

# Low order approximations in deconvolution and regression with errors in variables

Raymond J. Carroll

*Texas A&M University, College Station, USA*

and Peter Hall

*Australian National University, Canberra, Australia*

[Received September 2002. Revised April 2003]

**Summary.** We suggest two new methods, which are applicable to both deconvolution and regression with errors in explanatory variables, for nonparametric inference. The two approaches involve kernel or orthogonal series methods. They are based on defining a low order approximation to the problem at hand, and proceed by constructing relatively accurate estimators of that quantity rather than attempting to estimate the true target functions consistently. Of course, both techniques could be employed to construct consistent estimators, but in many contexts of importance (e.g. those where the errors are Gaussian) consistency is, from a practical viewpoint, an unattainable goal. We rephrase the problem in a form where an explicit, interpretable, low order approximation is available. The information that we require about the error distribution (the error-in-variables distribution, in the case of regression) is only in the form of low order moments and so is readily obtainable by a rudimentary analysis of indirect measurements of errors, e.g. through repeated measurements. In particular, we do not need to estimate a function, such as a characteristic function, which expresses detailed properties of the error distribution. This feature of our methods, coupled with the fact that all our estimators are explicitly defined in terms of readily computable averages, means that the methods are particularly economical in computing time.

**Keywords:** Density estimation; Measurement error; Nonparametric regression; Orthogonal series; Simulation–extrapolation

## 1. Introduction

Suppose that we observe the value of

$$W = X + U, \quad (1)$$

where the random variables  $X$  and  $U$  are independently distributed. We either know or have data on the distribution of  $U$ , and we wish to estimate the density or distribution of  $X$ . This is a classical deconvolution problem in statistics. Its contemporary applications date at least from work of Mendelsohn and Rice (1982), on the deconvolution of microfluorescence data, and have generated much methodological interest. Carroll and Hall (1988) addressed the problem of optimal deconvolution in the case where  $U$  has a normal distribution. They showed that there the fastest possible convergence rate is only logarithmic in sample size, the latter denoted by  $n$ , say. This implies that the problem of consistent estimation is, unless the variance of  $U$  is small, effectively insoluble in practical terms. Fan (1991) treated settings where the optimal

*Address for correspondence:* Raymond J. Carroll, Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA.  
E-mail: carroll@stat.tamu.edu

convergence rate is polynomial in  $n$ , and Fan (1992) discussed the contrary case where the rate is particularly poor. Even when the rate is polynomial, it is often particularly slow unless the density of  $U$  is so unsmooth as to contain a discontinuity. See also Efromovich (1997) and Wand (1998). Further references will be given later.

These results argue that the problem of inference about the density or distribution of  $X$  should be treated differently from in more standard statistical contexts. Since consistent estimation is so difficult in many important cases, it can be argued that we should not attempt to estimate the actual density,  $f_X$  say, of  $X$ . Instead we should estimate a function that, in a well-defined sense, approximates  $f_X$  and is estimable relatively efficiently. In this paper we suggest two approaches to this problem, based on kernel or orthogonal series methods. Both require some knowledge of the distribution of  $U$ . However, the necessary information is very rudimentary, being based only on low order moments, and is frequently available either from a sample drawn from the distribution of  $U$  or from replications of observations of  $W$ —i.e. small numbers of repeated observations of  $W$  for the same  $X$  but different values of  $U$ .

It should be emphasized that we are not estimating  $f_X$ , but estimating an approximation to  $f_X$ . From this viewpoint our approach is similar to that used in many dimension reduction problems: solving the problem at hand is infeasible, in both theory and practice, so we change the problem to one which captures the main features of interest and solve that instead. Since the target of our attention is no longer  $f_X$  then traditional measures of performance, e.g. the distance of our empirical approximation from  $f_X$  (see for example Carroll and Hall (1988) and Fan (1991)), are no longer relevant.

The first of our two methods is based on the observation that we can express the expected value of a kernel estimator of  $f_X$  as a series expansion in expectations of kernel estimators of derivatives of the density  $f_W$  of  $W$ , and that coefficients in the series depend only on moments of the distribution of  $U$ . By truncating the expansion we obtain a readily computable estimator of a low order approximation to  $f_X$ . Details of this technique will be given in Section 2. Of course, approximations to the distribution  $F_X$  of  $X$  can be found simply by integrating the density approximations.

The basis of the second method is a formula for expressing  $f_X$  in an orthogonal expansion with estimable coefficients. The coefficient estimators depend on the distribution of  $U$  only through its moments, but the functions in the orthogonal series can be virtually arbitrary. For example, they may be polynomials or trigonometric functions. Therefore, the type of orthogonal sequence can be chosen to reflect prior belief about the distribution of  $X$ . For example, if that distribution is believed to be supported on the whole real line then we might take the  $j$ th term in the series to be proportional to  $H_j(\tau x) \exp(-\frac{1}{2}\tau^2 x^2)$ , where  $H_j$  denotes the  $j$ th Hermite polynomial. The factor  $\exp(-\frac{1}{2}\tau^2 x^2)$  forces the density approximation to decrease to 0 in the tails; larger values of  $\tau$  accommodate lighter tails. However, if the  $X$ -distribution is known to be supported on a compact interval  $\mathcal{I} = [a, b]$ , and to descend to 0 at the ends of the interval, then we might take the orthogonal sequence to be of polynomials multiplied by  $\{1 - (2x - a - b)(b - a)^{-1}\}^c$ , for some  $c > 0$ , where the weight function is incorporated to force the tails of the density approximation to 0 at the ends of  $\mathcal{I}$ . The polynomials in the sequence are of course chosen to be orthogonal relative to this weight. Details of the method will be given in Section 4.

Both techniques enable estimation of derivatives of  $f_X$ , as well as  $f_X$  itself. And both have application to the problem of nonparametric regression with errors in the explanatory variables. There we observe  $Y$ , assumed to be generated by the model

$$Y = g(X) + \varepsilon, \tag{2}$$

where  $g$  is a smooth function and  $\varepsilon$  represents an error in the response variable  $Y$ . However, rather than observe  $X$  we have data only on  $W$ , given by equation (1). From the pairs  $(W, Y)$  we wish to estimate the function  $g$ . Both our methods for solving the problem represented by equation (1) can be used to construct estimators of  $g$ . In this setting they have the advantage, over competing approaches, of substantial computational simplicity. Details of our estimators of  $g$  will be given in Sections 3 and 4.

The main goal of this paper is to discuss inference under model (1), where we wish to estimate the density  $f_X$  (or the distribution  $F_X$ ) by using data on  $W$  and rudimentary moment-type information about the distribution  $F_U$  of  $U$ . In the past it has been quite common to suppose that the distribution of  $U$  is completely known, because the problem has been difficult to solve otherwise. Nevertheless, little evidence was generally available to support such an assumption. The methodology that is suggested in this paper allows the assumption to be removed.

Estimators in this problem have been discussed by, for example, Devroye (1989), Liu and Taylor (1989), Stefanski (1990), Stefanski and Carroll (1990), Zhang (1990), Hesse (1995a, b, 1999), Goldenshluger (1999), van Es and Uh (2000) and Cator (2001). Among the particular methodologies that have been developed, Masry (1993) and Fotopoulos (2000) have addressed applications to dependent data, Cordy and Thomas (1997) have proposed methods for deconvolution of a distribution function, Jongbloed and van Zuijlen (1998) have considered deconvolution when  $f_X$  has a discontinuity, Pensky and Vidakovic (1999) and Wang (1999) have treated wavelet-based methods, Zhang and Karunamuni (2000) have developed techniques for boundary bias correction, Yuan and Chen (2002) have treated the multivariate case and Youndje and Wells (2002) have suggested cross-validation methods for bandwidth choice.

