res.*, **166**, 303–312.
- Simon, S. L., Anspaugh, L. R., Hoffman, F. O., Scholl, A. E., Stone, M. B., Thomas, B. A. and Lyon, J. L. (2006) Update of dosimetry for the Utah Thyroid Cohort Study. *Radian Res.*, **165**, 208–222.
- 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 Phys.*, **68**, 460–471.
- Stefanski, L. A. and Cook, J. R. (1995) Simulation-extrapolation: the measurement error jackknife. *J. Am. Statist. Ass.*, **90**, 1247–1256.
- Stevens, W., Till, J. E., Thomas, D. C., Lyon, J. L., Kerber, R. A., Preston-Martin, S., Simon, S. L., Rallison, M. L. and Lloyd, R. D. (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, Salt Lake City.
- 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? *Am. 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. Sin.*, **13**, 1201–1210.