There is a particularly extensive literature on solving the problem that is posed by model (2), where we wish to estimate the function  $g$  from data on  $(Y, W)$  and information about  $F_U$ . We mention here only work that is directly related to the contributions of the present paper. Cook and Stefanski (1994) introduced the ‘simulation–extrapolation’ (SIMEX) method. Stefanski and Bay (1996) used a related approach to address the problem evinced by equation (1), whereas Stefanski and Cook (1995), Carroll *et al.* (1996, 1999) and Staudenmeyer and Ruppert (2004) developed theory for SIMEX and related approaches: the latter has an efficient bandwidth implementation. There is a wide variety of errors-in-variables problems related to that summarized by model (2), and to which solutions can be obtained by modifying the methods that are outlined in Sections 3 and 4. See, for example, the problems discussed by Wang *et al.* (1996), Gould *et al.* (1997), Luo *et al.* (1998), Holcomb (1999), Lin and Carroll (1999, 2000), Kim and Gleser (2000), Chen and Cowling (2001) and Novick and Stefanski (2002). Carroll *et al.* (1995), Thurigen *et al.* (2000) and Stefanski (2000) have reviewed research on problems involving measurement error.

## 2. Kernel methods for estimating $f_X$

Assume that data  $W_i = X_i + U_i$ , for  $1 \leq i \leq n$ , are generated by model (1). Suppose also that we know, or have data from which we can estimate, low order moments of the distribution of  $U$ . From this information we wish to estimate  $f_X$ . We shall assume that  $E(U) = 0$ : knowing the location of the distribution  $F_U$  of  $U$  is necessary to ensure identification of the distribution of  $X$ , given the distribution of  $W$ . However, in many cases of practical importance, even knowing  $F_U$  entirely has little effect on the fact that even optimal estimators of  $f_X$  converge so slowly that, for all practical purposes,  $f_X$  is not estimable. Instead we shall suggest estimable, low order approximations to  $f_X$ , depending on  $F_U$  only through its moments.

Let  $f_W$  denote the density of the distribution of  $W$ . Estimators of  $f_W$  and  $f_X$  are given by

$$\hat{f}_W(w) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{w - W_i}{h}\right),$$

$$\hat{f}_X(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)$$

respectively, where  $K$  is a kernel and  $h$  a bandwidth. Of course, we do not observe the  $X_i$ s, and so  $\hat{f}_X$  cannot be computed directly from data. Nevertheless we can aspire to producing a good approximation to it, or at least to its expected value. This motivates our methodology.

Let  $K^{(j)}$  denote the  $j$ th derivative of  $K$ , and define

$$\lambda_j(x) = E\left\{K^{(j)}\left(\frac{x - X}{h}\right)\right\},$$

$$\kappa_j(x) = E\left\{K^{(j)}\left(\frac{x - W}{h}\right)\right\},$$

$$u_j = E(U^j).$$

Assume for the time being that all moments of the distribution of  $U$  are finite, and that the kernel  $K$  is an analytic function. In particular, all derivatives of  $K$  are well defined on the whole real line. The analyticity condition is fulfilled by the Gaussian kernel. We claim that, for each  $j \geq 1$ ,

$$\lambda_j(x) = \kappa_j(x) + \sum_{r=1}^{\infty} \sum_{k_1=1}^{\infty} \dots \sum_{k_r=1}^{\infty} \frac{(-1)^{k_1+\dots+k_r+r}}{k_1! \dots k_r! h^{k_1+\dots+k_r}} u_{k_1} \dots u_{k_r} \kappa_{k_1+\dots+k_r+j}(x), \quad (3)$$

where the multiple infinite series converge absolutely. A proof of claim (3) will be given in Appendix A. Multiplying equation (3) throughout by  $h^{-j-1}$  we obtain an expression for  $E\{\hat{f}_X^{(j)}(x)\}$  in terms of  $E\{\hat{f}_W^{(j)}(x)\}$ :

$$E\{\hat{f}_X^{(j)}(x)\} = E\{\hat{f}_W^{(j)}(x)\} + \sum_{r=1}^{\infty} \sum_{k_1=1}^{\infty} \dots \sum_{k_r=1}^{\infty} \frac{(-1)^{k_1+\dots+k_r+r}}{k_1! \dots k_r!} u_{k_1} \dots u_{k_r} E\{\hat{f}_W^{(k_1+\dots+k_r+j)}(x)\}. \quad (4)$$

A ‘ $\nu$ th-order approximation’ to the right-hand side of equation (4) is one that takes account of moment products up to and including order  $\nu$ . Such an approximation is given by

$$E\{\hat{f}_W^{(j)}(x)\} + \sum_{r \geq 1, k_1 \geq 1, \dots, k_r \geq 1: k_1+\dots+k_r \leq \nu} \dots \sum_{k_r=1}^{\infty} \frac{(-1)^{k_1+\dots+k_r+r}}{k_1! \dots k_r!} u_{k_1} \dots u_{k_r} E\{\hat{f}_W^{(k_1+\dots+k_r+j)}(x)\}. \quad (5)$$

Also, the  $k$ th derivative of  $h^{k+1} \hat{f}_W(x)$  is an unbiased estimator of  $\kappa_k(x)$ . Therefore, if  $\hat{u}_j$  is an estimator of (or another type of approximation to)  $u_j$ , then

$$\hat{f}_W^{(j)}(x) + \sum_{r \geq 1, k_1 \geq 1, \dots, k_r \geq 1: k_1+\dots+k_r \leq \nu} \dots \sum_{k_r=1}^{\infty} \frac{(-1)^{k_1+\dots+k_r+r}}{k_1! \dots k_r!} \hat{u}_{k_1} \dots \hat{u}_{k_r} \hat{f}_W^{(k_1+\dots+k_r+j)}(x)$$

is an estimator of the quantity (5). Therefore the following is an estimator of the  $\nu$ th-order approximation to  $E\{\hat{f}_X^{(j)}(x)\}$ :

$$\hat{f}_{X,\nu}^{(j)}(x) = \hat{f}_W^{(j)}(x) + \sum_{r \geq 1, k_1 \geq 1, \dots, k_r \geq 1: k_1+\dots+k_r \leq \nu} \dots \sum_{k_r=1}^{\infty} \frac{(-1)^{k_1+\dots+k_r+r}}{k_1! \dots k_r!} \hat{u}_{k_1} \dots \hat{u}_{k_r} \hat{f}_W^{(k_1+\dots+k_r+j)}(x). \quad (6)$$

In particular, second-, fourth- and sixth-order approximations to  $E\{\hat{f}_X^{(j)}(x)\}$  are given by

$$\hat{f}_{X,2}^{(j)}(x) = \hat{f}_W^{(j)}(x) + \hat{u}_1 \hat{f}_W^{(j+1)}(x) + \frac{1}{2}(2\hat{u}_1^2 - \hat{u}_2) \hat{f}_W^{(j+2)}(x), \quad (7)$$

$$\begin{aligned} \hat{f}_{X,4}^{(j)}(x) = & \hat{f}_{X,2}^{(j)}(x) + \frac{1}{6}(6\hat{u}_1^3 - 6\hat{u}_1\hat{u}_2 + \hat{u}_3) \hat{f}_W^{(j+3)}(x) \\ & + \frac{1}{24}(24\hat{u}_1^4 - 36\hat{u}_1^2\hat{u}_2 + 8\hat{u}_1\hat{u}_3 + 6\hat{u}_2^2 - \hat{u}_4) \hat{f}_W^{(j+4)}(x), \end{aligned} \quad (8)$$

$$\begin{aligned} \hat{f}_{X,6}^{(j)}(x) = & \hat{f}_{X,4}^{(j)}(x) + \frac{1}{120}(120\hat{u}_1^5 - 240\hat{u}_1^3\hat{u}_2 + 60\hat{u}_1^2\hat{u}_3 \\ & + 180\hat{u}_1\hat{u}_2^2 + 10\hat{u}_1\hat{u}_4 - 20\hat{u}_2\hat{u}_3 + \hat{u}_5) \hat{f}_W^{(j+5)}(x) \\ & + \frac{1}{720}(720\hat{u}_1^6 - 1800\hat{u}_1^4\hat{u}_2 + 480\hat{u}_1^3\hat{u}_3 + 1080\hat{u}_1^2\hat{u}_2^2 - 90\hat{u}_1^2\hat{u}_4 \\ & + 12\hat{u}_1\hat{u}_5 - 360\hat{u}_1\hat{u}_2\hat{u}_3 - 90\hat{u}_2^3 + 20\hat{u}_3^2 + 30\hat{u}_2\hat{u}_4 - \hat{u}_6) \hat{f}_W^{(j+6)}(x). \end{aligned} \quad (9)$$

It is common to assume that the distribution of the error  $U$  is symmetric, in which case we would take  $\hat{u}_k = 0$  for odd  $k$ , and look only at approximations of even order. Then formula (6) simplifies to

$$\hat{f}_{X,2\nu}^{(j)}(x) = \hat{f}_W^{(j)}(x) + \sum_{r \geq 1} \dots \sum_{k_1 \geq 1, \dots, k_r \geq 1: k_1 + \dots + k_r \leq \nu} \frac{(-1)^r}{2k_1! \dots 2k_r!} \hat{u}_{2k_1} \dots \hat{u}_{2k_r} \hat{f}_W^{(2k_1 + \dots + 2k_r + j)}(x),$$

and equations (7)–(9) likewise assume simpler forms.

All these high order approximations can be represented as standard kernel estimators based on adjusted kernels. Indeed, if we define

$$K_\nu(x) = K(x) + \sum_{r \geq 1} \dots \sum_{k_1 \geq 1, \dots, k_r \geq 1: k_1 + \dots + k_r \leq \nu} \frac{(-1)^{k_1 + \dots + k_r + r}}{k_1! \dots k_r!} \hat{u}_{k_1} \dots \hat{u}_{k_r} K^{(k_1 + \dots + k_r)}(x),$$

then the estimator  $\hat{f}_{X,\nu}^{(j)}(x)$  on the left-hand side of equation (6) is identical to

$$\frac{1}{nh^{j+1}} \sum_{i=1}^n K_\nu^{(j)}\left(\frac{x - W_i}{h}\right).$$

In practice the moment estimators  $\hat{u}_k$  would usually be root  $n$  consistent for the respective true moments  $u_j$ . In this case  $\hat{f}_{X,\nu}^{(j)}(x)$ , defined at equation (6), would converge to its  $\nu$ th-order approximation, defined at formula (5), at the rate equal to the slowest of the rates at which the density estimators  $\hat{f}_W^{(k+j)}(x)$ , for  $k_1 + \dots + k_r \leq \nu$ , converge to their respective limits  $f_W^{(k+j)}(x)$ .

A  $\nu$ th-order approximation to the distribution (function) of  $X$  is given by

$$\hat{F}_{X,\nu}(x) = \int_{-\infty}^x \hat{f}_{X,\nu}(t) dt,$$

and in particular cases is obtainable directly from equations (7)–(9).

All these results, and in particular expressions (3)–(9), have multivariate forms, dealing with the case where  $U$ ,  $W$  and  $X$  are  $d$ -vectors. To illustrate the version of equation (3) in this setting, write  $U = (U^{(1)}, \dots, U^{(d)})$  and  $x = (x^{(1)}, \dots, x^{(d)})$ , and, given integers  $k_1, \dots, k_d \geq 0$ , define

$$w(k_1, \dots, k_d) = (-1)^{k_1 + \dots + k_d} \frac{E\{(U^{(1)})^{k_1} \dots (U^{(d)})^{k_d}\}}{h^{k_1 + \dots + k_d} k_1! \dots k_d!}.$$

This is the  $d$ -variate analogue of  $(-1)^k u_k / h^k k!$ , appearing in the product at expression (3). Given an integer  $s \geq 1$ , let  $k_1^{(s)}, \dots, k_d^{(s)} \geq 0$ , let  $\Sigma^{(s)}$  denote summation over  $k_1^{(s)}, \dots, k_d^{(s)}$  such that at least one  $k_i^{(s)} \geq 1$  and define

$$\lambda_{j_1 \dots j_d}(x) = h^{j_1 + \dots + j_d} \frac{\partial^{j_1 + \dots + j_d}}{(\partial x_1)^{j_1} \dots (\partial x_1)^{j_d}} E \left\{ K \left( \frac{x - X}{h} \right) \right\},$$

$$\kappa_{j_1 \dots j_d}(x) = h^{j_1 + \dots + j_d} \frac{\partial^{j_1 + \dots + j_d}}{(\partial x_1)^{j_1} \dots (\partial x_1)^{j_d}} E \left\{ K \left( \frac{x - W}{h} \right) \right\},$$

these being the  $d$ -variate analogues of  $\lambda_j(x)$  and  $\kappa_j(x)$  respectively. Then, analogously to defining  $\lambda_j$  in terms of the  $\kappa_k$ s by equation (3), we have

$$\lambda_{j_1 \dots j_d}(x) = \kappa_{j_1 \dots j_d}(x) + \sum_{r=1}^{\infty} (-1)^r \sum^{(1)} \dots \sum^{(r)} \left\{ \prod_{s=1}^r w(k_1^{(s)}, \dots, k_d^{(s)}) \right\} \\ \times \kappa_{j_1 + k_1^{(1)} + \dots + k_1^{(s)}, \dots, j_d + k_d^{(1)} + \dots + k_d^{(s)}}(x).$$

Substituting an estimator, or another type of approximation, for  $E\{(U^{(1)})^{k_1} \dots (U^{(d)})^{k_d}\}$  in the definition of  $w$ , writing  $\hat{w}(k_1, \dots, k_d)$  for the resulting approximation to  $w(k_1, \dots, k_d)$  and noting that  $\hat{f}_W^{(l_1, \dots, l_d)}$  (the partial derivative of  $\hat{f}_W$   $l_j$  times with respect to  $x_j$ , for  $1 \leq j \leq d$ ) is unbiased for  $h^{-(l_1 + \dots + l_d + d)} \kappa_{l_1, \dots, l_d}$ , we see that if we replace  $w$  by  $\hat{w}$ , and  $\kappa_{l_1, \dots, l_d}$  by  $h^{l_1 + \dots + l_d + d} \hat{f}_W^{(l_1, \dots, l_d)}$ , we obtain an empirical form of equation (3). This is the multivariate analogue of formula (6).

### 3. Kernel methods for estimating $g$

For simplicity we shall develop regression applications of only the second-order version of estimators defined in Section 2. Other cases are similar, but since the variance increases quickly with estimator order then the case where the order exceeds 2 is seldom of interest: later in the present section we shall discuss this point in more detail. We shall treat both Nadaraya–Watson and local linear estimators of the mean,  $g(x)$ , of  $Y$  given  $X = x$ . In more conventional settings, where these estimators are computed from independent data  $(X_1, Y_1), \dots, (X_n, Y_n)$  having the distribution of  $(X, Y)$ , their Nadaraya–Watson and local linear forms are given by the respective ratios

$$\tilde{g}_{\text{NW}} = \frac{T_0}{S_0},$$

$$\tilde{g}_{\text{LL}} = \frac{S_2 T_0 - S_1 T_1}{S_2 S_0 - S_1^2},$$
(10)

where

$$S_r(x) = \frac{1}{nh} \sum_{i=1}^n K_r \left( \frac{x - X_i}{h} \right),$$

$$T_r(x) = \frac{1}{nh} \sum_{i=1}^n Y_i K_r \left( \frac{x - X_i}{h} \right),$$
(11)

$K_r(u) = u^r K(u)$ ,  $h$  is a bandwidth and  $K$  is a kernel function.

In view of equation (7), if we were to use  $W_i$  instead of  $X_i$  in the formulae for  $S_r$  and  $T_r$  then we would commit an error which, to second order, could be corrected by replacing  $K_r$  by  $L_r(u) = K_r - (\hat{u}_2 / 2h^2) K_r^{(2)}$ , where  $\hat{u}_2$  denotes an estimator of the variance of  $U$ . This suggests that the first-order effect of the errors  $U_i$  on the biases of  $\tilde{g}_{\text{NW}}$  and  $\tilde{g}_{\text{LL}}$  may be removed by

replacing  $S_r$  and  $T_r$  in equations (10) and (11) by  $A_r$  and  $B_r$  respectively, where

$$A_r(x) = \frac{1}{nh} \sum_{i=1}^n L_r\left(\frac{x - W_i}{h}\right),$$

$$B_r(x) = \frac{1}{nh} \sum_{i=1}^n Y_i L_r\left(\frac{x - W_i}{h}\right).$$

A more detailed justification will be given in Appendix A.2. The resulting Taylor series expansion corrected (TAYLEX) estimators are

$$\hat{g}_{\text{NW}} = \frac{B_0}{A_0},$$

$$\hat{g}_{\text{LL}} = \frac{A_2 B_0 - A_1 B_1}{A_2 A_0 - A_1^2}. \tag{12}$$

A range of alternative formulae for  $L_r$  can be given, without contradicting the bias correction properties of the estimators  $\hat{g}_{\text{NW}}$  and  $\hat{g}_{\text{LL}}$ . In theory it is necessary only that the alternative formulae agree up to terms which are of smaller order than  $\sigma_U^2$ .

The motivation behind  $\hat{g}_{\text{NW}}$  and  $\hat{g}_{\text{LL}}$  is similar to that for the estimator  $\hat{f}_{X,2}$  which was introduced in Section 2. The estimators represent Taylor series corrections of the naïve estimators  $\tilde{g}_{\text{NW}}$  and  $\tilde{g}_{\text{LL}}$ . If it is assumed that the distribution of  $U$  is Gaussian then of course all moments of  $U$  are known up to a scale factor represented by  $\sigma_U^2 = \text{var}(U)$ . TAYLEX adjusts for scale up to terms of first order in  $\sigma_U^2$ .

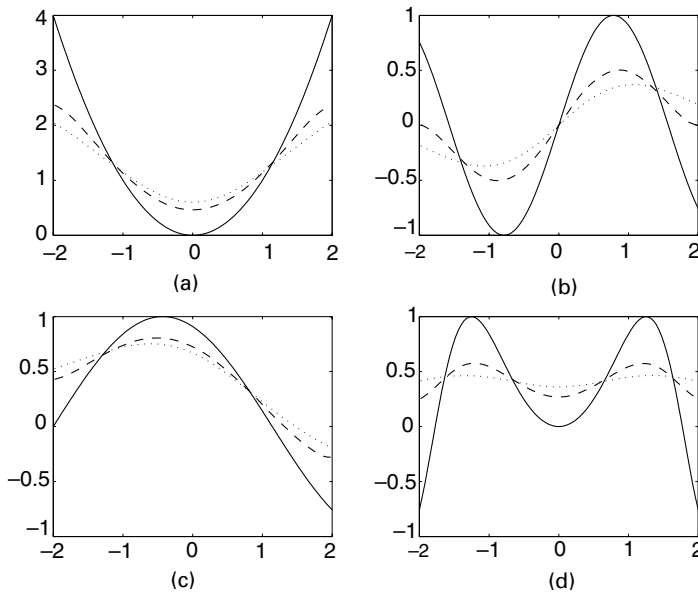
Analogously, the SIMEX method, based on polynomial interpolation of degree  $p$ , can be interpreted as an attempt to remove bias terms in  $\sigma_U^2, \dots, \sigma_U^{2p}$ . For mention of this property in a parametric context, see Cook and Stefanski's (1994) discussion of SIMEX. In nonparametric problems, using SIMEX for  $p \geq 3$  generally increases the variance so much that the reduction in bias does not compensate, and so SIMEX typically is applied only for  $p = 1$  or  $p = 2$ . Indeed, Carroll *et al.* (1999) showed that the variance of the SIMEX estimate is asymptotically the same as if measurement error were ignored, but multiplied by a factor which is independent of the regression function. They showed that, under certain conditions, using quadratic extrapolation to remove bias up to order  $\sigma_U^4$  leads to an estimator which is asymptotically nine times more variable than that based on linear extrapolation, which removes bias of the order  $\sigma_U^2$ . Additionally, the use of cubic extrapolation to remove bias up to order  $\sigma_U^6$  is asymptotically 52 times more variable than linear extrapolation. Our numerical experiments suggest that the factors are similar in the case of TAYLEX, and that taking  $p \geq 3$  generally makes TAYLEX too highly variable to be attractive.

The justification of TAYLEX that is given in Appendix A.2 will be based on theory for relatively small values of  $\sigma_U^2$ . Although analytic arguments suggest that small variances are not essential for low order Taylor series expansions to offer some reduction in bias, they nevertheless imply that Taylor approximations are likely to be better for smaller values of  $\sigma_U^2$ . Similar remarks can be made in the context of the SIMEX method, suggesting that TAYLEX should enjoy roughly comparable performance, a property which we shall confirm in Section 5.

When  $K(\cdot)$  is the standard Gaussian density, the TAYLEX kernel

$$L(u) = K(u) - \frac{\sigma_U^2}{2h^2} K^{(2)}(u)$$

is the deconvoluting kernel density for double-exponential (Laplace) errors; see Carroll and Hall (1988) and Fan and Truong (1993). In particular, the TAYLEX estimator is consistent in that



**Fig. 1.** Actual regression functions with and without measurement error (—, true function;  $\cdots$ , actual regression of  $Y$  on  $W$  when measurement errors are normally distributed;  $-\cdots-$ , regression of  $Y$  on  $W$  when measurement errors follow a double-exponential (Laplace) distribution): (a)  $g(x) = x^2$ ; (b)  $g(x) = \sin(2x)$ ; (c)  $g(x) = \sin(2 + x)$ ; (d)  $g(x) = \sin(x^2)$

case. For a bandwidth of size  $n^{-1/9}$  it gives estimators that enjoy the optimal  $L^p$  convergence rate  $n^{-2/9}$ . This suggests one reason why the TAYLEX method might work well even if the measurement errors have a Gaussian distribution. Specifically, in a wide variety of situations the observed regression functions  $E(Y|W)$  are fairly similar when the errors in variables are Gaussian and double exponential: see below for an illustration. This property, and the fact that TAYLEX estimators are consistent in the case of double-exponential errors, suggests that the TAYLEX estimator should produce a good approximation in the case of normal measurement errors.

To illustrate, consider the case where  $X$  is uniformly distributed on the interval  $[-2, 2]$ , and the measurement error variance is  $\sigma_U^2 = \frac{1}{2} \text{var}(X)$ . In this setting we treated four functions  $g$ :

- (a)  $g(x) = x^2$ ,
- (b)  $g(x) = \sin(2x)$ ,
- (c)  $g(x) = \sin(2 + x)$  and
- (d)  $g(x) = \sin(x^2)$ .

Figs 1(a), 1(b), 1(c) and 1(d) show  $g(x)$  and  $E(Y|W)$  in these respective cases, when the measurement errors are Gaussian and double exponential. The closeness of the two observed regression functions, computed in the cases of Gaussian and double-exponential errors, suggests that the TAYLEX second-order correction may, in practice, be even better than second order.

#### 4. Orthogonal series methods for estimating $f_X$ and $g$

In the orthogonal series case it is convenient to treat together the cases of deconvolution and nonparametric regression with errors in variables. In the first of these problems we are in the context of Section 2, where data  $W_i = X_i + U_i$ , for  $1 \leq i \leq n$ , are generated by model (1).

Moreover, it is assumed that we have estimators of, or other approximations to, moments of the distribution of  $U$ . Given a function  $\beta$  we may estimate  $E\{\beta(W)\}$  as  $n^{-1} \sum_i \alpha(W_i)$ . Therefore, if for each function  $\alpha$  we can express  $E\{\alpha(X)\}$  in terms of values of  $E\{\beta(W)\}$  for a range of functions  $\beta$ , and of the moments of the distribution of  $U$ , then we can estimate  $E\{\alpha(X)\}$ .

If we may do this for a sequence of functions  $\alpha_1, \alpha_2, \dots$  that are orthonormal on the support of  $X$ , then we can estimate the distribution  $F_X$ , and density  $f_X$ , of  $X$  by using orthogonal series methods. In particular, if the orthonormal sequence is complete then we may represent  $f_X$  in terms of its generalized Fourier expansion

$$f_X = \sum_{j \geq 1} a_j \alpha_j,$$

where  $a_j = E\{\alpha_j(X)\}$ . Then, given an estimator  $\hat{a}_j$  of  $a_j$ , our estimator of  $f_X$  would be

$$\hat{f}_X = \sum_{1 \leq j \leq m} \hat{a}_j \alpha_j,$$

where  $m$  was a smoothing parameter.

As in the case of the kernel-based approach that was suggested in Section 2, this method allows us to conduct explicit inference about  $\nu$ th-order approximations to  $F_X$  and  $f_X$ , where  $\nu$  denotes the maximum degree of moment products included in the approximation. However, relative to the kernel approach we now have an additional degree of freedom, which is available through the choice of the orthonormal sequence.

In the errors-in-variables problem, where data are generated by model (2), for any function  $\beta$  the variable  $Y\beta(W)$  is unbiased for  $E\{g(X)\beta(W)\}$ . In particular, given data  $(W_i, Y_i)$  on  $(W, Y)$  we can estimate  $E\{g(X)\beta(W)\}$  as the average value of  $Y_i \beta(W_i)$ . Therefore, if we can express  $E\{g(X)\alpha(X)\}$  in terms of quantities such as  $E\{g(X)\beta(W)\}$ , and of moments of the distribution of  $U$ , then we can estimate  $E\{g(X)\alpha(X)\}$ . This enables us to estimate  $f_X g$ . We have already seen how to estimate  $f_X$ , and so we can estimate  $g$ .

Indeed, if the generalized Fourier expansion of  $\gamma \equiv f_X g$  was  $\gamma = \sum_{j \geq 1} b_j \alpha_j$ , where  $b_j = E\{g(X)\alpha_j(X)\}$ , then, given an estimator  $\hat{b}_j$  of  $b_j$ , our estimator of  $g$  would be  $\hat{g} = \hat{\gamma} / \hat{f}_X$ , where  $\hat{\gamma} = \sum_{1 \leq j \leq m_1} \hat{a}_j \alpha_j$ ,  $m_1$  was a smoothing parameter and  $\hat{f}_X$  was defined as before.

We shall describe methods for estimating  $f_X g$ , since the simpler problem of estimating  $f_X$  can be addressed simply by taking  $g \equiv Y \equiv 1$ . Assume that  $g$  is uniformly bounded, that the function  $\alpha$  is infinitely differentiable, that  $E|\alpha^{(k)}(W)| < C a^k$  for constants  $a, C > 0$  and all integers  $k \geq 0$ , and that  $E(U) = 0$  and  $E\{\exp(tU)\} < \infty$  for  $t$  in some neighbourhood of the origin. Given an integer  $j \geq 1$ , let  $\sum_{(k_{2j-1}, k_{2j})}$  denote summation over integer pairs  $(k_{2j-1}, k_{2j})$  such that  $1 \leq k_{2j-1} < \infty, 0 \leq k_{2j} < \infty$  and  $k_{2j-1} + k_{2j}$  is even. We claim that, in this notation, there exists  $a_0 > 0$  such that, provided that  $0 < a < a_0$ ,

$$E\{g(X)\alpha(X)\} = E\{g(X)\alpha(W)\} - \sum_{r=0}^{\infty} \sum_{k=2}^{\infty} \sum_{(k_1, k_2)} \dots \sum_{(k_{2r-1}, k_{2r})} \frac{(-1)^{k_1+k_3+\dots+k_{2r-1}}}{k! k_1! k_2! \dots k_{2r}!} \times u_k u_{k_1+k_2} \dots u_{k_{2r-1}+k_{2r}} E\{g(X)\alpha^{(k+k_1+k_2+\dots+k_{2r})}(W)\}, \tag{13}$$

where  $u_j = E(U^j)$  and the infinite series converge absolutely. The contribution to the right-hand side of equation (13) when  $r = 0$  is interpreted as  $\sum_{k \geq 2} (k!)^{-1} u_k E\{g(X)\alpha^{(k)}(W)\}$ . A derivation of equation (13), and of its convergence properties, is outlined in Appendix A.3.

Identity (13) continues to hold, and the series converges although not always absolutely, under less stringent assumptions. Indeed, if  $\alpha$  is a polynomial of degree  $j$  then  $\alpha^{(k+k_1+k_2+\dots+k_{2r})}$  vanishes for  $k + k_1 + k_2 + \dots + k_{2r} > j$ . Therefore each of the series on the right-hand side of

equation (13), including that over  $r$ , is non-vanishing only for a finite number of indices, and so convergence is not an issue. A case in point is that of estimating  $f_X$  or  $F_X$  using Legendre polynomials, orthogonal on a compact interval.

The case of trigonometric functions is also relatively uncomplicated. There, if  $\alpha(x) = \sin(tx)$  or  $\cos(tx)$ , for a real number  $t$ , and if the distribution of  $U$  has a positive, real-valued characteristic function, as well as a finite moment-generating function in a neighbourhood of the origin, then claim (13) holds and the multiple infinite series converge, although not always absolutely. The assumption here about the distribution of  $U$  is conventional in problems of this type; see, for example, Fan (1991) and Fan and Truong (1993).

The  $\nu$ th-order approximation suggested by equation (13) involves truncating the multiple series on the right-hand side so that it includes only moment products up to and those including order  $\nu$ . In particular, defining  $a_g = E\{g(X)\alpha(X)\}$  we define its  $\nu$ th-order approximation to be

$$a_{g,\nu} = E\{g(X)\alpha(W)\} - \sum_{r \geq 0} \dots \sum_{k_1, \dots, k_{2r}: k_1 + \dots + k_{2r} \leq \nu} \frac{(-1)^{k_1 + k_3 + \dots + k_{2r-1}}}{k! k_1! k_2! \dots k_{2r}!} \times u_k u_{k_1 + k_2} \dots u_{k_{2r-1} + k_{2r}} E\{g(X)\alpha^{(k+k_1+k_2+\dots+k_{2r})}(W)\}. \tag{14}$$

Given respective estimators  $\hat{u}_j$  of the moments  $u_j$ , our estimator of  $a_{g,\nu}$  is

$$\hat{a}_{g,\nu} = \hat{a}_{g,W}^{(0)} - \sum_{r \geq 0} \dots \sum_{k_1, \dots, k_{2r}: k_1 + \dots + k_{2r} \leq \nu} \frac{(-1)^{k_1 + k_3 + \dots + k_{2r-1}}}{k! k_1! k_2! \dots k_{2r}!} \times u_k u_{k_1 + k_2} \dots u_{k_{2r-1} + k_{2r}} \hat{a}_{g,W}^{(k+k_1+k_2+\dots+k_{2r})},$$

where  $\hat{a}_{g,W}^{(k)} = n^{-1} \sum_{i \leq n} Y_i \alpha^{(k)}(W_i)$  is an unbiased estimator of  $a_{g,W}^{(k)} = E\{g(X)\alpha^{(k)}(W)\}$ .

## 5. Numerical properties

### 5.1. Introduction

We did simulations for regression and density estimation. In all cases,  $\mu_x = 0$ ,  $\sigma_x^2 = 4/3$ ,  $\sigma_u^2 = 1/3$  and  $\text{var}(Y|X) = \sigma_\varepsilon^2 = 0.05$ . The distributions for  $X$  were the normal, uniform, skew normal with index 5 and density function proportional to  $\phi(x)\Phi(5x)$  and normal mixture distributions with equal mixtures of normals with means  $\pm 1.2$  and standard deviation 0.5 rescaled to have the assumed variance. The distributions for  $\varepsilon$  were the normal and double-exponential distributions. In each setting there were 100 simulated data sets. The sample sizes were  $n = 100$  and  $n = 250$ . In the regression case, 15 different functions were used (Table 1). Staudenmeyer and Ruppert (2004) considered functions 10–12 and 14 and 15.

### 5.2. The regression and density estimation methods used in the simulations

#### 5.2.1. Density function estimator

For the naïve method, kernel density estimates were first fitted with local bandwidths computed via empirical bias bandwidth selection (EBBS), using a Gaussian kernel. The mean of the EBBS bandwidths was computed and used in a final kernel density estimate with the same kernel.

In the case of TAYLEX, for simplicity we used the same bandwidths as in the regression case; see below for details. For orthogonal series, the denominator of the regression estimate was used.

**Table 1.** Functions used in the simulations

Function	Formula
1	$x^2$
2	$\sin(2x)$
3	$\sin(2 + x)$
4	$x$
5	$\sin(x^2)$
6	$\{1 + \exp(4x)\}^{-1}$
7	$\sin(\pi x/2)[1 + 2x^2\{1 + \sin(\pi x/2)\}]^{-1}$
8	$\sin(\pi x/2)[1 + 2x^2\{1 + \sin(x)\}]^{-1}$
9	$0.2129 + 0.2300x + 0.1786x^2$
10	$1000\{(x + 2)/4\}_+^3\{1 - (x + 2)/4\}_+^3$
11	$10 \sin\{4\pi(x + 2)/4\}$
12	$15 \Phi(x/0.8)$
13	$x\{10 - 5x I(x > 0.5)\}$
14	$20[2 - \sin\{2\pi(x + 2)/4\}]^{-1}$
15	$(0.5 - x/4)_+\{10 - 5x I(x > 0.5)\}$

### 5.2.2. Regression function estimator

As well as TAYLEX and the orthogonal series methods we used the ‘naïve’ technique, which ignores measurement error entirely. There we fitted local linear regression local bandwidths computed via EBBS (Ruppert, 1997), using a Gaussian kernel (TAYLEX kernel with  $\sigma_u^2 = 0$ ). Then the mean of the EBBS bandwidths was computed and used in a local linear regression with a Gaussian kernel.

In the TAYLEX case, Nadaraya–Watson regression estimates were employed. Local bandwidths were computed by EBBS, and the bandwidth actually used was the average EBBS bandwidth multiplied by 0.75. Calculations by Staudenmeyer and Ruppert (2004) showed that there is a bias term of order  $O(\sigma_u^4)$  in the regression estimate, and holding this fixed suggests that the bandwidth should be of order smaller than the usual  $h^{-1/5}$ . The 0.75 correction is an *ad hoc* means of accomplishing this. Preliminary simulations suggest that this approach works well in decreasing both the bias and the mean-squared error.

To keep denominators from becoming unstable, the following modification was made to Nadaraya–Watson regression estimates. An estimator  $\hat{\mu}_x$  of the mean  $\mu_x$  and an estimator  $\hat{\sigma}_x^2$  of the variance  $\sigma_x^2$ , of  $X$ , were computed, the former as the sample mean of the  $W$ s and the latter as  $\text{var}(W) - \sigma_u^2$ . The denominator of the Nadaraya–Watson estimator was taken to be the maximum of the usual Nadaraya–Watson kernel density and 0.20 times the normal density with mean  $\hat{\mu}_x$  and variance  $\hat{\sigma}_x^2$ . In particular, if the Nadaraya–Watson kernel density estimate was negative, it was replaced as above. This greatly improved the stability of the method.

Orthogonal series estimators were computed for  $k = 4$ , which gave better results than  $k = 6$ . In the denominator we used the same bounding device as for the TAYLEX method. This greatly improved the stability.

### 5.3. Results for density estimation

Density function estimation results are given in Table 2. We see here that both TAYLEX and the orthogonal series estimator greatly outperform the naïve estimator, in terms of both bias and mean-squared error. This very clear result is different from what we found for regression (see below) and may be because the latter involves a division.

**Table 2.** Mean-squared error efficiencies for density estimation compared with ignoring measurement error†

$n$	$X$	$U$	Efficiencies for the following methods:	
			TAYLEX	Orthogonal series estimator
100	Normal	Normal	3.55	1.85
100	Normal	Double exponential	3.61	1.76
100	Uniform	Normal	1.55	0.95
100	Uniform	Double exponential	1.54	0.93
100	Skew normal	Normal	3.10	1.97
100	Skew normal	Double exponential	3.92	2.19
100	Mixture normal	Normal	1.69	3.37
100	Mixture normal	Double exponential	2.09	3.74
250	Normal	Normal	4.73	5.13
250	Normal	Double exponential	4.92	5.57
250	Uniform	Normal	1.84	1.84
250	Uniform	Double exponential	2.54	2.31
250	Skew normal	Normal	4.80	5.41
250	Skew normal	Double exponential	5.76	5.53
250	Mixture normal	Normal	1.97	6.87
250	Mixture normal	Double exponential	2.96	7.40

† $\sigma_x^2 = 4/3$  and  $\sigma_u^2 = 1/3$ . This calculation is based on 100 simulated data sets, with a grid from  $[-2, 2]$ .

**Table 3.** Simulation results when  $X$  and  $U$  are normally distributed†

Method	Results for the following functions:							
	Function 1		Function 3		Function 10		Function 12	
	RMSE	MAB	RMSE	MAB	RMSE	MAB	RMSE	MAB
Naïve Nadaraya–Watson	0.94	0.77	0.34	0.28	4.28	3.82	1.57	1.41
TAYLEX	0.67	0.44	0.25	0.17	3.34	2.69	1.22	0.68
Orthogonal series	1.27	0.14	0.47	0.05	6.51	0.88	3.52	0.39
SIMEX linear	0.44	0.44	0.17	0.10	2.18	1.51	0.93	0.32
SIMEX quadratic	0.58	0.12	0.24	0.05	2.70	0.68	1.75	0.19
SIMEX cubic	1.30	0.13	0.58	0.04	6.13	0.38	4.33	0.16

†See Table 1 for definitions of the functions; RMSE, root-mean-squared error; MAB, mean absolute bias;  $n = 100$ ,  $\sigma_\varepsilon^2 = 0.05$ ,  $\sigma_x^2 = 4/3$  and  $\sigma_u^2 = 1/3$ . There were 100 simulated data sets.

#### 5.4. Results for regression

TAYLEX lowers the bias: when averaged over the 15 functions, for each of the eight combinations of distributions for  $X$  and  $U$  it had approximately 35% less mean absolute bias than the estimator that ignores measurement error. It also outperforms the naïve estimator in terms of the mean-squared error. For the standard quadratic extrapolant function, TAYLEX and SIMEX have approximately the same performance; compare the second and fourth rows of Table 3. However, TAYLEX enjoys much faster computation. This would be a significant factor if either method were used in connection with simulation, for example employing bootstrap techniques to construct confidence bands or to choose the bandwidth.

SIMEX with a cubic extrapolant function is, as expected from theory, wildly variable; see the fifth row of Table 3. Nor did the orthogonal series approach perform well in this context, tending to suffer from high variability due to fluctuations in the denominator.

An alternative method is the version of SIMEX, using an efficient approach to bandwidth choice, developed by Staudenmeyer and Ruppert (2004). A set of MATLAB programs for implementation are available from [http://stat.tamu.edu/~carroll/matlab\\_programs.html](http://stat.tamu.edu/~carroll/matlab_programs.html). The execution time depends strongly on the number of grid points at which the function is to be evaluated: about 2 min for five grid points, about 5 min for 15 grid points, about 9 min for 25 grid points and about 18 min for 51 grid points (on a Pentium 4 computer running at 1.9 GHz). Therefore, we could only run a few selected simulations in the case where both  $X$  and  $U$  were normally distributed. The results, not given here, represent about a 30% improvement on the performance of either standard SIMEX or our TAYLEX method.

## 6. Discussion

In this paper, we have suggested kernel (TAYLEX) and orthogonal series methods that are applicable to both deconvolution and regression with errors in explanatory variables. They are based on defining a low order approximation and proceed by constructing relatively accurate estimators of that quantity rather than attempting to estimate the true target functions consistently. The information that we require about the error distribution (the error-in-variables distribution, in the case of regression) is only in the form of low order moments and so is readily obtainable by rudimentary analysis of indirect measurements of errors, e.g. through repeated measurements. In particular, we do not need to estimate a function, such as a characteristic function, which expresses detailed properties of the error distribution. This feature of our methods, coupled with the fact that all our estimators are explicitly defined in terms of readily computable averages, means that the methods are particularly economical in computing time.

Our results indicate that, for density estimation, both methods give considerable gains in mean-squared error efficiencies over the method that ignores the errors in variables. See Table 2, where these efficiencies typically range from 2.0 to 5.0 and all exceed 1.5 even for samples of size  $n = 100$ . For regression, the results are less impressive but the mean-squared error efficiencies still average 1.43 for  $n = 100$  and 1.61 for  $n = 250$ .

## Acknowledgements

Carroll's research was done while visiting the Centre for Mathematics and Its Applications at the Australian National University and was supported by a grant from the National Cancer Institute (CA-57030) and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30-ES09106). We thank the Associate Editor and two referees for their detailed and helpful comments.

## Appendix A: Technical arguments

### A.1. Proof of equation (3)

Put  $v_k = E(U/h)^k$ . By analyticity of  $K$ ,  $\kappa_j = \sum_{k \geq 0} (-1)^k (k!)^{-1} v_k \lambda_{j+k}$ , whence it follows that

$$\lambda_j = \kappa_j - \sum_{k=1}^{\infty} \frac{(-1)^k}{k!} v_k \lambda_{j+k}. \quad (15)$$

Substitute for  $\lambda_{j+k}$  in equation (15) by replacing  $j$  by  $j + k$  in the same formula, thereby obtaining

$$\begin{aligned} \lambda_j &= \kappa_j - \sum_{k_1=1}^{\infty} \frac{(-1)^{k_1}}{k_1!} v_{k_1} \left\{ \kappa_{j+k_1} - \sum_{k_2=1}^{\infty} \frac{(-1)^{k_2}}{k_2!} v_{k_2} \lambda_{j+k_1+k_2} \right\} \\ &= \kappa_j - \sum_{k_1=1}^{\infty} \frac{(-1)^{k_1}}{k_1!} v_{k_1} \kappa_{j+k_1} + \sum_{k_1=1}^{\infty} \sum_{k_2=1}^{\infty} \frac{(-1)^{k_1+k_2}}{k_1! k_2!} v_{k_1} v_{k_2} \lambda_{j+k_1+k_2}. \end{aligned}$$

Result (3) follows on iterating this argument.

### A.2. Derivation of the Taylor series expansion approximation

We shall show that the TAYLEX estimator  $\hat{g}_{LL}$ , defined at expression (12), corrects for second-order contributions of the error  $U$ . The case of  $\hat{g}_{NW}$  is simpler. However, in each instance the proof is more complex than noting the similarity to the result proved in Appendix A.1, since there is potential for interaction between terms in the numerator and denominator in the definition of  $\hat{g}_{LL}$ .

Since

$$K_r\left(\frac{x - W_i}{h}\right) = K_r\left(\frac{x - X_i}{h}\right) - \frac{U_i}{h} K'_r\left(\frac{x - X_i}{h}\right) + \frac{1}{2} \left(\frac{U_i}{h}\right)^2 K_r^{(2)}\left(\frac{x - X_i}{h}\right) + \dots,$$

then, if we could observe the  $U_i$ s, ‘corrected’ forms of  $S_r$  and  $T_r$ , with the corrections applied up to and including quadratic terms, would be respectively

$$\begin{aligned} S_{2,r}(x) &= S_{1,r}(x) + \frac{1}{nh} \sum_{i=1}^n \left\{ \frac{U_i}{h} K'_r\left(\frac{x - X_i}{h}\right) - \frac{1}{2} \left(\frac{U_i}{h}\right)^2 K_r^{(2)}\left(\frac{x - X_i}{h}\right) \right\}, \\ T_{2,r}(x) &= T_{1,r}(x) + \frac{1}{nh} \sum_{i=1}^n Y_i \left\{ \frac{U_i}{h} K'_r\left(\frac{x - X_i}{h}\right) - \frac{1}{2} \left(\frac{U_i}{h}\right)^2 K_r^{(2)}\left(\frac{x - X_i}{h}\right) \right\}, \end{aligned} \tag{16}$$

where  $S_{1,r}$  and  $T_{1,r}$  denote the versions of  $S_r$  and  $T_r$  at expression (11) obtained on replacing  $X_i$  there by  $W_i$ .

Consider substituting  $S_{2,r}$  and  $T_{2,r}$  for  $S_r$  and  $T_r$  respectively at expression (10), and Taylor series expansion of the result, to calculate a ‘corrected’ version of the formula for  $\tilde{g}_{LL}$  when, in defining the latter, we use  $W_i$  instead of  $X_i$  as design points. To compute the effect that the correction has on the bias, take the expectation of the Taylor series expansion with respect to the  $U_i$ s, conditional on the pairs  $(X_i, Y_i)$ . If  $U_i = \sigma_U V_i$ , where the distribution of  $V_i$  is fixed as  $\sigma_U$  decreases, we may express the effect that the correction has on the bias as a series in  $\sigma_U^2, \sigma_U^3, \dots$ .

We focus on the term in  $\sigma_U^2$  in this expansion. It is made up from two sources: ‘type A’ terms that are expected values of contributions that come directly from the quantities  $(U_i/h)^2$  in the formulae for  $S_{2,r}$  and  $T_{2,r}$  at equation (16) and ‘type B’ terms that are expected values of contributions that come from products of two quantities in  $U_i/h$  from those formulae. (Hence, type B terms involve an interaction between the numerator and denominator in  $\hat{g}_{LL}$ .) Type A terms are multiplied by the factor  $n(nh)^{-1}$ , which has been applied to each of the series at equation (16). However, type B terms, since they derive from the expectation of the product of two series, are multiplied by  $n(nh)^{-2}$ . To appreciate why, note that the off-diagonal terms vanish from the expectation of the product, owing to independence of the zero-mean variables  $U_i$ . Since  $nh$  is large then type B terms will contribute significantly less than type A terms.

Therefore, the dominant contributions to bias corrections of size  $\sigma_U^2$  come directly from the terms in  $(U_i/h)^2$  at equation (16). From this point, minor modifications of the argument in Appendix A.1 can be used to prove that these terms are removed by defining the estimator  $\hat{g}_{LL}$  with the kernel  $K$  replaced by  $L$ . The resulting estimator is just  $\hat{g}_{LL}$ .

### A.3. Derivation of equation (13)

By Taylor series expansion and the independence of  $U$  and  $X$ ,

$$E\{g(X) \alpha(W)\} = \sum_{k=0}^{\infty} \frac{1}{k!} E(U^k) E\{g(X) \alpha^{(k)}(X)\}.$$

Writing  $E\{g(X)\alpha^{(k)}(X)\}$  as  $E\{g(X)\alpha^{(k)}(W-U)\}$ , and, for  $k \geq 2$ , Taylor series expansion of  $\alpha^{(k)}(W-U)$  about  $W$ , then writing  $E\{U^{k_1}g(X)\alpha^{(k+k_1)}(W)\}$  as  $E\{U^{k_1}g(X)\alpha^{(k+k_1)}(X+U)\}$ , and Taylor series expansion of  $\alpha^{(k+k_1)}(X+U)$  about  $X$  and continuing this argument indefinitely, we obtain the formula

$$\begin{aligned}
 E\{g(X)\alpha(W)\} &= E\{g(X)\alpha(X)\} + \sum_{k=2}^{\infty} \frac{1}{k!} E(U^k) E\{g(X)\alpha^{(k)}(W)\} \\
 &+ \sum_{k=2}^{\infty} \sum_{k_1=1}^{\infty} \sum_{k_2=0}^{\infty} \frac{(-1)^{k_1}}{k!k_1!k_2!} E(U^k) E(U^{k_1+k_2}) E\{g(X)\alpha^{(k+k_1+k_2)}(W)\} \\
 &+ \sum_{k=2}^{\infty} \sum_{k_1=1}^{\infty} \sum_{k_2=0}^{\infty} \sum_{k_3=1}^{\infty} \sum_{k_4=0}^{\infty} \frac{(-1)^{k_1+k_3}}{k!k_1!k_2!k_3!k_4!} E(U^k) E(U^{k_1+k_2}) \\
 &\times E(U^{k_3+k_4}) E\{g(X)\alpha^{(k+k_1+k_2+k_3+k_4)}(W)\} + \dots
 \end{aligned} \tag{17}$$

Result (17) is equivalent to equation (13). (Note that we may drop terms in  $E(U^{k_{2j-1}+k_{2j}})$  for which  $k_{2j-1} + k_{2j}$  is odd.)

The argument above is valid provided that the multiple infinite series that it produces converges absolutely. This property is readily checked if

- (a)  $\sup |g| < \infty$ ,
- (b)  $E|\alpha^{(k)}(W)| < C\alpha^k$  and
- (c)  $E\{\exp(tU)\} < \infty$  for all  $t$  in some neighbourhood of the origin.

The series at equation (17) can also be shown to converge, to the limit on the left-hand side, when  $\alpha(x) = \sin(tx)$  or  $\cos(tx)$ .

## References

Carroll, R. J. and Hall, P. (1988) Optimal rates of convergence for deconvolving a density. *J. Am. Statist. Ass.*, **83**, 1184–1186.

Carroll, R. J., Küchenhoff, H., Lombard, F. and Stefanski, L. A. (1996) Asymptotics for the SIMEX estimator in nonlinear measurement error models. *J. Am. Statist. Ass.*, **91**, 242–250.

Carroll, R. J., Maca, J. D. and Ruppert, D. (1999) Nonparametric regression with errors in covariates. *Biometrika*, **86**, 541–554.

Carroll, R. J., Ruppert, D. and Stefanski, L. A. (1995) *Measurement Error in Nonlinear Models*. London: Chapman and Hall.

Cator, E. A. (2001) Deconvolution with arbitrarily smooth kernels. *Statist. Probab. Lett.*, **54**, 205–214.

Chen, S. X. and Cowling, A. (2001) Measurement errors in line transect surveys where detectability varies with distance and size. *Biometrics*, **57**, 732–742.

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

Cordy, C. B. and Thomas, D. R. (1997) Deconvolution of a distribution function. *J. Am. Statist. Ass.*, **92**, 1459–1465.

Devroye, L. (1989) Consistent deconvolution in density estimation. *Can. J. Statist.*, **17**, 235–239.

Efromovich, S. (1997) Density estimation for the case of supersmooth measurement error. *J. Am. Statist. Ass.*, **92**, 526–535.

van Es, B. and Uh, H. W. (2000) Multi bandwidth kernel estimators for nonparametric deconvolution problems: asymptotics and finite sample performance. *J. Nonparam. Statist.*, **13**, 107–128.

Fan, J. (1991) On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.*, **19**, 1257–1272.

Fan, J. (1992) Deconvolution with supersmooth distributions. *Can. J. Statist.*, **20**, 155–169.

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

Fotopoulos, S. B. (2000) Invariance principles for deconvolving kernel density estimation for stationary sequences of random variables. *J. Statist. Plannng Inf.*, **86**, 31–50.

Goldenshluger, A. (1999) On pointwise adaptive nonparametric deconvolution. *Bernoulli*, **5**, 907–925.

Gould, W. R., Stefanski, L. A. and Pollock, K. H. (1997) Effects of measurement error on catch-effort estimation. *Can. J. Fish. Aquat. Syst.*, **54**, 898–906.

Hesse, C. H. (1995a) Deconvolving a density from partially contaminated observations. *J. Multiv. Anal.*, **55**, 246–260.

Hesse, C. H. (1995b) Deconvolving a density from contaminated dependent observations. *Ann. Inst. Statist. Math.*, **47**, 645–663.

Hesse, C. H. (1999) Data-driven deconvolution. *J. Nonparam. Statist.*, **10**, 343–373.

- Holcomb, J. P. (1999) Regression with covariates and outcome calculated from a common set of variables measured with error: estimation using the SIMEX method. *Statist. Med.*, **18**, 2847–2862.
- Jongbloed, G. and van Zuijlen, M. (1998) Isotonic inverse estimators for nonparametric deconvolution. *Ann. Statist.*, **26**, 2395–2406.
- Kim, J. and Gleser, L. J. (2000) SIMEX approaches to measurement error in ROC studies. *Commun. Statist. Theory Meth.*, **29**, 2473–2491.
- Lin, X. H. and Carroll, R. J. (1999) SIMEX variance component tests in generalized linear mixed measurement error models. *Biometrics*, **55**, 613–619.
- Lin, X. H. and Carroll, R. J. (2000) Nonparametric function estimation for clustered data when the predictor is measured without/with error. *J. Am. Statist. Ass.*, **95**, 520–534.
- Liu, M. C. and Taylor, R. (1989) A consistent nonparametric density estimator for the deconvolution problem. *Can. J. Statist.*, **17**, 427–438.
- Luo, M., Stokes, L. and Sager, T. (1998) Estimation of the CDF of a finite population in the presence of a calibration sample. *Environ. Ecol. Statist.*, **5**, 277–289.
- Masry, E. (1993) Strong consistency and rates for deconvolution of multivariate densities of stationary processes. *Stoch. Processes Appl.*, **47**, 53–74.
- Mendelsohn, J. and Rice, J. A. (1982) Deconvolution of microfluorometric histograms with  $B$  splines. *J. Am. Statist. Ass.*, **77**, 748–753.
- Novick, S. J. and Stefanski, L. A. (2002) Corrected score estimation via complex variable simulation extrapolation. *J. Am. Statist. Ass.*, **97**, 472–481.
- Pensky, M. and Vidakovic, B. (1999) Adaptive wavelet estimator for nonparametric density deconvolution. *Ann. Statist.*, **27**, 2033–2053.
- Ruppert, D. (1997) Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *J. Am. Statist. Ass.*, **92**, 1049–1062.
- Staudenmeyer, J. and Ruppert, D. (2004) Local polynomial regression and simulation–extrapolation. *J. R. Statist. Soc. B*, **66**, 17–30.
- Stefanski, L. (1990) Rates of convergence of some estimators in a class of deconvolution problems. *Statist. Probab. Lett.*, **9**, 229–235.
- Stefanski, L. (2000) Measurement error models. *J. Am. Statist. Ass.*, **95**, 1353–1358.
- Stefanski, L. A. and Bay, J. M. (1996) Simulation extrapolation deconvolution of finite population cumulative distribution function estimators. *Biometrika*, **83**, 407–417.
- Stefanski, L. and Carroll, R. J. (1990) Deconvoluting kernel density estimators. *Statistics*, **21**, 169–184.
- Stefanski, L. A. and Cook, R. J. (1995) Simulation-extrapolation: the measurement error jackknife. *J. Am. Statist. Ass.*, **90**, 1247–1256.
- Thurigen, D., Spiegelman, D., Blettner, M., Heurer, C. and Brenner, H. (2000) Measurement error correction using validation data: a review of methods and their applicability in case-control studies. *Statist. Meth. Med. Res.*, **9**, 447–474.
- Wand, M. P. (1998) Finite sample performance of deconvolving kernel density estimators. *Statist. Probab. Lett.*, **37**, 131–139.
- Wang, N., Carroll, R. J. and Liang, K. Y. (1996) Quasilikelihood estimation in measurement error models with correlated replicates. *Biometrics*, **52**, 401–411.
- Wang, Y. Z. (1999) Change-points via wavelets for indirect data. *Statist. Sin.*, **9**, 103–117.
- Youndje, T. and Wells, M. T. (2002) Least squares cross-validation for the kernel deconvolution density estimator. *C. R. Math. Acad. Sci.*, **334**, 509–513.
- Yuan, M. and Chen, J. Q. (2002) Deconvolving multidimensional density from partially contaminated observations. *J. Statist. Plannng Inf.*, **104**, 147–160.
- Zhang, C.-H. (1990) Fourier methods for estimating mixing densities and distributions. *Ann. Statist.*, **18**, 806–831.
- Zhang, S. P. and Karunamuni, R. J. (2000) Boundary bias correction for nonparametric deconvolution. *Ann. Inst. Statist. Math.*, **52**, 612–629.